Methods in Molecular Biology 1802

Springer Protocols

Sebastian Boegel *Editor*

HLA Typing

Methods and Protocols

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

HLA Typing

Methods and Protocols

Edited by

Sebastian Boegel

Johannes Gutenberg University of Mainz, Mainz, Germany

Humana Press

Editor
Sebastian Boegel
Johannes Gutenberg University of Mainz
Mainz, Germany

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-8545-6 ISBN 978-1-4939-8546-3 (eBook)
https://doi.org/10.1007/978-1-4939-8546-3

Library of Congress Control Number: 2018943706

Printed on acid-free paper

This Humana Press imprint is published by the registered company Springer Science+Business Media, LLC part of Springer Nature.
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface

The human leukocyte antigens (HLA) complex is located on chromosome 6 and encodes a series of genes, including its most prominent members, i.e., the highly polymorphic classical HLA class I (HLA-A, -B, -C) and class II (HLA-DR, -DQ, -DP) molecules. These molecules are glycoproteins with the primary task to bind and present self, abnormal self (e.g., neo-epitopes arising from mutations), and foreign peptide antigens derived from intracellular (mainly HLA class I) or from extracellular proteins (mainly HLA class II) on the surface of nucleated cells. HLA-dependent peptide presentation is critical to an effective adaptive immune system in which the antigens bound to HLA molecules are recognized by T lymphocytes, leading to an immune response.

The HLA system is characterized by a vast amount of allelic variants resulting in a wide variety of HLA allele combinations, i.e., the HLA type, that differ between individuals. Determination of an individual's composition of HLA alleles—HLA typing—is essential for clinical work: the HLA type is a key parameter in solid organ and hematopoietic cell transplantation (Chapter 1) and certain HLA alleles were shown to have a high degree of association with various autoimmune diseases (Chapter 2). Two resources, which are reviewed and described in this book, play key roles in the development and application of HLA typing methods and immunological research in general: (1) the Immuno Polymorphism Database (IPD) catalogues the vast amount of sequence variants (Chapter 3) and (2) the Allele Frequency Net database contains information about HLA allele frequencies throughout many different populations (Chapter 4).

High-throughput DNA and RNA sequencing enables the rapid generation of billions of short nucleic acid sequence fragments. In the last years, the field has seen a rapid evolution of both laboratory and *in silico* methods to determine the HLA type in a massively high-throughput fashion. As the costs of sequencing continue to fall, we anticipate that HLA typing, especially using next generation sequencing, will be the future of HLA typing.

Thus, the focus of this volume is to gather a variety of protocols using high-throughput methods for HLA typing. Chapters 5–10 describe *wet lab* protocols comprising different methodologies and sequencing platforms. Chapters 11–18 summarize *in silico* tools, which are able to determine the HLA type from high-throughput data. This comprises the imputation of the HLA type from SNPs using microarray data (Chapter 11) and identification of the HLA alleles from (standard or targeted) DNA and RNA next generation sequencing data (Chapters 12–18), including a webserver (Chapter 18), as well as a software tool for HLA haplotype frequency estimation (Chapter 19).

This volume has not been possible without the contributions of the leading experts in the fields of HLA typing using high-throughput data, HLA sequence analysis, bioinformatics, and immunogenomics. I am very grateful for the discussions I had throughout the process of editing this book and for their consent to share their expertise and knowledge with the scientific community.

I wish to thank Springer and the Series Editor Prof. John M. Walker for giving me the opportunity to assemble this outstanding collection of manuscripts and for their excellent support throughout the process of editing this book. Last but not least, I especially wish to thank my friend and mentor Dr. John C. Castle for his guidance and valuable scientific input.

Mainz, Germany *Sebastian Boegel*

Contents

Contributors

Jashan P. Abraham • *Anthony Nolan Research Institute, London, UK*

Sheetal Ambardar • *Centre for Functional Genomics and Bioinformatics, Institute of TransDisciplinary Health Sciences and Technology, Bangalore, Karnataka, India*

Yu Bai • *Regeneron Pharmaceuticals, Tarrytown, NY, USA*

Dominic J. Barker • *Anthony Nolan Research Institute, London, UK*

Aleksandra Biedrzycka • *Institute of Nature Conservation, Polish Academy of Sciences, Kraków, Poland*

Gergely Bodis • *Bioscientia Institut für Medizinische Diagnostik GmbH, Ingelheim, Germany; Acura Rheumatology Center Rhineland Palatine, Bad Kreuznach, Germany*

Sebastian Boegel • *TRON gGmbH–Translational Oncology at Johannes Gutenberg-University Medical Center gGmbH, Mainz, Germany*

Thomas Bukur • *TRON gGmbH–Translational Oncology at Johannes Gutenberg-University Medical Center gGmbH, Mainz, Germany*

John C. Castle • *TRON gGmbH–Translational Oncology at Johannes Gutenberg-University Medical Center gGmbH, Mainz, Germany; Agenus Inc, Lexington MA, USA*

Claire H. Edgerly • *Department of Pathology and Laboratory Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA*

Philip K. Ehrenberg • *U.S. Military HIV Research Program (MHRP), Walter Reed Army Institute of Research, Silver Spring, MD, USA*

Andre Franke • *Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany*

Wen Fury • *Regeneron Pharmaceuticals, Tarrytown, NY, USA*

Aviva Geretz • *U.S. Military HIV Research Program (MHRP), Walter Reed Army Institute of Research, Silver Spring, MD, USA; Henry M. Jackson Foundation for the Advancement of Military Medicine (HJF), Bethesda, MD, USA*

Gurpreet Ghattaoraya • *Institute of Integrative Biology, University of Liverpool, Liverpool, UK*

Faviel F. Gonzalez-Galarza • *Department of Molecular Immunobiology, Faculty of Medicine, Centre for Biomedical Research, Autonomous University of Coahuila, Torreón, Coahuila, Mexico*

Malali Gowda • *Centre for Functional Genomics and Bioinformatics, Institute of TransDisciplinary Health Sciences and Technology, Bangalore, Karnataka, India*

Koichiro Higasa • *Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan*

Hidetoshi Inoko • *Department of Molecular Life Science, Division of Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, Isehara, Kanagawa, Japan*

Andrew R. Jones • *Institute of Integrative Biology, University of Liverpool, Liverpool, UK*

Simonas Juzenas • *Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany; Laboratory of Clinical and Molecular Gastroenterology, Institute for Digestive Research, Lithuanian University of Health Sciences, Kaunas, Lithuania*

Shuji Kawaguchi • *Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan*
Carl Kingsford • *Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*
Jerzy K. Kulski • *Department of Molecular Life Science, Division of Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, Isehara, Kanagawa, Japan; School of Psychiatry and Clinical Neurosciences, The University of Western Australia, Crawley, WA, Australia*
James Lan • *Nephrology and Kidney Transplantation, University of British Columbia, Vancouver General Hospital, Vancouver, BC, Canada*
Kathrin Lang • *DKMS Life Science Lab, Dresden, Germany*
Vinzenz Lange • *DKMS Life Science Lab, Dresden, Germany*
Heewook Lee • *Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA*
Chang Liu • *Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA*
Giuseppe Maccari • *Anthony Nolan Research Institute, London, UK; The Pirbright Institute, Surrey, UK*
Steven G. E. Marsh • *Anthony Nolan Research Institute, London, UK; UCL Cancer Institute, University College London, London, UK*
Fumihiko Matsuda • *Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan*
Antony McCabe • *Institute of Integrative Biology, University of Liverpool, Liverpool, UK*
Eduardo J. Melo dos Santos • *Human and Medical Genetics, Institute of Biological Sciences, Federal University of Para, Belém, PA, Brazil*
Derek Middleton • *Transplant Immunology Laboratory, Royal Liverpool and Broadgreen University Hospital, Liverpool, UK; Institute of Infection and Global Health, University of Liverpool, Liverpool, UK*
Magdalena Migalska • *Evolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland*
Philipp Quenzel • *DKMS Life Science Lab, Dresden, Germany*
James Robinson • *Anthony Nolan Research Institute, London, UK; UCL Cancer Institute, University College London, London, UK*
Ugur Sahin • *TRON gGmbH–Translational Oncology at Johannes Gutenberg-University Medical Center gGmbH, Mainz, Germany*
Jürgen Sauter • *DKMS gemeinnützige GmbH, Tübingen, Germany*
Christian Schäfer • *DKMS gemeinnützige GmbH, Tübingen, Germany*
Alexander H. Schmidt • *DKMS Life Science Lab, Dresden, Germany; DKMS, Tübingen, Germany; DKMS gemeinnützige GmbH, Tübingen, Germany*
Gerhard Schöfl • *DKMS Life Science Lab, Dresden, Germany*
Andreas Schwarting • *Acura Rheumatology Center Rhineland Palatine, Bad Kreuznach, Germany; Division of Rheumatology and Clinical Immunology, Department of Internal Medicine I, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany*
Alvaro Sebastian • *Sixth Researcher, Poznan, Poland; Instituto Aragonés de Empleo (INAEM), Zaragoza, Spain; Evolutionary Biology Group, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland*

Takashi Shiina • *Department of Molecular Life Science, Division of Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, Isehara, Kanagawa, Japan*
Vineeth Surendranath • *DKMS Life Science Lab, Dresden, Germany*
Shingo Suzuki • *Department of Molecular Life Science, Division of Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, Isehara, Kanagawa, Japan*
András Szolek • *Applied Bioinformatics, Department for Computer Science, University of Tübingen, Tübingen, Germany*
Louise Takeshita • *Institute of Integrative Biology, University of Liverpool, Liverpool, UK*
Rasmi Thomas • *U.S. Military HIV Research Program (MHRP), Walter Reed Army Institute of Research, Silver Spring, MD, USA; Henry M. Jackson Foundation for the Advancement of Military Medicine (HJF), Bethesda, MD, USA*
Victoria Toth • *Bioscientia Institut für Medizinische Diagnostik GmbH, Ingelheim, Germany; Acura Rheumatology Center Rhineland Palatine, Bad Kreuznach, Germany*
Melanie Vollstedt • *Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany*
David Wang • *Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, USA*
Eric T. Weimer • *Department of Pathology and Laboratory Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA*
Michael Wittig • *Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany*
Ryo Yamada • *Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan*
Xiao Yang • *Grail, Inc., Menlo Park, CA, USA*
Yuxin Yin • *Department of Pathology & Laboratory Medicine, UCLA Immunogenetics Center, Los Angeles, CA, USA*
Qiuheng Zhang • *Department of Pathology & Laboratory Medicine, UCLA Immunogenetics Center, Los Angeles, CA, USA*
Xiuwen Zheng • *Department of Biostatistics, University of Washington, Seattle, WA, USA*

Chapter 1

The Past, Present, and Future of HLA Typing in Transplantation

Claire H. Edgerly and Eric T. Weimer

Abstract

The HLA region is the most polymorphic genes in the human genome and is associated with an increasing number of disease states. Historically, HLA typing methodology has been governed by phenotypic determination. This practice has evolved into the use of molecular methods such as real-time PCR, sequence-specific oligonucleotides, and sequencing-based methods. Numerous studies have identified HLA matching as a key determinate to improve patient outcomes from transplantation. Solid-organ transplants focus on HLA-DRB1 in renal organ allocation while hematopoietic cell transplants focus on HLA-A, -B, -C, -DRB1 matching. The role of HLA typing in the future will be driven by HLA expression, understanding of HLA haplotypes, and rapid HLA typing.

Key words Single molecule sequencing, NGS, HLA, Immunogenetics, RNASeq, GVHD, Rejection

1 Introduction

It has been 63 years since the first successful renal transplant performed at Peter Bent Brigham Hospital [1]. In the subsequent decades, transplantation expanded to include liver, heart, lung, pancreas, and bone marrow transplants [2–6]. HLA were not recognized to play an important role in transplantation until around 1952 by Dausset's group [7]. Following the structural identification of HLA, there have been numerous advances in not only HLA typing technology but the utilization of HLA typing in transplantation [8–12]. This advancement has meant that HLA typing has gone from identification of HLA proteins to identification of HLA gene polymorphisms. HLA gene polymorphisms were first identified within regions that encode the antigen-binding domains of the classic HLA genes (HLA-A, -B, -C, -DRB1, -DQB1, -DPB1), where the majority of the HLA diversity lies [13, 14]. Now HLA gene polymorphisms within the untranslated regions, additional exonic, and intronic sequences are interrogated to gleam additional

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_1,

information about HLA regulation, splicing, and protein expression [15–25].

2 Solid-Organ Transplantation

HLA matching in solid-organ transplant such as kidney and liver confers multiple benefits, and conversely, mismatches are associated with worse outcomes [26–28]. However, there are tradeoffs in the drive for a matched donor, such as prolonged wait times and potential disease progression. The continuing refinement of immunosuppressive regimens also contributes to greater tolerance for some mismatches. The approach to HLA matching for solid-organ transplant has gone through several rounds of evolution in recent years.

The Kidney Allocation System (KAS) introduced in 2014 aimed to bring increased fairness and efficiency to the distribution of deceased donor kidneys. Highly sensitized patients, who had routinely been subject to extended waiting periods for an appropriate organ, were given priority for organs [29–33]. A review conducted 1 year after implementation (Stewart) showed that the KAS had produced a significant change in allocation methods, including markedly increased rates of transplantation in highly sensitized patients with PRAs of 99–100%, as well as patients who had been on dialysis for 10+ years [34]. The review also noted downsides of the new policy, such as the inevitable adverse effects of long travel times for organs allocated to patients at long distances.

HLA matching in solid-organ transplantation has traditionally centered on the HLA-A, -B and –DR loci, with matches at these loci associated with better overall survival. These benefits appeared magnified in highly sensitized patients (high cPRA) [28]. Advances in immunosuppression have led to decreased emphasis on HLA matching in solid-organ transplants [28, 35]. However, there are still benefits to HLA matching. HLA-DR has been shown to have the strongest contribution to graft outcome in solid-organ transplant [8, 36–39]. This also had the benefit of allowing for easier matches, as there is less diversity in DRB1 antigens than those in HLA-A or HLA-B. However, HLA laboratories in the US are required by the United Network for Organ Sharing (UNOS) to report HLA-A, -B, -C, Bw4, Bw6, -DRB1, -DRB3/4/5, –DQA1, -DQB1, -DPB1.

For many years, the gold standard for low-resolution HLA typing was serologic typing, which used sera collected from volunteers with known HLA antibodies, complement, and a vital dye [40]. If the cells put into the wells have antigens matching the antibodies within the wells, complement is activated, the cells die and the vital dye is taken up. Serologic typing can be performed quickly and was therefore helpful in urgent situations such as

deceased donor typing. However, antigen differences that may have significant effects in vivo are not always identified in serologic evaluation [40, 41]. Another problem associated with serologic typing is identifying sources of specific antibodies, a challenge that increases as more clinically significant HLA antigens are identified. As a result, only HLA typing by molecular methods is acceptable for solid-organ transplant. The majority of HLA laboratories use either real-time PCR or sequence-specific oligonucleotide (SSO) to achieve low-resolution HLA typing for solid-organ transplants. The speed, accuracy, and convenience of both technologies makes them excellent choices for deceased donors HLA typing as well.

3 Hematopoietic Cell Transplantation (HCT)

In patients with severe hematologic disease (benign or malignant), hematopoietic cell transplantation (HCT) can be a life-saving intervention. Due to the extensive diversity of HLA genes in worldwide populations, the search for an appropriate match often starts within the patient's family, specifically sibling donors. Because HLA alleles are inherited as haplotypes from each parent, the possibility of a full HLA match is 25% for each sibling, and the possibility of a "haploidentical" (single haplotype in common) is 50%. Low-resolution typing is generally sufficient to identify potential matches.

The advent of high-resolution (allele level) HLA typing introduced a new level of specificity to this process. In addition to direct DNA sequencing, other options for molecular typing include sequence-specific polymerase chain reaction (SSP-PCR) and sequence-specific oligonucleotide probes (SSOP). Molecular methods of typing allow much greater resolution in HLA typing and can provide information as specific as the amino acid level differences [42–45]. As of October 2017, the majority (approximately 85%) of American Society for Histocompatibility and Immunogenetics (ASHI) accredited HLA laboratories use Sanger sequencing to achieve high-resolution HLA typing for hematopoietic cell transplantation (HCT).

Although an identically matched related donor is ideal for HCT, such a match is only found in about 15–30% of recipients [46]. When an appropriate family member donor is not available, HCT using a volunteer unrelated donor can be a viable option when an HLA-appropriate donor is identified. As established by a 2007 study comparing characteristics of more than 3800 patient-donor pairs, high-resolution (allele-level) matching of HLA-A, -B, -C, and –DRB1 alleles is correlated with better overall survival, with mismatches linked to increased rates of treatment-related mortality, acute GVHD, and decreased survival [47, 48]. In this study, high-resolution matching at all of these loci was considered 8/8 "fully

matched." No difference was seen between high- vs low-resolution mismatches. This study confirmed an earlier study by Speiser et al. [12] that demonstrated the clinical value in high-resolution HLA matching for HCT patients.

HLA-DQ and –DP status were examined in secondary analyses, and –DQ single mismatches (antigen or allele) were not found to significantly affect outcome. In the study, a –DP mismatch did not affect survival, but showed a slight increase in risk of acute GVHD. Matching HLA-DPB1 introduces an additional challenge, since about 85% of pairs with an 8/8 match have at least one –DP mismatch [48].

The importance of matching at the HLA-DPB1 locus has been debated for some time. Mismatches at the -DPB1 locus were not found to be associated with worse outcomes in a 2012 retrospective analysis of 141 donor-recipient pairs [49]. However, in another study, non-permissive mismatches were associated with increased graft rejection and overall mortality [50]. The permissive vs non-permissive classification was developed based on cross-reactivity patterns seen in alloreactive T cells.

Recently, a 2015 study identified an HLA-DPB1 variant in the regulatory region that is associated with increased expression of HLA-DPB1, which correlated with an increased risk of GVHD when transplanted into a mismatched recipient [21]. Additionally, it was found that HLA-DPB1 genotyping accurately predicted the expression variant enabling potential use of HLA-DPB1 expression in permissive and nonpermissive mismatching Schemes [23].

A 2013 study multiple mismatches at low-expression loci (LEL) (HLA-DRB3/4/5, -DQB1, -DPB1) are influential in some clinical scenarios [17]. Mismatches at these loci were not significant in 8/8 matched transplants due to linkage disequilibrium. Consistent with linkage disequilibrium, LEL mismatches could be predicted based on HLA-DRB1 matching status. Additionally, three or more mismatches at the LEL were associated with increased risks of mortality and transplant-related mortality.

4 Outlook for HLA Typing by NGS

Since around 2015, clinical HLA laboratories have been using next-generation sequencing (NGS) technology to provide HLA typing for patients being evaluated for hematopoietic cell transplantation and their potential donors. The adoption of NGS within clinical HLA laboratories has been slowed by the relatively high initial cost for instrumentation ($45–90,000) and validation, integration with existing laboratory information systems (LIS), and concern for technical issues associated with NGS. The majority of the commercial vendors have absorbed most of the direct

costs associated with validation of the indirect cost for NGS validation is also high and more difficult to offset. The difficulties of validation are lost when you evaluate the gains in workflow, laboratory costs, and improvement in turn-around-time for reporting results [51–53].

The laboratories that support hematopoietic cell and solid-organ transplantation programs have used NGS to provide HLA typing for both programs. Use of NGS for this program is for different reasons. The high-throughput nature of NGS enables more efficient high-resolution HLA typing of up to 96 patients full HLA typing (HLA-A, -B, -C, -DRB1, -DQB1, −DQA1, -DPB1, -DPA1) at one time with very few ambiguities (~0.3%) [52, 54–56]. While the clinical utility of HLA-DQA1 and HLA-DPA1 in HLA matching has yet to be determined some HLA laboratories retain this information for future analysis. In contrast, using NGS for HLA typing to support solid-organ transplant programs enables identification of allele-level resolution of recipient and donors that can be beneficial for interpretation of donor-specific antibodies post-transplant, which are identified at the allele level.

There are several important issues that need to be addressed to improve our understanding of HLA genotypes and function in transplantation. First, laboratories lack the ability to phase entire HLA class I or II haplotypes. Second, HLA expression is clinically poorly utilized and undervalued. Third, laboratories are incapable of generating high-resolution HLA genotypes in time for decease donor organ allocation. Addressing these issues has the potential to greatly improve not only our understanding of how the HLA loci interact with each other but also improve recipient outcomes by enable more efficient HLA matching.

One way of addressing the first issue is implementation of single molecule sequencing. At present, there are two companies producing single molecule sequencers: Pacific Biosciences and Oxford Nanopore Technologies. Both of these technologies enable long-read (up to 100kbp reported) sequencing as well as rapid sequencing [57]. There have been relatively few studies examining HLA genotyping using single molecule sequencing [58–60]. Several large commercial laboratories have adopted Pacific Biosciences sequencers as their primary sequencing method, however, due to the cost of the instrumentation most academic laboratories have yet to adopt such technology.

In addition to single molecule sequencing, RNA sequencing (RNASeq) has numerous advantages over the current DNA-based approaches. RNASeq data provides accurate HLA genotyping [15, 16], higher throughput, entire transcriptome information, and relative HLA expression, which recent reports suggest effects patient outcome [20, 21, 42]. The ability of RNASeq data to provide information outside of HLA such as KIR, MICA, or MICB

typing enables a single sequencing assay to address additional biomarkers as our knowledge advances. The clear limitation of RNASeq data is the lack of intron information and the lack of microRNA. MicroRNA can differentially regulate HLA expression in pregnancy, cancer, and autoimmune diseases [19, 24, 61].

The next major shift in HLA typing by NGS may be the transition to high-resolution HLA typing for decease donors. Traditionally, this population has been HLA typed at low resolution due to the time necessary to achieve high-resolution HLA typing. With the emergence of nanopore technology real-time, rapid sequencing is now possible [59, 62–69]. As discussed above, nanopore sequencing can provide long-read genomic data capable of sequencing an entire HLA haplotype. Additionally, MinION sequences nucleotides (nt) at a faster rate (1400 nt/min) compared with Ion Torrent (1 nt/min) and MiSeq (0.17 nt/min) [62, 70, 71]. Lastly, nanopore sequencers are the generation of real-time sequencing data such that sequencing can be stopped once sufficient data is obtained further reducing the time necessary for hrHLA genotyping and gathered anywhere with the mobile sequencers (SmidgION).

The future of HLA typing by NGS is one that includes expansion of immunogenetics beyond the traditional HLA region to other regions, includes ability to phase entire HLA haplotypes, and utilizes an assay that is user-friendly and mobile. As always laboratories will be faced with many challenges developing these processes and always keep high-quality patient care at the forefront.

References

1. Merrill JP, Murray JE, Harrison JH, Guild WR (1956) Successful homotransplantation of the human kidney between identical twins. J Am Med Assoc 160(4):277–282
2. Barnard CN (1967) The operation. A human cardiac transplant: an interim report of a successful operation performed at Groote Schuur hospital, cape town. S Afr Med J 41(48):1271–1274
3. Hardy JD, Webb WR, Dalton ML Jr, Walker GR Jr (1963) Lung homotransplantation in man. JAMA 186:1065–1074
4. Lillehei RC, Idezuki Y, Kelly WD, Najarian JS, Merkel FK, Goetz FC (1969) Transplantation of the intestine and pancreas. Transplant Proc 1(1):230–238
5. Starzl TE, Brettschneider L, Penn I, Bell P, Groth CG, Blanchard H, Kashiwagi N, Putnam CW (1969) Orthotopic liver transplantation in man. Transplant Proc 1(1):216–222
6. Phillips RA, Cowan DH (1972) Human bone marrow transplantation. Med Clin North Am 56(2):433–451
7. Dausset J, Nenna A (1952) Presence of leuko-agglutinin in the serum of a case of chronic agranulocytosis. C R Seances Soc Biol Fil 146(19–20):1539–1541
8. Dyer PA, Claas FH (1997) A future for HLA matching in clinical transplantation. European journal of immunogenetics: official journal of the British society for histocompatibility and. Immunogenetics 24(1):17–28
9. Erlich H (2012) HLA DNA typing: past, present, and future. Tissue Antigens 80(1):1–11. https://doi.org/10.1111/j.1399-0039.2012.01881.x
10. Cereb N, Kim HR, Ryu J, Yang SY (2015) Advances in DNA sequencing technologies for high resolution HLA typing. Hum Immunol 76(12):923–927. https://doi.org/10.1016/j.humimm.2015.09.015

11. Olerup O, Zetterquist H (1992) HLA-DR typing by PCR amplification with sequence-specific primers (PCR-SSP) in 2 hours: an alternative to serological DR typing in clinical practice including donor-recipient matching in cadaveric transplantation. Tissue Antigens 39:225–235. https://doi.org/10.1111/j.1399-0039.1992.tb01940.x
12. Speiser DE, Tiercy JM, Rufer N, Grundschober C, Gratwohl A, Chapuis B, Helg C, Loliger CC, Siren MK, Roosnek E, Jeannet M (1996) High resolution HLA matching associated with decreased mortality after unrelated bone marrow transplantation. Blood 87(10):4455–4462
13. Middleton D (1999) History of DNA typing for the human MHC. Rev Immunogenet 1(2):135–156
14. Lawlor DA, Zemmour J, Ennis PD, Parham P (1990) Evolution of class-I MHC genes and proteins: from natural selection to thymic selection. Annu Rev Immunol 8:23–63. https://doi.org/10.1146/annurev.iy.08.040190.000323
15. Boegel S, Lower M, Schafer M, Bukur T, de Graaf J, Boisguerin V, Tureci O, Diken M, Castle JC, Sahin U (2012) HLA typing from RNA-Seq sequence reads. Genome Med 4(12):102. https://doi.org/10.1186/gm403
16. Buchkovich ML, Brown CC, Robasky K, Chai S, Westfall S, Vincent BG, Weimer ET, Powers JG (2017) HLAProfiler utilizes k-mer profiles to improve HLA calling accuracy for rare and common alleles in RNA-seq data. Genome Med 9(1):86. https://doi.org/10.1186/s13073-017-0473-6
17. Fernandez-Vina MA, Klein JP, Haagenson M, Spellman SR, Anasetti C, Noreen H, Baxter-Lowe LA, Cano P, Flomenberg N, Confer DL, Horowitz MM, Oudshoorn M, Petersdorf EW, Setterholm M, Champlin R, Lee SJ, de Lima M (2013) Multiple mismatches at the low expression HLA loci DP, DQ, and DRB3/4/5 associate with adverse outcomes in hematopoietic stem cell transplantation. Blood 121(22):4603–4610. https://doi.org/10.1182/blood-2013-02-481945
18. Greene JM, Wiseman RW, Lank SM, Bimber BN, Karl JA, Burwitz BJ, Lhost JJ, Hawkins OE, Kunstman KJ, Broman KW, Wolinsky SM, Hildebrand WH, O'Connor DH (2011) Differential MHC class I expression in distinct leukocyte subsets. BMC Immunol 12:39. https://doi.org/10.1186/1471-2172-12-39
19. Kaur G, Gras S, Mobbs JI, Vivian JP, Cortes A, Barber T, Kuttikkatte SB, Jensen LT, Attfield KE, Dendrou CA, Carrington M, McVean G, Purcell AW, Rossjohn J, Fugger L (2017) Structural and regulatory diversity shape HLA-C protein expression levels. Nat Commun 8:15924. https://doi.org/10.1038/ncomms15924
20. Kulkarni S, Savan R, Qi Y, Gao X, Yuki Y, Bass SE, Martin MP, Hunt P, Deeks SG, Telenti A, Pereyra F, Goldstein D, Wolinsky S, Walker B, Young HA, Carrington M (2011) Differential microRNA regulation of HLA-C expression and its association with HIV control. Nature 472(7344):495–498. https://doi.org/10.1038/nature09914
21. Petersdorf EW, Malkki M, O'HUigin C, Carrington M, Gooley T, Haagenson MD, Horowitz MM, Spellman SR, Wang T, Stevenson P (2015) High HLA-DP expression and graft-versus-host disease. N Engl J Med 373(7):599–609. https://doi.org/10.1056/NEJMoa1500140
22. Rene C, Lozano C, Eliaou JF (2016) Expression of classical HLA class I molecules: regulation and clinical impacts: Julia Bodmer award review 2015. HLA 87(5):338–349. https://doi.org/10.1111/tan.12787
23. Schone B, Bergmann S, Lang K, Wagner I, Schmidt AH, Petersdorf EW, Lange V (2018) Predicting an HLA-DPB1 expression marker based on standard DPB1 genotyping: linkage analysis of over 32,000 samples. Hum Immunol 79(1):20–27. https://doi.org/10.1016/j.humimm.2017.11.001
24. Seliger B (2017) Immune modulatory microRNAs as a novel mechanism to revert immune escape of tumors. Cytokine Growth Factor Rev 36:49–56. https://doi.org/10.1016/j.cytogfr.2017.07.001
25. Sellares J, Reeve J, Loupy A, Mengel M, Sis B, Skene A, de Freitas DG, Kreepala C, Hidalgo LG, Famulski KS, Halloran PF (2013) Molecular diagnosis of antibody-mediated rejection in human kidney transplants. Am J Transplant 13(4):971–983. https://doi.org/10.1111/ajt.12150
26. Cecka JM, Reed EF, Zachary AA (2015) HLA high-resolution typing for sensitized patients: a solution in search of a problem? Am J Transplant 15(4):855–856. https://doi.org/10.1111/ajt.13169
27. Montgomery RA, Leffell MS, Zachary AA (2013) Transplantation of the sensitized patient: histocompatibility testing. Methods Mol Biol 1034:117–125. https://doi.org/10.1007/978-1-62703-493-7_6
28. Zachary AA, Leffell MS (2016) HLA mismatching strategies for solid organ transplantation - a balancing act. Front Immunol 7:575. https://doi.org/10.3389/fimmu.2016.00575
29. Kamoun M, Phelan D, Noreen H, Marcus N, Klingman L, Gebel HM (2017) HLA compat-

ibility assessment and management of highly sensitized patients under the new kidney allocation system (KAS): a 2016 status report from twelve HLA laboratories across the U.S. Hum Immunol 78(1):19–23. https://doi.org/10.1016/j.humimm.2016.10.023

30. Smith JM, Biggins SW, Haselby DG, Kim WR, Wedd J, Lamb K, Thompson B, Segev DL, Gustafson S, Kandaswamy R, Stock PG, Matas AJ, Samana CJ, Sleeman EF, Stewart D, Harper A, Edwards E, Snyder JJ, Kasiske BL, Israni AK (2012) Kidney, pancreas and liver allocation and distribution in the United States. Am J Transplant 12(12):3191–3212. https://doi.org/10.1111/j.1600-6143.2012.04259.x
31. Stewart DE, Garcia VC, Aeder MI, Klassen DK (2017) New insights into the alleged kidney donor profile index labeling effect on kidney utilization. Am J Transplant 17(10):2696–2704. https://doi.org/10.1111/ajt.14379
32. Stewart DE, Klassen DK (2016) Kidney transplants from HLA-incompatible live donors and survival. N Engl J Med 375(3):287–288. https://doi.org/10.1056/NEJMc1604523#SA2
33. Stewart DE, Klassen DK (2017) Early experience with the new kidney allocation system: a perspective from UNOS. Clin J Am Soc Nephrol 12(12):2063–2065. https://doi.org/10.2215/cjn.06380617
34. Stewart DE, Kucheryavaya AY, Klassen DK, Turgeon NA, Formica RN, Aeder MI (2016) Changes in deceased donor kidney transplantation one year after KAS implementation. Am J Transplant 16(6):1834–1847. https://doi.org/10.1111/ajt.13770
35. Grgic I, Chandraker A (2017) Significance of biologics in renal transplantation: past, present, and future. Curr Opin Organ Transplant 23(1):51–62. https://doi.org/10.1097/mot.0000000000000496
36. Williams RC, Opelz G, McGarvey CJ, Weil EJ, Chakkera HA (2016) The risk of transplant failure with HLA mismatch in first adult kidney allografts from deceased donors. Transplantation 100(5):1094–1102. https://doi.org/10.1097/TP.0000000000001115
37. Tinckam KJ, Liwski R, Pochinco D, Mousseau M, Grattan A, Nickerson P, Campbell P (2015) cPRA increases with DQA, DPA, and DPB unacceptable antigens in the Canadian cPRA calculator. Am J Transplant 15(12):3194–3201. https://doi.org/10.1111/ajt.13355
38. Tinckam KJ, Rose C, Hariharan S, Gill J (2016) Re-examining risk of repeated HLA mismatch in kidney transplantation. J Am Soc Nephrol 27(9):2833–2841. https://doi.org/10.1681/ASN.2015060626
39. Roberts JP, Wolfe RA, Bragg-Gresham JL, Rush SH, Wynn JJ, Distant DA, Ashby VB, Held PJ, Port FK (2004) Effect of changing the priority for HLA matching on the rates and outcomes of kidney transplantation in minority groups. N Engl J Med 350(6):545–551. https://doi.org/10.1056/NEJMoa025056
40. Tinckam KJ (2012) Basic histocompatibility testing methods. pp 21–42. doi:https://doi.org/10.1007/978-1-4614-0008-0_2
41. Eng HS, Leffell MS (2011) Histocompatibility testing after fifty years of transplantation. J Immunol Methods 369(1–2):1–21. https://doi.org/10.1016/j.jim.2011.04.005
42. Crivello P, Zito L, Sizzano F, Zino E, Maiers M, Mulder A, Toffalori C, Naldini L, Ciceri F, Vago L, Fleischhauer K (2015) The impact of amino acid variability on alloreactivity defines a functional distance predictive of permissive HLA-DPB1 mismatches in hematopoietic stem cell transplantation. Biol Blood Marrow Transplant 21(2):233–241. https://doi.org/10.1016/j.bbmt.2014.10.017
43. Pidala J, Wang T, Haagenson M, Spellman SR, Askar M, Battiwalla M, Baxter-Lowe LA, Bitan M, Fernandez-Vina M, Gandhi M, Jakubowski AA, Maiers M, Marino SR, Marsh SG, Oudshoorn M, Palmer J, Prasad VK, Reddy V, Ringden O, Saber W, Santarone S, Schultz KR, Setterholm M, Trachtenberg E, Turner EV, Woolfrey AE, Lee SJ, Anasetti C (2013) Amino acid substitution at peptide-binding pockets of HLA class I molecules increases risk of severe acute GVHD and mortality. Blood 122(22):3651–3658. https://doi.org/10.1182/blood-2013-05-501510
44. Zino E, Frumento G, Marktel S, Sormani MP, Ficara F, Di Terlizzi S, Parodi AM, Sergeant R, Martinetti M, Bontadini A, Bonifazi F, Lisini D, Mazzi B, Rossini S, Servida P, Ciceri F, Bonini C, Lanino E, Bandini G, Locatelli F, Apperley J, Bacigalupo A, Ferrara GB, Bordignon C, Fleischhauer K (2004) A T-cell epitope encoded by a subset of HLA-DPB1 alleles determines nonpermissive mismatches for hematologic stem cell transplantation. Blood 103(4):1417–1424. https://doi.org/10.1182/blood-2003-04-1279
45. Zino E, Vago L, Di Terlizzi S, Mazzi B, Zito L, Sironi E, Rossini S, Bonini C, Ciceri F, Roncarolo MG, Bordignon C, Fleischhauer K (2007) Frequency and targeted detection of HLA-DPB1 T cell epitope disparities relevant in unrelated hematopoietic stem cell transplantation. Biol Blood Marrow Transplant 13(9):1031–1040. https://doi.org/10.1016/j.bbmt.2007.05.010

46. Gragert L, Eapen M, Williams E, Freeman J, Spellman S, Baitty R, Hartzman R, Rizzo JD, Horowitz M, Confer D, Maiers M (2014) HLA match likelihoods for hematopoietic stem-cell grafts in the U.S. registry. N Engl J Med 371(4):339–348. https://doi.org/10.1056/NEJMsa1311707
47. Tie R, Zhang T, Yang B, Fu H, Han B, Yu J, Tan Y, Huang H (2017) Clinical implications of HLA locus mismatching in unrelated donor hematopoietic cell transplantation: a meta-analysis. Oncotarget 8(16):27645–27660. https://doi.org/10.18632/oncotarget.15291
48. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, Fernandez-Vina M, Flomenberg N, Horowitz M, Hurley CK, Noreen H, Oudshoorn M, Petersdorf E, Setterholm M, Spellman S, Weisdorf D, Williams TM, Anasetti C (2007) High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood 110(13):4576–4583. https://doi.org/10.1182/blood-2007-06-097386
49. Touzeau C, Gagne K, Sebille V, Herry P, Chevallier P, Follea G, Devys A, Moreau P, Mohty M, Cesbron Gautier A (2012) Investigation of the impact of HLA-DPB1 matching status in 10/10 HLA matched unrelated hematopoietic stem cell transplantation: results of a French single center study. Hum Immunol 73(7):711–714. https://doi.org/10.1016/j.humimm.2012.03.013
50. Crocchiolo R, Zino E, Vago L, Oneto R, Bruno B, Pollichieni S, Sacchi N, Sormani MP, Marcon J, Lamparelli T, Fanin R, Garbarino L, Miotti V, Bandini G, Bosi A, Ciceri F, Bacigalupo A, Fleischhauer K (2009) Nonpermissive HLA-DPB1 disparity is a significant independent risk factor for mortality after unrelated hematopoietic stem cell transplantation. Blood 114(7):1437–1444. https://doi.org/10.1182/blood-2009-01-200378
51. Weimer ET (2016) Clinical validation of NGS technology for HLA: an early adopter's perspective. Hum Immunol 77(10):820–823. https://doi.org/10.1016/j.humimm.2016.06.014
52. Weimer ET, Montgomery M, Petraroia R, Crawford J, Schmitz JL (2016) Performance characteristics and validation of next-generation sequencing for human leucocyte antigen typing. J Mol Diagn 18(5):668–675. https://doi.org/10.1016/j.jmoldx.2016.03.009
53. Lange V, Bohme I, Hofmann J, Lang K, Sauter J, Schone B, Paul P, Albrecht V, Andreas JM, Baier DM, Nething J, Ehninger U, Schwarzelt C, Pingel J, Ehninger G, Schmidt AH (2014) Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics 15:63. https://doi.org/10.1186/1471-2164-15-63
54. Duke JL, Lind C, Mackiewicz K, Ferriola D, Papazoglou A, Gasiewski A, Heron S, Huynh A, McLaughlin L, Rogers M, Slavich L, Walker R, Monos DS (2016) Determining performance characteristics of an NGS-based HLA typing method for clinical applications. HLA 87(3):141–152. https://doi.org/10.1111/tan.12736
55. Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, Slavich L, Walker R, Hsiao T, McLaughlin L, D'Arcy M, Gai X, Goodridge D, Sayer D, Monos D (2010) Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum Immunol 71(10):1033–1042. https://doi.org/10.1016/j.humimm.2010.06.016
56. Profaizer T, Lazar-Molnar E, Pole A, Delgado JC, Kumanovics A (2017) HLA genotyping using the Illumina HLA TruSight next-generation sequencing kits: a comparison. Int J Immunogenet 44(4):164–168. https://doi.org/10.1111/iji.12322
57. Boza V, Brejova B, Vinar T (2017) DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads. PLoS One 12(6):e0178751. https://doi.org/10.1371/journal.pone.0178751
58. Laver TW, Caswell RC, Moore KA, Poschmann J, Johnson MB, Owens MM, Ellard S, Paszkiewicz KH, Weedon MN (2016) Pitfalls of haplotype phasing from amplicon-based long-read sequencing. Sci Rep 6:21746. https://doi.org/10.1038/srep21746
59. Ammar R, Paton TA, Torti D, Shlien A, Bader GD (2015) Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. F1000Res 4:17. https://doi.org/10.12688/f1000research.6037.1
60. Albrecht V, Zweiniger C, Surendranath V, Lang K, Schofl G, Dahl A, Winkler S, Lange V, Bohme I, Schmidt AH (2017) Dual redundant sequencing strategy: full-length gene characterisation of 1056 novel and confirmatory HLA alleles. Hla 90(2):79–87. https://doi.org/10.1111/tan.13057
61. Chen Q, Luo G, Zhang X (2017) MiR-148a modulates HLA-G expression and influences tumor apoptosis in esophageal squamous cell carcinoma. Exp Ther Med 14(5):4448–4452. https://doi.org/10.3892/etm.2017.5058
62. Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M (2015) Improved data analysis for the MinION nanopore sequencer. Nat Methods 12(4):351–356. https://doi.org/10.1038/nmeth.3290

63. Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17(1):239. https://doi.org/10.1186/s13059-016-1103-0
64. Karlsson E, Larkeryd A, Sjodin A, Forsman M, Stenberg P (2015) Scaffolding of a bacterial genome using MinION nanopore sequencing. Sci Rep 5:11996. https://doi.org/10.1038/srep11996
65. Loose M, Malla S, Stout M (2016) Real-time selective sequencing using nanopore technology. Nat Methods 13(9):751–754. https://doi.org/10.1038/nmeth.3930
66. Manrao EA, Derrington IM, Laszlo AH, Langford KW, Hopper MK, Gillgren N, Pavlenok M, Niederweis M, Gundlach JH (2012) Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. Nat Biotechnol 30(4):349–353. https://doi.org/10.1038/nbt.2171
67. Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W (2016) Nanopore sequencing detects structural variants in cancer. Cancer Biol Ther 17(3):246–253. https://doi.org/10.1080/15384047.2016.1139236
68. Bolisetty MT, Rajadinakaran G, Graveley BR (2015) Determining exon connectivity in complex mRNAs by nanopore sequencing. Genome Biol 16:204. https://doi.org/10.1186/s13059-015-0777-z
69. Branton D, Daniel B, Deamer DW, Andre M, Hagan B, Benner SA (2008) The potential and challenges of nanopore sequencing. Nat Biotechnol 26(10):1146–1153. https://doi.org/10.1038/nbt.1495
70. Wei S, Williams Z (2016) Rapid short-read sequencing and aneuploidy detection using MinION Nanopore technology. Genetics 202(1):37–44. https://doi.org/10.1534/genetics.115.182311
71. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4(4):265–270. https://doi.org/10.1038/nnano.2009.12

Chapter 2

Role of Human Leukocyte Antigens (HLA) in Autoimmune Diseases

Gergely Bodis, Victoria Toth, and Andreas Schwarting

Abstract

The aim of this review is to provide a brief overview of the role and current clinical relevance of HLA-B27 in spondyloarthritis and HLA-B51 in Behcet's disease as well as HLA-DQ2/DQ8 in celiac disease and HLA-DRB1 in rheumatoid arthritis and to discuss possible future implications.

Key words Spondyloarthritis, Behcet's disease, Rheumatoid arthritis, Autoimmunity, HLA-B27, HLA-B51, HLA-DQ2/DQ8m HLA-DRB1

1 Introduction

The human leukocyte antigen (HLA) system, which corresponds to the major histocompatibility complex (MHC) in humans, plays a pivotal role in the antigen presentation of intracellular and extracellular peptides and the regulation of innate and adaptive immune responses.

Since the discovery of HLA 60 years ago it has contributed to the understanding of the immune system as well as of the pathogenesis of several diseases. Aside from its essential role in determining donor-recipient immune compatibility in organ transplantation, HLA genotyping is meanwhile performed routinely as part of the diagnostic work-up of certain autoimmune diseases. Considering the ability of HLA to influence thymic selection as well as peripheral anergy of T cells, its role in the pathogenesis of autoimmunity is understandable.

2 MHC Class I: HLA-B27 and Spondyloarthritis (SpA)

The clinical entities of the spondyloarthritis (SpA) group are inflammatory diseases with distinctive axial and/or peripheral joint involvement, enthesitis and frequently accompanied by inflammatory eye disease, especially anterior uveitis. SpA belongs

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_2,

to the most common rheumatic diseases with a prevalence of 0.4–1.3% in the US with similar prevalence in Europe and lower rates in African and Asian populations [1, 2]. It also poses a major burden both socially and economically due to work disability occurring in 18.5–21% of SpA patients [3, 4].

One of the common denominators among distinct entities of the SpA family is the frequent association with MHC Class I molecules, particularly with HLA-B27, which has obtained significance in the routine diagnostic work-up in the last decades. HLA-B27 belongs to the MHC class I molecules, its main function is the presentation of intracellular peptides to CD8 positive T Lymphocytes. These MHC class I-restricted T cells possess cytotoxic or regulatory function, their activation leading accordingly either to tolerance, when the presented peptides are recognized as "self," to activation of cell-mediated immunity in case "non-self" antigens are presented or to a maladaptive autoimmune response if the "self" antigen is misrecognized.

The prevalence of HLA-B27 shows a pronounced north-south gradient in the normal population: it is lowest in the equatorial region (~0%) and highest in northern countries (30–40%). This geographical difference may be attributable to HLA-B27 carriers being more susceptible to malaria and also showing a more severe disease course. This susceptibility might have led to the negative selection of HLA-B27 positive individuals in areas endemic for malaria [5]. The population of Papua New Guinea and Eskimos seem to have the highest prevalence of HLA-B27, with 13–53% [6, 7] and 25–50% [8–10], respectively. Among Caucasians the prevalence is 6–10%. The prevalence is lower in Chinese (2–8%) [11, 12], Arab (2–5%) [13], African-American (2–4%) [14], and Japanese (0.4%) [15] populations. In the natives of South America, equatorial and southern Africa, and the Aboriginal people of Australia HLA-B27 is virtually absent [16, 17]. The prevalence of SpA corresponds to the distribution of HLA-B27 alleles in various populations.

A potential genetic susceptibility to ankylosing spondylitis (AS) was first recognized in 1950 [18] and the strong link to HLA-B27 was discovered in 1973 by two research groups [19, 20]. AS is a radiographic axial SpA primarily with spinal and sacroiliac joint involvement, which is characterized by enthesitis with chronic inflammation, that subsequently results in fibrosis and ossification of the involved sites. HLA-B27 has the strongest association with AS among disease entities of the SpA group, especially in Caucasians with 88–96% of patients being positive [19, 20]. Asian AS patients carry HLA-B27 less frequently. Among African Americans, HLA-B27 is present in 50% of patients with AS [21]. 30–80% of patients with reactive arthritis (ReA) and 20–35% of patients with psoriatic arthritis are HLA-B27 positive [22, 23].

Ethnic differences seem to exist regarding disease susceptibility conferred by HLA-B27. The relative risk of developing SpA in HLA-B27 positive individuals is increased 20–100-fold in the Causasian population [24–26], however, a study carried out on the Indonesian population found no increased relative risk among HLA-B27 positive subjects [27]. 10–30% of HLA-B27 positive first-degree relatives of HLA-B27 positive AS patients also develop the disease [28].

The risk of HLA-B27 positive individuals to develop ReA is 5–10 times greater than that of the general population [29]. The role of gene dosage is inconclusive. On one hand, higher relative risk of developing AS in HLA-B27 homozygotes was observed in Finnish AS patients. Interestingly, homozygous patients showed a less severe disease course [30]. On the other hand, an earlier study carried out in the Netherlands and a more recent study in Korea found no significant difference between homozygous and heterozygous patients [31, 32].

Similar to other MHC class I molecules, HLA-B27 is a heterotrimer derived from a heavy chain encoded by the HLA genes, a β2-microglobulin light chain, and the presented peptide. HLA-B27 differs from other HLA-B-molecules on possessing a free cysteine at residue 67 (Cys67) allowing the molecule to create stable homodimers without β2-microglobulin due to the formation of disulfide bonds [33]. HLA-B27 shows a marked genotypic and phenotypic polymorphism with at least 132 alleles and 105 subtypes. Non-synonymous nucleotide substitutions affecting the antigen-binding cleft can lead to differences in antigen presentation and ultimately in disease association [34].

Furthermore, the frequency of HLA-B27 subtypes varies among different ethnic groups as well. One of the benefits of genetic HLA-B27 testing is that the different subtypes can be determined more reliably and reproducibly, all of which have different levels of associations with disease [35]. The B*27:05 subtype is the most common in Caucasians, other subtypes evolved from this ancestral type by gene conversion, reciprocal combination, and point mutation. It is however probably not linked to SpA in the African population. B*27:02 shows a strong association in the Mediterranean population, while B*27:04 is a common subtype in Asian SpA patients [36]. A recent meta-analysis composed of 8993 AS patients and 19,254 healthy controls confirmed the significant association of B*27:02 and B*27:04 with AS. B*27:03, *27:06, and *27:09 are considered to be protective subtypes, although SpA cases in patients carrying these subtypes have been reported. B*27:03 and B*27:06 are common in Southeast Asia, while *27:09 is frequently found in Sardinia and Italy. Some rare subtypes also seem to contribute to the risk of SpA, including B*27:01, B*27:07, B*27:08, B*27:10, B*27:13, B*27:14, B*27:15, B*27:19, and B*27:25. Changes of the primary structure of the

HLA-B27 protein may explain the different levels of disease associations, especially the variations affecting the antigen-binding cleft and as a consequence possibly the peptide specificity. There has been a difference of two amino acids observed between the F pockets of the peptide-binding groove of B*27:06 and B*27:04, as well as a difference of one amino acid between B*27:09 and B*27:05. These minor changes of the amino acid sequences result in significantly different risk profiles [37–41].

Despite intensive research in the last decades the pathomechanism of SpA and the contribution of HLA-B27 to disease still remains unclear as neither one of the existing hypotheses can fully describe and explain the underlying mechanisms.

The unfolded protein response hypothesis is based on the ability of HLA-B27 heavy chains to form stable homodimers owing to the free thiole groups of Cys67 [33]. These complexes are retained in the endoplasmic reticulum in the absence of β2-microglobulin, misfold, and accumulate in the endoplasmic reticulum leading to stress response and inflammation [42]. The protective B*27:06 and B*27:09 subtypes are less prone to misfold than subtypes associated with disease risk. While this observation might seem to support this hypothesis, it is undermined by the fact that the disease-associated B*27:07 subtype has been shown to fold equally efficiently [43].

β2-microglobulin-free HLA-B27 heavy chain homodimers can also be found on the cell membrane, where they can also interact with CD4 positive T lymphocytes, NK cells, and myelomonocytic cells that express killer-immunglobulin-like receptors (KIR) and leukocyte immunglobulin-like receptors (LILR). β2-microglobulin can also be released from HLA-B27 molecules on the cell surface and be deposited in synovial tissue suggesting a possible role in the pathogenesis of SpA [44].

The molecular mimicry and arthritogenic peptides hypothesis proposes that owing to the properties of the antigen-binding cleft HLA-B27 can present certain microbial peptides similar to self-antigens. The immune response triggered by the displayed microbial peptides causes HLA-B27-restricted CD8 positive T-Lymphocytes to cross-react with these arthritogenic peptides triggering chronic inflammation [45]. Indeed, several such microbial peptides have been identified. The nitrogenase enzyme of *Klebsiella pneumoniae* shares a sequence of six consecutive amino acids with HLA-B27 [46]. Another *K. pneumoniae* protein, the pullulanase enzyme, and certain outer surface proteins of Yersinia enterocolitica and pseudotuberculosis, Shigella flexneri and Salmonella typhimurium also possess sequences homologous with HLA-B27 [47, 48]. A recent study aiming to identify such arthritogenic peptides assessed the peptide repertoire of eight frequent HLA-B27 subtypes (HLA-B*27:02–09). They identified

more than 7500 endogenous peptides presented by these B27 subtypes. However, most peptides that are presented by the risk-subtypes could also bind to B*27:06 and B*27:09, which are considered to be protective. This significant overlap of presented peptides between the subtypes leads the authors to the conclusion that the different risk profiles among subtypes may be due to quantitative changes affecting antigen sensitivity of autoreactive T cells and most likely not to qualitative changes of the HLA-B27 peptide repertoire [49]. Additionally, several studies proposed a link between the interaction of HLA-B27 with the intestinal microbiome and the pathogenesis of related diseases. HLA-B27 may affect the composition of the gut flora. Indeed, dysbiotic changes have been described in patients with SpA. HLA-B27 transgenic rats had increased proportions of Prevotellaceae and loss of Rikenellaceae in the intestinal flora. If these HLA-B27 transgenic animals were kept in a germ-free environment, they did not develop arthritis. Recolonization of the gut with Bacteroides vulgatus resulted in inflammatory changes [50, 51]. However, introducing Lactobacillus and fusiform bacteria to the gut of germ-free animals had shown no such effect. These observations suggest that the modulation of the complex interplay between the immune system and microbiome can influence and in some cases prevent disease manifestation in this animal model. These changes might be connected with the unfolded protein response: the endoplasmic reticulum stress response could lead to intestinal inflammation, impaired barrier function, and loss of oral tolerance. The resulting increased translocation of microbial antigens could, on one hand, induce extraintestinal inflammation, on the other hand prime autoreactive T-Lymphocytes [52–54]. However, a significant role of microbial antigens or arthritogenic peptides in the pathogenesis of SpA has not been unequivocally demonstrated.

HLA-B27 determination has obtained clinical significance in the past decades in the routine diagnostic work-up of SpA due to its strong genetic association with disease. HLA-B27 determination has a sensitivity of 83–96%, specificity of 90–96%, and a likelihood ratio of 9.0 for AS in Caucasians with inflammatory back pain [55]. HLA-B27 positivity is part of the Assessment of Spondyloarthritis International Society (ASAS) classification criteria for axial and peripheral SpA as well as the Amor Criteria for the diagnosis of SpA [56]. Current German guidelines also recommend HLA-B27 determination in case of clinical suspicion of SpA. However, screening of the general population is not recommended, as a positive result merely indicates genetic susceptibility. Accordingly, only a minority of HLA-B27 carriers will develop a disease of the SpA spectrum. Generally, testing for HLA-B27 should not be repeated, although in case of serological typing cross-reactivity with other HLA-B molecules as well as false negative results were reported [57, 58].

The data provided by the Recognising and Diagnosing Ankylosing Spondylitis Reliably (RADAR) study has led to the development of a strategy for primary care physicians when to refer patients with early-onset (<45 years) chronic back pain to rheumatologic evaluation: the selection could be based on either HLA-B27 positivity, inflammatory back pain or sacroiliitis on MRI. The authors concluded that a referral strategy based on these three criteria can lead to the diagnosis of axial SpA in 35% of cases [59].

In addition to being a pivotal part of the diagnostic work-up a prognostic value has been attributed to HLA-B27 as well. In patients with AS HLA-B27 positivity is associated with earlier disease onset, higher disease activity, risk of peripheral joint involvement, symmetric sacroiliitis, severity of MRI findings in sacroiliitis, and positive family history [60–63], although there have been conflicting reports [64, 65]. Undiagnosed patients with early inflammatory back pain—especially in case of non-radiologic axial SpA—benefit from the combination of MRI of the sacroiliac joints and HLA-B27 determination. Severe sacroiliitis in HLA-B27 positive patients is highly specific for the development of AS. Patients with mild or no sacroiliitis on MRI have a low risk of developing AS regardless of the HLA-B27 status [66]. Higher NSAID use and higher need for biologicals has been observed in HLA-B27 positive patients with AS [67]. On the other hand, TNF-alpha inhibitors show a greater therapeutic effect in HLA-B27 positive patients with AS [68, 69].

Psoriatic arthritis (PsA), a further entity of the SpA group, is a multifaceted chronic inflammatory joint disease, which is associated with cutaneous psoriasis in the majority of patients. It generally manifests as an asymmetrical oligoarthritis, although polyarticular as well as axial forms also commonly occur. In patients with PsA HLA-B27 is associated with axial manifestation and possibly also with distal phalangeal joint involvement, and this association seems to be independent of psoriasis [70]. However, PsA is not associated with HLA-Cw6, which is present in 10–60% of patients with psoriasis [71], although earlier studies using serologic methods described a possible connection. Interestingly, 61% of PsA patients with symmetric sacroiliitis carried HLA-B*27:05, as opposed to 9.8% of patients with asymmetric sacroiliitis, where the haplotype HLA-B*08:01 – C*07:01 was more prevalent.

Patients possessing B*27:05 and especially the B*27:05-C*01:02 haplotype had a higher risk of dactylitis. The B*27:05 and C*01:02 alleles were associated with enthesitis in PsA.

Patients with a synovial-predominant pattern carried the B*08:01-C*07:01 haplotype more frequently, which predisposes to joint deformity. The B*27:05-C*02:02, B*37:01-C*06:02, and B*08:01-C*07:01 haplotypes are associated with a more

severe disease course [72]. Therefore, the different MHC class I alleles play a role in determining whether the patients with PsA develop asymmetrical or symmetrical sacroiliitis as well as enthesitis and dactylitis.

Uveitis is strongly linked to HLA-B27 as well. Conversely, the presence of an isolated HLA-B27 positive uveitis confers a high risk for developing SpA [73]. HLA B27 positivity correlates with worse prognosis and more severe disease course in ReA compared to B27 negative patients [74]. Therefore, the diagnostic role of HLA B27 is mostly to support the clinical suspicion of SpA in addition to providing prognostic information.

In addition to SpA, HLA-B27 has also been linked to other disease entities. Although rheumatoid arthritis is not associated with HLA-B27 in itself, an elevated risk of atlanto-axial subluxation has been described in carriers [75]. HLA-B27 is also considered to have a protective role in several viral diseases such as HIV, Hepatitis C, Influenza, Epstein-Barr virus, Herpes simplex virus, and Puumala Hantavirus infection, although it increases the risk of contracting malaria [5, 76–78]. Interestingly, a recent study reported the HLA-B27 molecule sharing a homology of four consecutive amino acids with an immunodominant peptide of E1 glycoprotein of Chikungunya virus. This leads the authors to the conclusion that HLA-B27 positivity might also play a role in persistent arthralgia following Chikungunya infection, its importance, however, remains to be seen [79].

Recent genome-wide associated studies performed in large groups of patients pinpointed the association of several non-B27 HLA as well as non-HLA genes with SpA [80, 81]. Nevertheless, routine genetic testing of non-B27 HLA as part of the diagnostic work-up of SpA is likely premature [82].

3 MHC Class I: HLA-B51 and Adamantiades-Behcet's Disease (BD)

Behcet's disease (BD) belongs to the group of variable vessel vasculitides according to the 2012 Revised International Chapel Hill Consensus Conference Nomenclature of Vasculitides, characterized by inflammatory eye disease (uveitis), oral and genital ulcers. Similarly to SpA, BD also shows a marked geographical distribution with Mediterranean and Asian populations being most affected, hence the name "silk road disease." The prevalence of BD is 17–42/10000 in Turkey, 2.1–420/100,000 in Asian and North African populations, and 0.3–7.5/100,000 in Western Europe and the United States [83–85].

Patients with BD often carry an MHC Class I molecule, HLA-B51 [86], especially those of Turkish or Asian origin,

whereas the association in Caucasian patients is weaker. In a recent meta-analysis, the prevalence of HLA-B51 in BD patients ranged between 50 and 72% [87] compared to 10–15% in healthy controls in high-risk populations.

Further genetic and environmental factors are likely to play an additional role in the pathogenesis of BD. A change of BD phenotype, in particular a decrease in HLA-B51 frequency, has recently been reported in Japanese patients, accordingly [88].

HLA-B51 is one of two distinct split-antigens of the HLA-B5 serotype and is primarily associated with BD risk, although some case-reports have also found a possible link with the second split-antigen, HLA-B52 [89, 90]. Several HLA-B51 subtypes have been described, of which especially B*51:01, B*51:02(01), B*51:08, B*51:09, and B*51:22 seem to be associated with BD risk [91, 92].

The role of HLA-B51 in the pathogenesis of BD is not fully understood. Selective binding of certain peptides and the activation of CD8 positive T-Lymphocytes and NK cells due to interactions of the HLA-B51-heterotrimer and T-cell receptors as well as killer immunoglobulin-like receptors are likely to be implicated. Gamma-delta T cells also play a role in the pathogenesis of BD. Furthermore, active BD was associated with significant in vivo activation of Vδ1 and Vδ2 gamma-delta T cells, while an over proportional activation of Vδ1 gamma-delta T cells has been seen exclusively in HLA-B51 positive patients [93]. Additionally, a possible role of HLA-B51 in neutrophil hyperfunction in BD has been described, the spontaneous activation of HLA-B51 positive neutrophils leads to perivascular tissue injury and promotes a Th1 immune response [94, 95]. These findings suggest that HLA-B51 is involved in the activation of CD8 positive T cells, gamma-delta T cells, NK cells, and neutrophils.

As for the clinical significance, HLA typing is not part of the International Criteria for Behçet disease, however, the testing is available in several medical laboratories. The estimated sensitivity is 51% and the specificity 71% [96]. It is important to understand the limitations and the diagnostic conclusiveness of both positive and negative results. It is not meant to diagnose BD, but rather support the diagnosis. The screening of high-risk populations is not recommended, as the majority of HLA-B51 carriers do not develop BD. Conversely, BD cannot be excluded in the absence of HLA-B51.

A possible prognostic value has been attributed to HLA-B51: HLA-B51 carriers have been shown to have a higher risk of genital ulcers as well as ocular or skin involvement. Male patients are more likely to be HLA-B51 positive [87]. Certain subtypes may be associated with different risk profiles, for example Turkish HLA-B*51:03 positive patients are at a higher risk of neurological

involvement, and HLA-B*51:09 may lower the risk of developing papulopustular lesions [92].

The testing of additional BD susceptibility HLA-alleles, such as HLA-A03, A26, B15, 27, 57 [97] [98], is currently not considered to be clinically or diagnostically relevant due to their low specificity.

4 MHC Class II: HLA-DQ2/DQ8 and Celiac Disease (CD)

Celiac disease (CD) is one of the most common organ-specific autoimmune diseases with a prevalence of 1% that primarily affects the small intestines following gluten exposure [99]. It has strong links to both genetic and environmental factors, latter being gluten exposure. As for the genetic factors, a strong association exists with MHC class II alleles, HLA-DQ2 and DQ8. A link with non-MHC genes has also been described in genome-wide association studies [100, 101].

90% of Caucasian patients with CD express HLA-DQ2.5cis encoded by HLA-DQA1*0501-DQB1*0201 or DQ2.5trans on HLA-DQA1*0505 DQB1*0301/DQA1*0201-DQB1*0202 haplotypes. 5% of these patients carry HLA-DQ8 with the HLA-DQA1*0301-DQB1*0302 haplotype. Patients negative for these HLA molecules mostly possess HLA-DQA1*0201-DQB1*0202 haplotypes (HLA-DQ2.2) [102]. HLA-DQ2.2 carriers have an inconsequential risk of developing CD. By contrast, approximately 20–30% of the healthy Caucasian population are HLA-DQ2 positive. A gender-specific distribution of HLA alleles has been described: female patients with celiac disease are infrequently DQ2.5/DQ8 negative. However, not all DQ2 carriers develop CD. Non-HLA genes are likely to play an additional role, as seen in identical twins, who show a higher concordance rate (70%) than HLA identical siblings. HLA-DQ2 or –DQ8 are necessary, but not sufficient for the development of CD. The estimated risk of DQ2 / DQ8 carriers is 36–53% [103].

Zygosity is a strong determinant of gluten peptide presentation and disease risk. Additionally, homozygous patients have been shown to have a more severe disease course. Especially HLA-DQ2.5 homozygotes may exhibit an augmented immune response following infections of the gastrointestinal tract due to elevated interferon gamma concentrations, which has been shown to regulate HLA-DQ expression indirectly [104, 105].

Different haplotypes recognize different ligands with different affinity, resulting in different risk profiles. A comparison of DQ2.5 ligands with DQ2.2 ligands revealed that DQ2.5 can present a broader spectrum of gliadin peptides than DQ2.2. Gliadin peptides can withstand gastrointestinal digestion; therefore, DQ2 molecules can recognize these resistant immunodominant epitopes

and present them to CD4 positive T-lymphocytes in the intestinal mucosa. The affinity of DQ2 to gliadin peptides is further increased by the tissue transglutaminase enzyme, which deamidates glutamine residues of gliadin peptides. The resulting glutamic acid residues display an increased affinity to DQ2 molecules [106, 107]. Activation of both the innate and adaptive immune system leads to a humoral response against tissue transglutaminase, as well as to TNF alpha and interferon gamma secretion, which ultimately result in tissue damage and disease manifestations. Additionally, possible links with infections have been described. Due to a homology of an amino acid sequence between the 54 kDa E1b protein of human adenovirus type 12 and gliadin, exposure to the virus may promote autoimmunity in genetically susceptible individuals [108]. Hepatitis C, Giardia lamblia, Campylobacter jejuni, Rotavirus, and Enterovirus have also been implicated as possible triggers of CD. It has been postulated that HLA-DQ molecules may also have an impact on the intestinal microbiome. Patients with celiac disease have a different composition of the intestinal microbiome with decreased proportion of Actinobacteria (especially Bifidobacterium genus) and elevated proportion of Firmicutes, Proteobacteria, and Staphylococcus spp. [109, 110]. There is a permanent interaction between microbes and Th17, Treg and B-lymphocytes. HLA-DQ molecules are likely to influence this interaction depending on the displayed ligands and lead to either tolerance of certain microbial strains or immune response against them. It is therefore possible that the effect of HLA molecules on the pathogenesis of celiac disease may be, to some extent, due to the altered microbiome.

Regarding the clinical and diagnostic relevance, the testing of DQ2/DQ8 has an excellent negative predictive value of 99%: a negative test virtually excludes CD [111], while a positive test merely indicates genetic susceptibility. An advantage of HLA testing is that a gluten-free diet is not necessary for optimal diagnostic conclusiveness in marked contrast to autoantibody testing and histology [112]. According to the European Society for Pediatric Gastroenterology, Hepatology, and Nutrition guidelines for the diagnosis of coeliac disease testing of HLA-DQ2/DQ8 should be included in the diagnostic work-up of celiac disease in children. Small-bowel biopsy may not be necessary in pediatric population in case of symptomatic patients with significantly elevated tTG and EMA antibody titers and HLA-DQ2/DQ8 positivity [113].

A similar straightforward approach has been suggested for all age groups with suspected CD. This so-called four out of five rule allows the diagnosis of CD if four of the following five criteria are met: typical symptoms, significant elevation of CD antibodies, HLA-DQ2 or HLA-DQ8, typical biopsy result, response to gluten-free diet [114].

US Guidelines do not recommend HLA-DQ2/8 testing in the routine diagnostic work-up of CD; however, it may be useful in selected clinical situations, such as a discrepancy between histologic and serologic results [115].

HLA testing may also be useful to exclude CD in high-risk individuals, such as first-degree relatives of CD patients, patients with autoimmune diabetes mellitus, selective IgA deficiency, Down syndrome, Turner syndrome, Williams syndrome, or autoimmune thyroiditis. Testing could also be considered in case of unexplained iron deficiency anemia or early-onset osteoporosis [116, 117]. With the help of HLA testing the number of invasive diagnostic procedures may be reduced in this high-risk population.

5 MHC Class II: HLA-DRB1, Shared Epitope Hypothesis, and Rheumatoid Arthritis (RA)

Rheumatoid arthritis (RA) is a common chronic systemic autoimmune disease, which primarily presents with a symmetrical polyarthritis and has a prevalence of 0.5–1% worldwide [118]. Although the exact pathogenesis of RA remains unclear, its complex association with MHC class II molecules has been described [119]. The contribution of HLA genes to susceptibility is estimated to account for 50% of risk [120]. Remarkably, RA seems to be associated with such HLA-DRB1 alleles, which share sequences of five amino acids in position 70–74 of the antigen-binding groove of HLA-DR-β-chains, as described by the shared epitope hypothesis [121] [122]. The shared epitopes (for example QKRAA, QRRAA, RKRAA, RRRAA) are present in 70–90% of Caucasian patients with seropositive RA in contrast to the prevalence of 20–30% in the general population and patients with seronegative RA [123]. The highest relative risk of developing RA has been attributed to HLA-DRB1*0401 and *0404, which can be detected in 50–61% and 27–37% of seropositive patients, respectively. DRB1*0404 may also be associated with seronegative RA. Latter population also exhibits HLA-DRB1*0101 [124, 125]. The HLA-DRB1*0401/*0404 genotype is associated with elevated risk of disease, earlier onset, seropositivity, accelerated joint damage, and the presence of rheumatoid nodules [126]. Ethnic differences have been reported, the prevalence of DRB-1 in African American patients is lower (25%) [127]. HLA-DRB1*0405 is the most frequent allele in Asian RA patients [127, 128]. HLA-DRB1*1402 is associated with RA in Native American patients [127]. A large European meta-analysis has shown that HLA-DRB1*13:01 provides protection against Anti-citrullinated protein antibodies (ACPA) positive disease but not against ACPA negative RA [129].

Homozygosity or compound heterozygosity for HLA-DRB1 alleles containing one of the shared epitope sequences is associated with increased risk of developing RA. Patients carrying two shared epitope-containing HLA-DRB1*04 alleles—especially homozygosity for HLA-DRB1*0401—have a higher risk of extraarticular manifestations including rheumatoid vasculitis [130].

Several theories of the pathogenetic role of the shared epitopes have been proposed such as molecular mimicry, antigen presentation of arthritogenic peptides, as well as a role in the positive selection of specific autoreactive T-lymphocytes in the thymus. However, the exact mechanisms remain unclear.

Recent studies have shown that polymorphisms in certain amino acid positions of MHC molecules can better account for genetic susceptibility than solely the shared epitope hypothesis. Amino acids in positions 11, 13, 70, 71, and 74 of the DR-β chain show strong independent correlation with relative risk. Interestingly, only changes of the latter three affect the shared epitope motif. Positions 70 and 71 have a significant role in presentation of vimentin, alpha-enolase, and collagen as well as in modulating the interaction with T-cell receptors [131]. Positions 67 and 86 may also affect the binding of possible arthritogenic peptides [132]. It has been suggested that three of the above-mentioned amino acid positions (11/13, 71, 74) and position nine of HLA-B and HLA-DPB1 account for the majority of HLA-associated genetic susceptibility to RA [133]. HLA-DRB1 haplotypes also influence disease severity, mortality, and therapy response. Valine in amino acid position 11 of HLA-DRB1 has the strongest genetic association with radiologic damage independent of shared epitope status as well as with clinical and laboratory markers of inflammation and overall mortality. Positions 71 and 74 are also associated with erosive damage [134, 135]. These findings underline the additional importance of non-shared epitope polymorphisms.

Shared epitopes are associated significantly only with ACPA positive RA [136]. Citrullinated antigens bind preferentially to HLA-DRB1 with shared epitope sequences leading to the activation of autoreactive T cells and subsequently to the expansion of autoantibody-secreting B-lymphocytes. Patients carrying shared epitope motifs were more frequently ACPA positive in a dosage-dependent manner. Cigarette smoking, a major environmental risk factor of RA, has been shown to induce citrullination of proteins in the lung and its harmful effect may be due to interactions with HLA-DRB1 molecules and trigger the excessive immune reaction [137]. Indeed, heavy smoking increased risk of developing ACPA positive RA in the presence of shared epitope-containing HLA-DRB1 alleles, although no significant association has been found in patients with ACPA negative RA [138, 139].

Regarding the clinical usefulness of genetic testing in patients with suspected or diagnosed RA HLA-DRB1 analysis is neither included in current classification criteria nor recommended as a diagnostic tool by the current ACR/EULAR guidelines [140]. In a case-control study HLA-DRB1*0401 and *0404 had a sensitivity of 60% and a specificity of 64% as indicators for the future development of RA. The combination of anti-CCP2 antibodies and the testing of these two HLA-haplotypes proved to be the best approach to detect RA susceptibility [141]. Other authors found no benefit of additional shared epitope testing owing to its strong association with ACPA [142]. Genotyping of shared epitopes and HLA-DRB1 variants may provide valuable information regarding the choice of treatment options. A triple therapy containing methotrexate, hydroxychloroquine and sulfasalazine is more effective in the presence of shared epitopes than methotrexate monotherapy. No significant difference between therapy regimes was observed however in shared epitope negative patients [143]. The efficacy of TNF-alpha inhibitors may also be influenced by HLA-DRB1 haplotypes. TNFi response was not associated with the presence of shared epitopes but rather with amino acid position 11. Patients with valine at this position had significantly better EULAR responses independent of zygosity and ACPA status compared to noncarriers [134, 144].

Nevertheless, HLA-DRB1 analysis is available in several commercial medical laboratories. Results should be interpreted with caution: on one hand a positive result merely indicates genetic predisposition and is not suitable for the diagnosis of RA. On the other hand, RA can certainly not be excluded in case of absence of shared epitopes, especially in non-Caucasian patients.

References

1. Reveille JD (2011) Epidemiology of Spondyloarthritis in North America. The American Journal of Medical Sciences 341(4):284–286
2. Stolwijk C, Boonen A, Tubergen AV et al (2012) Epidemiology of Spondyloarthritis. Rheum Dis Clin N Am 38(3):441–476
3. Rohekar S, Pope J (2010) Assessment of work disability in seronegative spondyloarthritis. Clin Exp Rheumatol 28:35–40
4. Ramonda R, Marchesoni A, Carletto A et al (2016) Patient-reported impact of spondyloarthritis on work disability and working life: the ATLANTIS survey. Arthritis Res Ther 18:78
5. Mathieu A, Cauli A, Fiorillo MT et al (2008) HLA-B27 and ankylosing spondylitis geographic distribution versus malaria endemic: casual or causal liaison? Ann Rheum Dis 67:138–140
6. Richens JE, Prasad ML, Bhatia K et al (1986) Arthritis and HLA-B27 in Papua New Guinea. Br Med J (Clin Res Ed) 293(6556):1209
7. Bhatia K, Richens J, Prasad ML, Koki G (1988) High prevalence of the haplotype HLA-A11, B27 in arthritis patients from the highlands of Papua New Guinea. Tissue Antigens 31(2):103–106
8. Gofton JP, Chalmers A, Price GE et al (1984) HL-A 27 and ankylosing spondylitis in B.C. Indians. J Rheumatol 11(5):572–573
9. Boyer GS, Templin DW, Cornoni-Huntley JC et al (1994) Prevalence of spondyloarthropathies in Alaskan Eskimos. J Rheumatol 21(12):2292–2297
10. Erdesz S, Shubin SV, Shoch BP et al (1994) Spondyloarthropathies in circumpolar populations of Chukotka (Eskimos and Chukchi): epidemiology and clinical characteristics. J Rheumatol 21(6):1101–1104

11. Liu X, Li YR, Hu LH et al (2010) High frequencies of HLA-B27 in Chinese patients with suspected of ankylosing spondylitis. Rheumatol Int 30(10):1305–1309
12. Ho HH, Chen JY (2013) Ankylosing spondylitis: Chinese perspective, clinical phenotypes, and associated extra-articular systemic features. Curr Rheumatol Rep 15:344
13. Mustafa KN, Hammoudeh M, Khan MA (2012) HLA-B27 prevalence in Arab populations and among patients with ankylosing spondylitis. J Rheumatol 39:1675–1677
14. Khan MA (1987) Race-related differences in HLA association with ankylosing spondylitis and Reiter's disease in American blacks and whites. J Natl Med Assoc 70(1):41–42
15. Tanaka H, Akaza T, Juji T (1996) Report of the Japanese central bone marrow data center. Clin Transpl 10:139–144
16. Khan MA (1995) HLA-B27 and its subtypes in world populations. Curr Opin Rheumatol 7:263–269
17. Tikly M, Njobvu P, McGill P (2014) Spondyloarthritis in sub Saharan Africa. Curr Rheumatol Rep 16(6):421
18. Riecker HH, Neel JV, Test A (1950) The inheritance of spondylitis rhizomelique (ankylosing spondylitis) in the K. Family. Ann Intern Med 33(5):1254–1273
19. Schlosstein L, Terasaki PI, Bluestone R et al (1973) High association of an HL-A antigen, W27, with ankylosing spondylitis. N Engl J Med 288:704–706
20. Brewerton DA, Hart FD, Nicholls A et al (1973) Ankylosing spondylitis and HL-A 27. Lancet 1(7809):904–907
21. Akkoc N, Khan MA (2006) Epidemiology of Ankylosing spondylitis and related Spondyloarthropathies. In: Ankylosing spondylitis and the Spondyloarthropathies. Elsevier, Amsterdam, pp 117–131
22. Kopplin LJ, Mount G, Suhler EB (2016) Review for disease of the year: epidemiology of HLA-B27 associated ocular disorders. Ocul Immunol Inflamm 24(4):470–475
23. Queiro R, Morante I, Cabezas I et al (2016) HLA-B27 and psoriatic disease: a modern view of an old relationship. Rheumatology (Oxford) 55(2):221–229
24. Braun JBM, Remlinger G (1998) Prevalence of spondylarthropathies in HLA-B27 positive and negative blood donors. Arthritis & Rheumatology 41(1):58–67
25. Reveille JD, Hirsch R, Dillon CF et al (2012) The prevalence of HLA-B27 in the US: data from the US National Health and nutrition examination survey, 2009. Arthritis & Rheumatology 64(5):1407–1411
26. Costantino F, Talpin A, Said-Nahal R et al (2013) Prevalence of Spondyloarthritis in reference to HLA-B27 in the French population: results of the GAZEL cohort. Ann Rheum Dis 74(4):689–693
27. Nasution AR, Mardjuadi A, Suryadhana NG et al (1993) Higher relative risk of spondyloarthropathies among B27 positive Indonesian Chinese than native Indonesians. J Rheumatol 20:988–990
28. van der Linden S, Valkenburg H, Cats A (1983) The risk of developing ankylosing spondylitis in HLA-B27 positive individuals: a family and population study. Br J Rheumatol 22(4 Suppl 2):18–19
29. Feltkamp TE (1995) Factors involved in the pathogenesis of HLA-B27 associated arthritis. Scand J Rheumatol 101:213–217
30. Jaakkola E, Herzberg I, Laiho K et al (2006) Finnish HLA studies confirm the increased risk conferred by HLA-B27 homozygosity in ankylosing spondylitis. Ann Rheum Dis 65(6):775–780
31. van Der Linden SM, Valkenburg HA, De Jongh BM et al (1984) The risk of developing ankylosing spondylitis in HLA-B27 positive individuals. A comparison of relatives of spondylitis patients with the general population. Arthritis Rheum 27(3):241–249
32. Kim TJ, Na KS, Lee HJ et al (2009) HLA-B27 homozygosity has no influence on clinical manifestations and functional disability in ankylosing spondylitis. Clin Exp Rheumatol 27:574–579
33. Dangoria NS, DeLay ML, Kingsbury DJ et al (2002) HLA-B27 misfolding is associated with aberrant intermolecular disulfide bond formation (dimerization) in the endoplasmic reticulum. J Biol Chem 277:23459–23468
34. Khan MA (2013) Polymorphism of HLA-B27: 105 subtypes currently known. Curr Rheumatol Rep 15:362
35. Frankenberger B, Breitkopf S, Albert E et al (1997) Routine molecular genotyping of HLA-B27 in spondyloarthropathies overcomes the obstacles of serological typing and reveals an increased B *2702 frequency in ankylosing spondylitis. J Rheumatol 24(5):899–903
36. Lin J, Lü H, Feng C (1996) Ankylosing spondylitis and heterogeneity of HLA-B27 in Chinese. Chinese Medical Journal (Engl) 109(4):313–316
37. Taurog DJ (2007) The mystery of HLA B27: if it isn't one thing, it's another. Arthritis Rheum 56(8):2478–2481
38. Hill AVS, Allsop CEM, al KD (1991) HLA class I typing by PCR: HLA-B27 and an African B27 subtype. Lancet 337:640–642
39. Cauli A, Vacca A, Dessole G et al (2008) HLA-B* 2709 and lack of susceptibility to sacroiliitis: further support from the clinic. Clin Exp Rheumatol 26(6):1111–1112

40. Yang T, Duan Z, Wu S et al (2014) Association of HLA-B27 genetic polymorphisms with ankylosing spondylitis susceptibility worldwide: a meta-analysis. Modern rheumatology / the Japan Rheumatism Association 24(1):150–161
41. Lin H, Gong YZ (2017) Association of HLA-B27 with ankylosing spondylitis and clinical features of the HLA-B27-associated ankylosing spondylitis: a meta-analysis. Rheumatol Int 37(8):1267–1280
42. Mear JP, Schreiber KL, Munz C et al (1999) Misfolding of HLA-B27 as a result of its B suggests a novel mechanism for its role in susceptibility to spondyloarthropathies. J Immunol 163(12):6665–6670
43. Galocha B, López de Castro JA (2008) Folding of HLA-B27 subtypes is determined by the global effect of polymorphic residues and shows incomplete correspondence to ankylosing spondylitis. Arthritis & Rheumatology 58:401–412
44. Sheehan NJ (2010) HLA-B27: what's new? Rheumatology (Oxford) 49:621–631
45. Ebringer A (1983) The cross-tolerance hypothesis, HLA-B27 and ankylosing spondylitis. Br J Rheumatol 22(4 Suppl 2):53–66
46. Schwimmbeck PL, Oldstone MB (1988) Molecular mimicry between human leukocyte antigen B27 and Klebsiella. Consequences for spondyloarthropathies. Am J Med 85(6A):51–53
47. Lahesmaa R, Skurnik M, Vaara M et al (1991) Molecular mimickry between HLA B27 and Yersinia, Salmonella, Shigella and Klebsiella within the same region of HLA α1-helix. Clinical & Experimental Immunology 86:399–404
48. Fielder M, Pirt SJ, Tarpey I et al (1995) Molecular mimicry and ankylosing spondylitis: possible role of a novel sequence in pullulanase of *Klebsiella pneumoniae*. FEBS Lett 369:243–248
49. Schittenhelm RB, Sian TC, Wilmann PG et al (2015) Revisiting the Arthritogenic peptide theory: quantitative not qualitative changes in the peptide repertoire of HLA–B27 Allotypes. Arthritis & Rheumatology 67:702–713
50. Hoentjen F, Tonkonogy SL, Qian BF et al (2007) CD4(+) T lymphocytes mediate colitis in HLA-B27 transgenic rats monoassociated with nonpathogenic Bacteroides vulgatus. Inflamm Bowel Dis 13:317–324
51. Rath HC, Wilson KH, Sartor RB (1999) Differential induction of colitis and gastritis in HLA-B27 transgenic rats selectively colonized with Bacteroides vulgatus or Escherichia coli. Infect Immun 67:2969–2974
52. Taurog JD, Richardson JA, Croft JT et al (1994) The germfree state prevents development of gut and joint inflammatory disease in HLA-B27 transgenic rats. J Exp Med 180(6):2359–2364
53. Costello ME, Elewaut D, Kenna TJ et al (2013) Microbes, the gut and ankylosing spondylitis. Arthritis Res Ther 15:214
54. Lin P, Bach M, Asquith M et al (2014) HLA-B27 and human β2-microglobulin affect the gut microbiota of transgenic rats. PLoS One 9:e105684
55. Rudwaleit M, van der Heijde D, Khan MA et al (2004) How to diagnose axial spondyloarthritis early. Ann Rheum Dis 63(5):535–543
56. Rudwaleit M, van der Heijde D, Landewe R et al (2009) The development of assessment of SpondyloArthritis international society classification criteria for axial spondyloarthritis (part II): validation and final selection. Ann Rheum Dis 68(6):777–783
57. Kirveskari J, Kellner H, Wuorela M et al (1997) False-negative serological HLA-B27 typing results may be due to altered antigenic epitopes and can be detected by polymerase chain reaction. Br J Rheumatol 36(2):185–189
58. Levering WH, Wind H, Sintnicolaas K et al (2003) Flow cytometric HLA-B27 screening: cross-reactivity patterns of commercially available anti-HLA-B27 monoclonal antibodies with other HLA-B antigens. Cytometry Part B Clinical Cytometry 54:28–38
59. Sieper J, Srinivasan S, Zamani O et al (2013) Comparison of two referral strategies for diagnosis of axial spondyloarthritis: the Recognising and diagnosing Ankylosing spondylitis reliably (RADAR) study. Ann Rheum Dis 72:1621–1627
60. Linssen A, Feltkamp TE (1988) B27 positive diseases versus B27 negative diseases. Ann Rheum Dis 47(5):431–439
61. Feldtkeller E, Khan MA, van der Heijde D et al (2003) Age at disease onset and diagnosis delay in HLA-B27 negative vs. positive patients with ankylosing spondylitis. Rheumatol Int 23:61–66
62. Marzo-Ortega H, McGonagle D, O'Connor P et al (2009) Baseline and 1-year magnetic resonance imaging of the sacroiliac joint and lumbar spine in very early inflammatory back pain. Relationship between symptoms, HLA-B27 and disease extent and persistence. Ann Rheum Dis 68:1721–1727
63. Chung HY, Machado P, van der Heijde D et al (2011) HLA-B27 positive patients differ from HLA-B27 negative patients in clinical presentation and imaging: results from the DESIR cohort of patients with recent onset axial spondyloarthritis. Ann Rheum Dis 70:1930–1936
64. Khan MA, Kushner I, Braun WE (1977) Comparison of clinical features in HLA-B27

positive and negative patients with ankylosing spondylitis. Arthritis Rheum 20:909–912
65. Hamersma J, Cardon LR, Bradbury L et al (2001) Is disease severity in ankylosing spondylitis genetically determined? Arthritis Rheum 44:1396–1400
66. Bennett AN, McGonagle D, O'Connor P et al (2008) Severity of baseline magnetic resonance imaging-evident sacroiliitis and HLA-B27 status in early inflammatory back pain predict radiographically evident ankylosing spondylitis at eight years. Arthritis & Rheumatology 58:3413–3418
67. Freeston J, Barkham N, Hensor E et al (2007) Ankylosing spondylitis, HLA-B27 positivity and the need for biologic therapies. Joint Bone Spine 74(2):140–143
68. Rudwaleit M, Listing J, Brandt J et al (2004) Prediction of a major clinical response(BASDAI 50) to tumour necrosis factor alpha blockers in ankylosing spondylitis. Ann Rheum Dis 63:665–670
69. Vastesaeger N, Van Der Heijde D, Inman R et al (2011) Predicting the outcome of ankylosing spondylitis therapy. Ann Rheum Dis 70:973–981
70. Brewerton DA, Caffrey M, Nicholls A et al (1974) HL-A 27 and the arthropathies associated with ulcerative colitis and psoriasis. Lancet 1:956–958
71. Guðjónsson JE, Valdimarsson H, Kárason A et al (2002) HLA-Cw6-positive and HLA-Cw6-negative patients with psoriasis vulgaris have distinct clinical features. J Investig Dermatol 118:362–365
72. FitzGerald O, Haroon M, Giles JT et al (2015) Concepts of pathogenesis in psoriatic arthritis: genotype determines clinical phenotype. Arthritis Res Ther 17(1):115
73. Rosenbaum JT (1992) Acute anterior uveitis and spondyloarthropathies. Rheum Dis Clin N Am 18:143–151
74. Schiellerup P, Krogfelt KA, Locht H (2008) A comparison of self-reported joint symptoms following infection with different enteric pathogens: effect of HLA-B27. J Rheumatol 35(3):480–487
75. Ollier W, Pepper L, Thomson W (1994) HLA-B27 as a marker for developing subluxations of the cervical spine in RA. Arthritis & Rheumatology 37(suppl):A1017
76. den Uyl D, van der Horst-Bruinsma IE, van Agtmael M (2004) Progression of HIV to AIDS: a protective role for HLA-B27? AIDS Rev 6(2):89–96
77. Mustonen J, Partanen J, Kanerva M et al (1998) Association of HLA B27 with benign clinical course of Nephropathia Epidemica caused by Puumala hantavirus. Scand J Immunol 47(3):277–279
78. Neumann-Haefelin C (2013) HLA-B27-mediated protection in HIV and hepatitis C virus infection and pathogenesis in spondyloarthritis: two sides of the same coin? Curr Opin Rheumatol 25:426–433
79. Reddy V, Desai A, Krishna SS et al (2017) Molecular mimicry between Chikungunya virus and host components: a possible mechanism for the arthritic manifestations. PLoS Negl Trop Dis 11(1):e0005238
80. Australo-Anglo-American Spondyloarthritis Consortium (TASC), Reveille JD, Sims AM et al (2010) Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. Nat Genet 42(2):123–127
81. International Genetics of Ankylosing Spondylitis Consortium (IGAS), Cortes A, Hadler J et al (2013) Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. Nat Genet 45(7):730–738
82. Reveille JD (2015) Biomarkers for diagnosis, monitoring of progression, and treatment responses in ankylosing spondylitis and axial spondyloarthritis. Clin Rheumatol 34:1009–1018
83. Azizlerli G, Kose AA, Sarica R et al (2003) Prevalence of Behcet's disease in Istanbul, Turkey. Int J Dermatol 42:803–806
84. Mahr A, Belarbi L, Wechsler B et al (2008) Population-based prevalence study of Behcet's disease: differences by ethnic origin and low variation by age at immigration. Arthritis & Rheumatology 58(12):3951–3959
85. Çölgeçen E, Özyurt K, Ferahbaş A et al (2015) The prevalence of Behçet's disease in a city in central Anatolia in Turkey. Int J Dermatol 54:286–289
86. Ohno S, Ohguchi M, Hirose S et al (1982) Close association of HLA-Bw51 with Behçet's disease. Arch Ophthalmol 100:1455–1458
87. Maldini C, Lavalley MP, Cheminant M et al (2012) Relationships of HLA-B51 or B5 genotype with Behcet's disease clinical characteristics: systematic review and meta-analyses of observational studies. Rheumatology (Oxford) 51(5):887–900
88. Kirino Y, Ideguchi H, Takeno M et al (2016) Continuous evolution of clinical phenotype in 578 Japanese patients with Behçet's disease: a retrospective observational study. Arthritis Res Ther 18:217
89. Sugisaki K, Saito R, Takagi T et al (2005) HLA-B52-positive vasculo-Behçet disease: usefulness of magnetic resonance angiography, ultrasound study, and computed tomographic angiography for the early evaluation of multiarterial lesions. Mod Rheumatol 15(1):56–61
90. Arber N, Klein T, Meiner Z et al (1991) Close association of HLA-B51 and B52 in Israeli

patients with Behçet's syndrome. Ann Rheum Dis 50:351–353
91. Verity DH, Wallace GR, Vaughan RW et al (2003) Behçet's disease: from Hippocrates to the third millennium. Br J Ophthalmol 87:1175–1183
92. Demirseren DD, Ceylan GG, Akoglu G et al (2014) HLA-B51 subtypes in Turkish patients with Behcet's disease and their correlation with clinical manifestations. Genet Mol Res 13:4788–4796
93. Yasouka H, Yamaguchi Y, Mizuki N et al (2008) Preferential activation of circulating CD8+ and γδ T cells in patients with active Behçet's disease and HLA-B51. Clin Exp Rheumatol 26(Suppl. 50):S59–S63
94. Takeno M, Kariyone A, Yamashita N et al (1995) Excessive function of peripheral blood neutrophils from patients with Behcet's disease and from HLA-B51 transgenic mice. Arthritis Rheum 38:426–433
95. Eksioglu-Demiralp E, Direskeneli H, Kibaroglu A et al (2001) Neutrophil activation in Behcet's disease. Clin Exp Rheumatol 19(5 Suppl 24):S19–S24
96. International Team for the Revision of the International Criteria for Behçet's Disease (ITR-ICBD) (2014) The international criteria for Behçet's disease (ICBD): a collaborative study of 27 countries on the sensitivity and specificity of the new criteria. J Eur Acad Dermatol Venereol 28:338–347
97. Kuranov AB, Kötter I, Henes JC et al (2014) Behçet's disease in HLA-B*51 negative Germans and Turks shows association with HLA-Bw4-80I. Arthritis Res Ther 16(3):R116
98. Ortiz-Fernández L, Carmona F-D, Montes-Cano M-A et al (2016) Genetic analysis with the Immunochip platform in Behçet disease. Identification of residues associated in the HLA class I region and new susceptibility loci. PLoS One 11(8):e0161305
99. Gujral N, Freeman HJ, Thomson AB (2012) Celiac disease: prevalence, diagnosis, pathogenesis and treatment. World J Gastroenterol 18:6036–6059
100. van Heel DA, Franke L, Hunt KA et al (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet 39:827–829
101. Garner C, Ahn R, Ding YC et al (2014) Genome-wide association study of celiac disease in North America confirms FRMD4B as new celiac locus. PLoS One 9(7):e101428
102. Karell K, Louka AS, Moodie SJ et al (2003) European genetics cluster on celiac disease. HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European genetics cluster on celiac disease. Hum Immunol 64:469–477
103. Fasano A (2016) Genetics of celiac disease. http://emedicine.medscape.com/article/1790189-overview
104. Vader W, Stepniak D, Kooy Y et al (2003) The HLA-DQ2 gene dose effect in celiac disease is directly related to the magnitude and breadth of gluten-specific T cell responses. Proc Natl Acad Sci U S A 100(21):12390–12395
105. Abraham G, Rohmer A, Tye-Din JA et al (2015) Genomic prediction of celiac disease targeting HLA-positive individuals. Genome Med 7:72
106. Shan L, Molberg O, Parrot I et al (2002) Structural basis for gluten intolerance in celiac sprue. Science 297:2275–2279
107. Arentz-Hansen H, Körner R, Molberg Ø et al (2000) The intestinal T cell response to α-Gliadin in adult celiac disease is focused on a single Deamidated glutamine targeted by tissue transglutaminase. J Exp Med 191(4):603–612
108. Kagnoff MF, Austin RK, Hubert JJ et al (1984) Possible role for a human adenovirus in the pathogenesis of celiac disease. J Exp Med 160(5):1544–1557
109. De Palma G, Capilla A, Nova E et al (2012) Influence of milk-feeding type and genetic risk of developing coeliac disease on intestinal microbiota of infants: the PROFICEL study. PLoS One 7:e30791
110. Olivares M, Neef A, Castillejo G et al (2015) The HLA-DQ2 genotype selects for early intestinal microbiota composition in infants at high risk of developing coeliac disease. Gut 64:406–417
111. Sollid LM, Lie BA (2005) Celiac disease genetics: current concepts and practical applications. Clinical Gastroenterology Hepatology 3:843–851
112. Alaedini A, Green PH (2005) Narrative review: celiac disease: understanding a complex autoimmune disorder. Ann Intern Med 142:289–298
113. Husby S, Koletzko S, Korponay-Szabo IR et al (2012) European Society for Pediatric Gastroenterology, Hepatology, and nutrition guidelines for the diagnosis of coeliac disease. J Pediatr Gastroenterol Nutr 54:136–160
114. Catassi C, Fasano A (2010) Celiac disease diagnosis: simple rules are better than complicated algorithms. Am J Med 123:691–693
115. Rubio-Tapia A, Hill ID, Kelly CP et al (2013) ACG clinical guidelines: diagnosis and management of celiac disease. Am J Gastroenterol 108:656–676
116. Rostom A, Murray JA, Kagnoff MF (2006) American gastroenterological association

(AGA) institute technical review on the diagnosis and management of celiac disease. Gastroenterology 131:1981–2002
117. Hill ID, Dirks MH, Liptak GS et al (2005) Guideline for the diagnosis and treatment of celiac disease in children: recommendations of the north American Society for Pediatric Gastroenterology, Hepatology and nutrition. J Pediatr Gastroenterol Nutr 40:1–19
118. Alamanos Y, Drosos AA (2005) Epidemiology of adult rheumatoid arthritis. Autoimmun Rev 4:130–136
119. Stastny P (1976) Mixed lymphocyte cultures in rheumatoid arthritis. J Clin Investig 57:1148–1157
120. Barton A, Worthington J (2009) Genetic susceptibility to rheumatoid arthritis: an emerging picture. Arthritis Rheum 61:1441–1446
121. Gregersen PK, Silver J, Winchester RJ (1987) The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. Arthritis & Rheumatology 30:1205–1213
122. Silver J, Goyert SM (1985) Epitopes are the functional units of Ia molecules and form the molecular basis for disease susceptibility. In: Ferrone S, Solheim BG, Moller E (eds) HLA Class II Antigens. Springer-Verlag, Berlin
123. Stastny P (1978) Association of the B-cell alloantigen DRw4 with rheumatoid arthritis. N Engl J Med 298:869–871
124. Gonzalez-Gay MA, Hajeer AH, Dababneh A et al (2001) Seronegative rheumatoid arthritis in elderly and polymyalgia rheumatica have similar patterns of HLA association. J Rheumatol 28:122–125
125. Weyand CM, Klimiuk PA, Goronzy JJ (1998) Heterogeneity of rheumatoid arthritis: from phenotypes to genotypes. Semin Immunopathol 20(1–2):5–22
126. MacGregor A, Ollier W, Thomson W et al (1995) HLA-DRB1* 0401/0404 genotype and rheumatoid arthritis: increased association in men, young age at onset, and disease severity. J Rheumatol 22(6):1032–1036
127. Hughes LB, Morrison D, Kelley JM et al (2008) The HLA-DRB1 shared epitope is associated with susceptibility to rheumatoid arthritis in African Americans through European genetic admixture. Arthritis & Rheumatology 58:349–358
128. Lee HS, Lee KW, Song GG et al (2004) Increased susceptibility to rheumatoid arthritis in Koreans heterozygous for HLA–DRB1*0405 and *0901. Arthritis Rheum 50:3468–3475
129. van der Woude D, Lie BA, Lundström E et al (2010) Protection against anti-citrullinated protein antibody-positive rheumatoid arthritis is predominantly associated with HLA-DRB1*1301: a meta-analysis of HLA-DRB1 associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in four European populations. Arthritis & Rheumatology 62:1236–1245
130. Turesson C, Schaid DJ, Weyand CM et al (2005) The impact of HLA-DRB1 genes on extra-articular disease manifestations in rheumatoid arthritis. Arthritis Res Ther 7(6):R1386–R1393
131. Anderson KM, Roark CL, Portas M et al (2016) A molecular analysis of the shared epitope hypothesis: binding of arthritogenic peptides to DRB1*04 alleles. Arthritis & Rheumatology 68:1627–1636
132. Roark CL, Anderson KM, Aubrey MT et al (2016) Arthritogenic peptide binding to DRB1*01 alleles correlates with susceptibility to rheumatoid arthritis. J Autoimmun 72:25–32
133. Raychaudhuri S et al (2012) Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet 44:291–296
134. Viatte S, Plant D, Han B et al (2015) Association of HLA-DRB1 haplotypes with rheumatoid arthritis severity, mortality, and treatment response. JAMA 313:1645–1656
135. Ling SF, Viatte S, Lunt M et al (2016) HLA-DRB1 amino acid positions 11/13, 71, and 74 are associated with inflammation level, disease activity, and the health assessment questionnaire score in patients with inflammatory polyarthritis. Arthritis & Rheumatology 68:2618–2628
136. Huizinga TW, Amos CI, van der Helm-van Mil AH et al (2005) Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins. Arthritis & Rheumatology 52(11):3433–3438
137. Klareskog L, Stolt P, Lundberg K et al (2006) A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA–DR (shared epitope)–restricted immune reactions to autoantigens modified by citrullination. Arthritis & Rheumatology 54:38–46
138. Kim K, Jiang X, Cui J et al (2015) Interactions between amino-acid-defined MHC class II variants and smoking for seropositive rheumatoid arthritis. Arthritis & Rheumatology 67(10):2611–2623
139. Jiang X, Kallberg H, Chen Z et al (2016) An Immunochip-based interaction study of contrasting interaction effects with smoking in ACPA-positive versus ACPA-negative rheu-

matoid arthritis. Rheumatology (Oxford) 55(1):149–155
140. Aletaha D, Neogi T, Silman AJ et al (2010) 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European league against rheumatism collaborative initiative. Ann Rheum Dis 69:1580–1588
141. Berglin E, Padyukov L, Sundin U et al (2004) A combination of autoantibodies to cyclic citrullinated peptide (CCP) and HLA-DRB1 locus antigens is strongly associated with future onset of rheumatoid arthritis. Arthritis Res Ther 6(4):R303–R308
142. Van der Cruyssen B, Hoffman IEA, Peene I et al (2007) Prediction models for rheumatoid arthritis during diagnostic investigation: evaluation of combinations of rheumatoid factor, anti-citrullinated protein/peptide antibodies and the human leucocyte antigen-shared epitope. Ann Rheum Dis 66(3):364–369
143. O'Dell JR, Nepom BS, Haire C et al (1998) HLA-DRB1 typing in rheumatoid arthritis: predicting response to specific treatments. Ann Rheum Dis 57(4):209–213
144. Danila MI, Hughes LB, Bridges SL (2008) Pharmacogenetics of etanercept in rheumatoid arthritis. Pharmacogenomics 9:1011–1015

Chapter 3

The IPD Databases: Cataloguing and Understanding Allele Variants

Jashan P. Abraham, Dominic J. Barker, James Robinson, Giuseppe Maccari, and Steven G. E. Marsh

Abstract

The IMGT/HLA Database has provided a repository for information regarding polymorphism in the genes of the immune system since 1998. In 2003, it was absorbed into the Immuno Polymorphism Database (IPD). The IPD project has enabled us to create and maintain a platform for curating and publishing locus-specific databases which are either involved directly with, or relate to, the function of the Major Histocompatibility Complex across a number of species. In collaboration with specialist groups and nomenclature committees individual sections have been curated prior to their submission to the IPD for online publication. The IPD consists of five core databases, with the primary database being the IMGT/HLA Database. With the work of various nomenclature committees, the HLA Informatics Group, and alongside the European Bioinformatics Institute, we provide access to this data through the website (http://www.ebi.ac.uk/ipd/) to the public domain. The IPD project continually develops new tools in conjunction with on-going scientific developments—such as Next-Generation Sequencing—to maintain efficiency and usability in response to user feedback and requests. The website is updated on a regular basis to ensure that new and confirmatory sequences are distributed to the immunogenetics community, as well as the wider research and clinical communities.

Key words HLA, KIR, MHC, Alleles, Variants, Transplantation, Database

1 Introduction

The Immuno Polymorphism Database (IPD) is comprised of a set of specialized databases regarding the study of polymorphic genes within the immune system. We have collaborated with specialist groups and nomenclature committees that curate individual sections prior to their submission to IPD for online publication [1]. The IPD project stores all gathered data in sets of related databases and currently consists of five databases: the IPD-IMGT/HLA Database, which is comprised of the sequences of the human Major Histocompatibility Complex (MHC); the IPD-KIR Database, containing allelic sequences of the human Killer-cell Immunoglobulin-like Receptors (KIR); the IPD-MHC Database

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_3,

providing access to non-human MHC sequences; IPD-HPA Database which records human platelet antigens; and finally the IPD-ESTDAB, which allows accessibility to the European Searchable Tumour Cell-Line Database, a cell bank of melanoma cell lines characterized immunologically.

2 Allele Databases

The IPD project's major use is as a database repository for polymorphic gene sequence data. It combines polymorphic gene systems curated by various nomenclature committees with tools and infrastructure designed to publish and maintain the data, developed by a core bioinformatics team. The project's aim is facilitating the nomenclature committees' gene-specific guidelines and definitions on naming genes and alleles into the curation of submissions to IPD. Every allele is defined as a unique nucleotide sequence within the database and may vary in length, from the full length of the gene, from the 5′ untranslated region (UTR) to 3′ UTR, or to an obligatory number of exons, dependent on the classification of the sequence. Database entries do not entail comparative descriptions between two sequences. However, detailed information on the process used to obtain the sequence is mandatory, particularly for any novel sequences submitted. Allele variants that present a Single Nucleotide Polymorphism (SNP) compared to a reference sequence must be verified by proof of the sequence, the methodology used to obtain this sequence and details of the sample and source used to produce the sequence, as well as evidence of multiple sequencing.

3 The IPD-IMGT/HLA Database

The IPD-IMGT/HLA Database was established to provide a locus-specific database (LSDB) for the allelic sequences of the genes in the HLA system—the human MHC. The database was first released in 1998 and was later incorporated as a module of IPD in 2012. With over 220 genes [2], the MHC is one of the most complex and polymorphic regions of the human genome [3]. The central genes of interest in the system are 21 polymorphic HLA genes within the 6p21.3 region on the short arm of human chromosome 6. The HLA proteins encoded by these genes mediate response to infectious disease, and influences the outcome of cell and organ transplants in humans. Polymorphism in MHC genes can attain an extremely high level—there are almost 5000 variants observed in HLA-B alone. Variation to this extent may be considered hyperpolymorphic when contrasted with other

gene systems. The MHC has been identified as comprising of three distinct regions. The class I region, at the telomeric end of the MHC, encodes genes for the classical HLA class I molecules, HLA-A, -B and –C, and the non-classical molecules HLA-E, -F, and -G. These molecules, co-dominantly expressed on the cell surface, are responsible for presenting intracellularly derived peptides to CD8 positive T cells. HLA class II genes, located within the class II region at the centromeric end of the MHC, encode the classical HLA-DR, –DQ, and –DP molecules, as well as the non-classical HLA-DM and -DO. The expression of HLA class II genes is limited to where CD4 positive T cells are presented extracellularly derived peptides by professional antigen presenting cells only. Numerous non-HLA immune system genes are found within the class III region, located between the class I region and the class II regions. The pressing urgency for a curated LSDB ranging over these polymorphic variants is obvious, given the nomenclature now covers well over 50 genes and 17,000 alleles. The 16th of December 1998 marked the IMGT/HLA Database's first public release [4]. The database has been updated every 3 months subsequently, as the IPD-IMGT/HLA Database, spanning 75

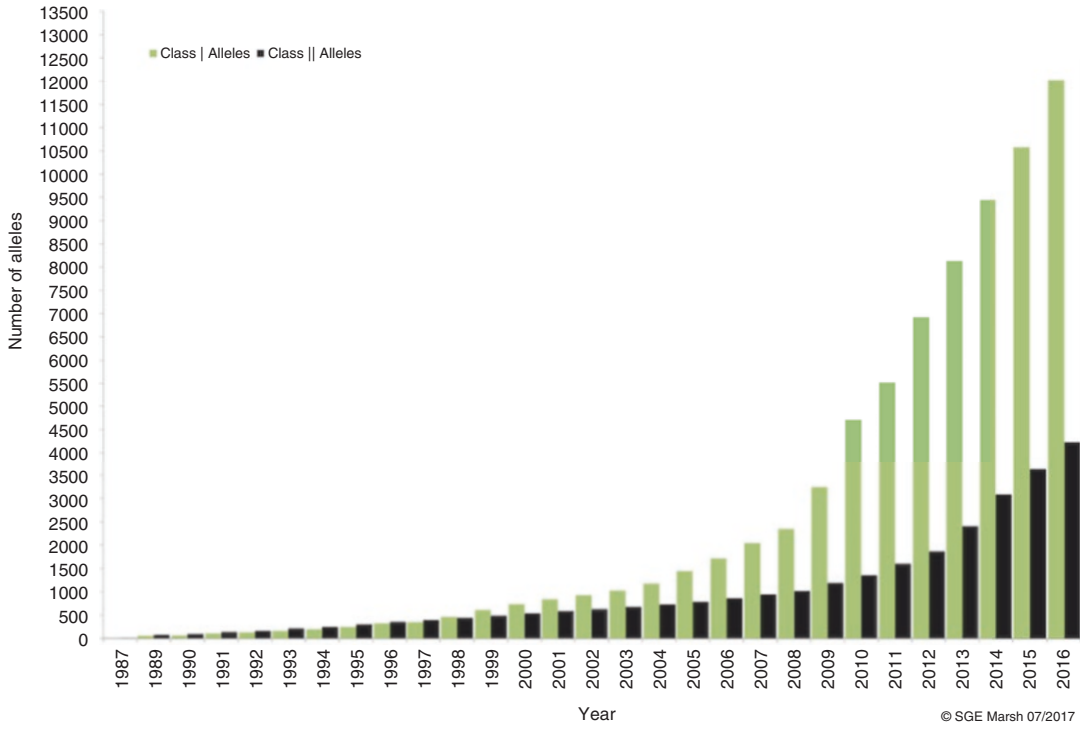


Fig. 1 Growth of the IPD-IMGT/HLA Database. The cumulative number of HLA allele sequences in the IPD-IMGT/HLA Database is shown for class I (green) and class II (black) alleles. The slope of the line reflects the rate of acquisition

releases, (Fig. 1), to include all sequences designated as publicly available which have been given official names by the WHO Nomenclature Committee at the time of release publication.

The sustained use of the IPD-IMGT/HLA Database by the transplantation community is a vital force in its progression and on-going success. HLA molecules are instrumental in solid organ [5, 6] and hematopoietic stem cell transplantation (HSCT) [7], with the success of these procedures being strongly correlated with the degree of HLA matching between the donor and recipient [8]. HLA matching has been shown to be critical in determining patient outcomes subsequent to respective hematological disorders being treated by receiving unrelated donor hematopoietic stem cells [9]. Thus, there has been progressive improvement in the achieved levels of resolutions via HLA class I and class II typing methods. Distinguishing synonymous and non-synonymous mutations, within the protein coding domains of HLA class I and II nucleotide sequences, is the current focus of HLA typing. These peptide-binding domains ultimately interact with variable lymphocyte receptors. The steady improvement in typing resolution has made the construction and progression of an accurate and expansive nucleotide sequence database for all polymorphic HLA class I and II genes essential. A broad, non-specialist sequence database being used for these sequences can lead to issues with consistency in keyword descriptions, erroneous sequences, and uncorrected errors.

A SNP between a recipient and their donor could have a major impact on the outcome and success of a transplant. Therefore, it is vital for an HLA sequence repository to maintain high standards of both control and curation—hence the minimum set of criteria required for any submission made to the IPD-IMGT/HLA Database. These criteria have been set in order to ensure all submissions and the clinically relevant data meet the highest standard possible. Submitted sequences that do not meet these standards are not accepted, though they may be found within generalist databases such as the European Nucleotide Archive (ENA), GenBank and the DNA Data Bank of Japan (DDBJ). Submissions are then checked to ensure the appropriate steps have been taken to correctly identify novel polymorphisms. This process uses in-house pipelines, utilizing both Basic Local Alignment Search Tool (BLAST) [10] and Clustal [11], both searching and aligning the submitted sequence against known and unnamed but submitted sequences at amino acid, coding DNA sequence (CDS) and genomic levels. The search results should identify any discrepancy between the submitted sequence and those sequences that already exist and provide further detailed analysis that allows the appropriate naming of an allele. In addition to the automated analysis, curators assess each sequence to validate their constitution and the

mechanism of their generation. Modern bioinformatics techniques are employed for the curation, annotation, and analysis of sequences submitted to the IPD-IMGT/HLA Database. Furthermore, all sequences are checked throughout the various stages of the submission process by experts, to ensure and improve accuracy. To ease the dissemination of newly curated sequences to the wider community, novel alleles are added periodically once official names have been assigned to them. Four times a year, all named and publicly available alleles are added to the public copy of the IPD-IMGT/HLA Database, with its online resources updated and the wider community is notified of the release. Users may then download the most recent version of the database for local resources, ensuring clinical testing and analysis is undergone using the most recent data available.

4 HLA Polymorphism and Next-Generation Sequencing

Despite recent demand to increase both the length and accuracy of sequences, reduced cost of sequencing where possible has also been desired. This has led to the development of "Next-Generation Sequencing" (NGS) techniques that have allowed affordable, routine high-throughput sequencing approaches [12, 13]. High-throughput sequencing, through parallel sequencing in which the same molecule(s) is sequenced multiple times in the same experiment, generates billions of sequence bases and leads to vast amounts of newly available data. This has in turn led to new HLA gene sequencing approaches within immunogenetics, potentially allowing for greater accuracy and coverage [8, 14–17]. These developments have great implications for both the clinical applications of HSCT as well as wider immunogenetics research. DNA sequencing is employed by HSCT to identify potential transplant donors with patients, as matching HLA is a vital component of matching potential donors to patients receiving allogeneic transplants for hematological disorders [8, 9]. The recent NGS method developments have seen a significant expansion in the IPD-IMGT/HLA Database' user base, with additional interest in the highly curated datasets that are needed for analysis of the data produced by the next-generation technologies. Historically, the data that comprised the IPD-IMGT/HLA Database has been generated by techniques with a focus on the more variable regions of the HLA molecule—specifically exons 2 and 3 of class I and exon 2 of class II. Therefore, while the database may have held a large number of polymorphic sequences, these sequences may have been limited to particular regions of the DNA. Fig. 2 shows two coverage plots for HLA-B—currently the most polymorphic HLA locus—and details genomic DNA (gDNA) sequences. The majority of coverage being across

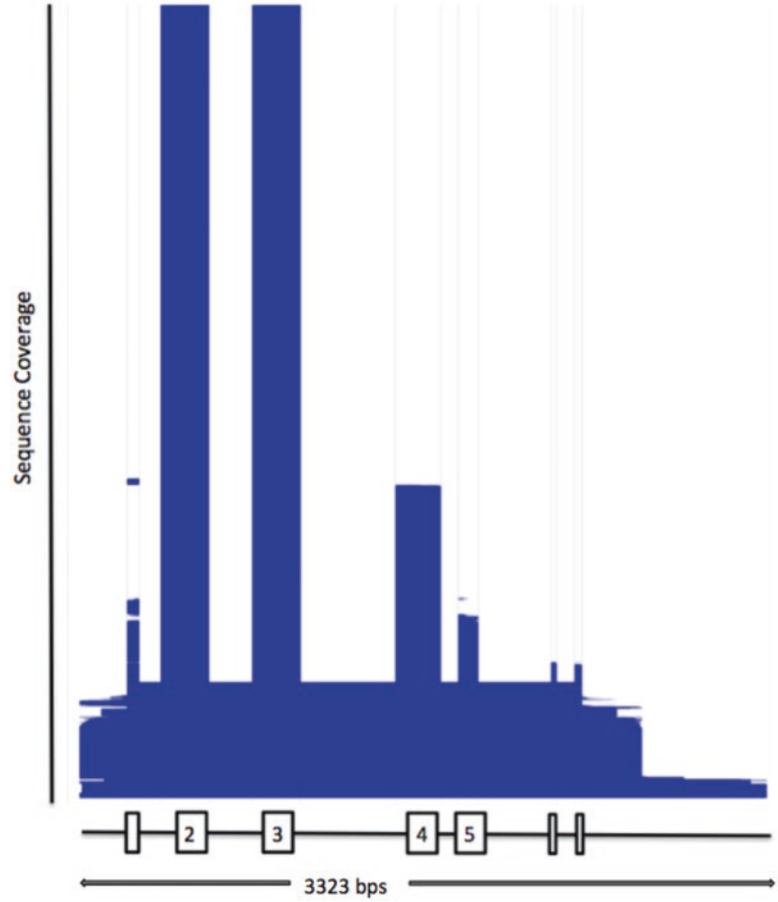


Fig. 2 Sequence Coverage of HLA-B genomic DNA sequences in the IPD-IMGT/HLA Database as of July 2017. The blue areas represent the sequenced regions. The white areas represent the unsequenced regions. The sequences are ordered by the length of sequence covered, the plots clearly show the exon 2 and 3 regions which are mandatory requirements for submission to the IPD-IMGT/HLA Database (version 3.29)

exons 2 and 3 can clearly be seen, especially given sequencing these exons constitutes the minimum requirement for acceptance. Flanking exons' coverage is far lower, with <25% of sequences containing exon 1 data and <35% for exon 4 data. Average coverage for gDNA is <10% of the alleles in a given gene.

However, more recently there is a strong trend toward the submission of genomic sequences—covering all of an allele's exons, introns, and UTR's—being provided for the IPD-IMGT/HLA Database (Fig. 3). Alongside this, a vastly increased number of total sequences can be expected to be entered into the database as the generation of long-read length sequences continues. As well as

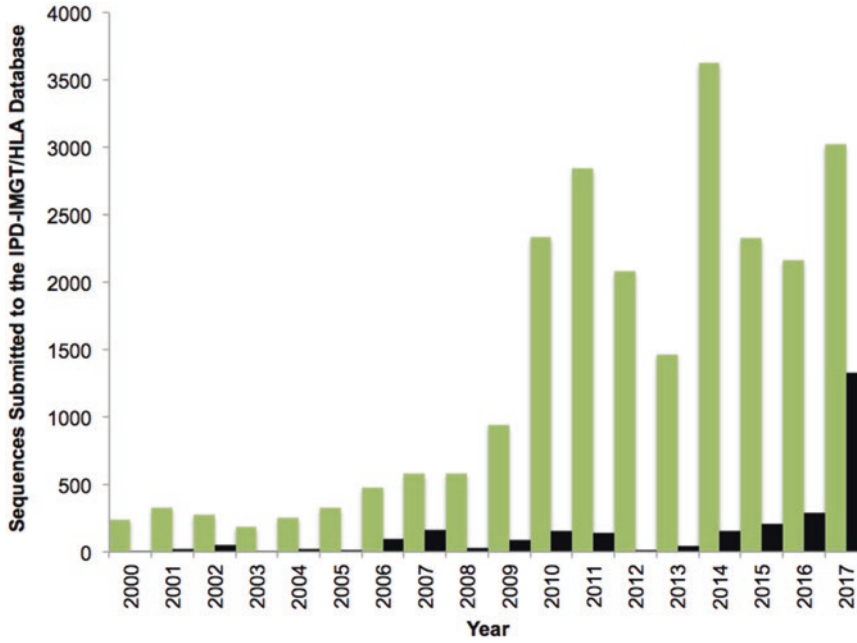


Fig. 3 Annual submissions to the IPD-IMGT/HLA Database with genomic sequences (black) against total sequences received (green). Recent years have seen a vast rise in the number of submitted genomic sequences

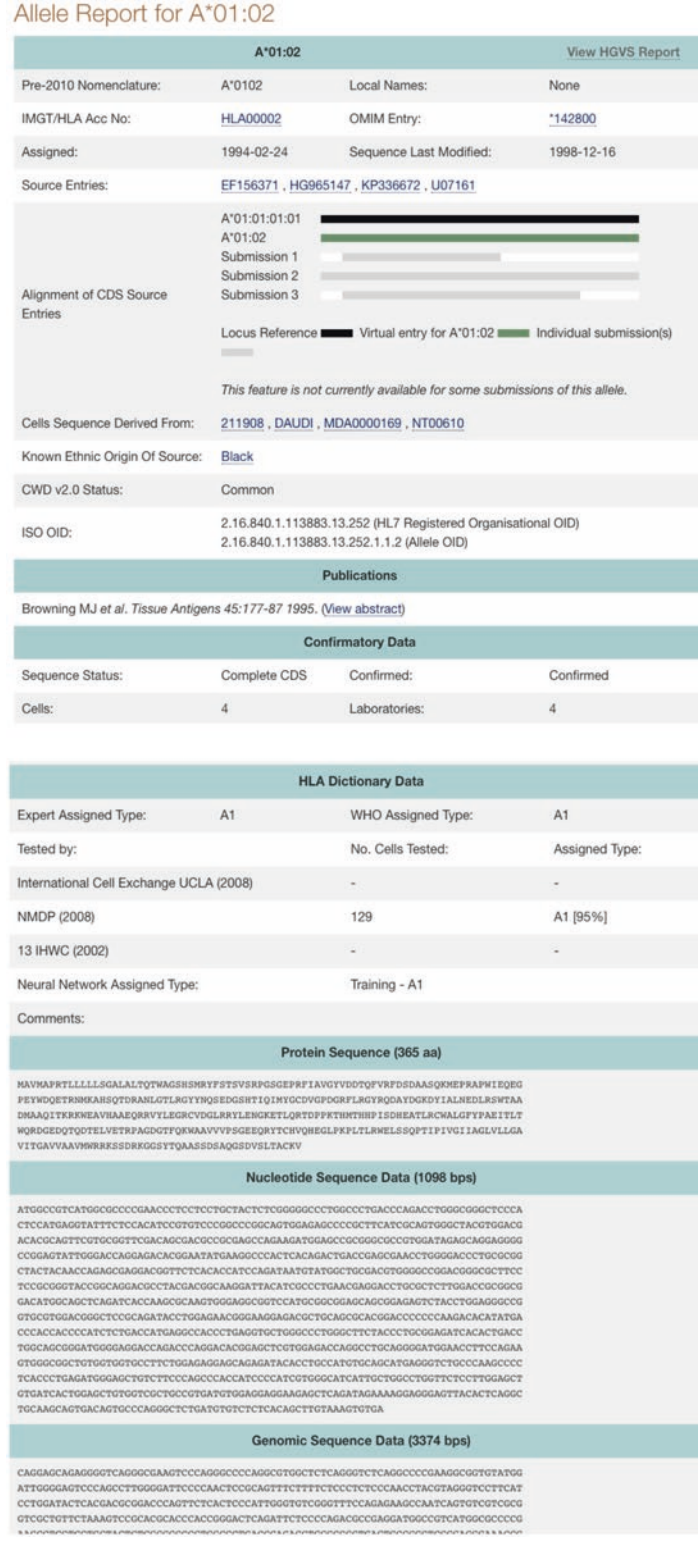
extending existing sequences and completing missing gaps in certain sequences, there is likely to be an increase in the number of novel sequences. Despite this, the full impact of this increase in data upon the clinical setting and its utility is not currently known. Given current practice, matching—using exons 2 and 3 for class I and exon 2 only for class II—is unlikely to result in any significant impact regardless of the increasing volume of genomic data. This increase in data likely requires analysis of polymorphisms beyond these limited regions, to ultimately reveal the clinical importance of full-length genomic sequencing. Given the changes implemented in the 2010 HLA nomenclature update [18], it is likely the suitability of the database, as well as the nomenclature itself, will maintain its fitness regardless of the expected voluminous influx of sequences. The IPD-IMGT/HLA Database itself, as well as the underlying infrastructure which maintains it, is currently under review, and new tools and analysis pipelines are in development to assist NGS-generated sequence curation. NGS-generated data requires an accurate reference database in order to fully assess sequence quality, be it the long reads associated with Pacific Biosciences SMRT technology [12, 13] or the shorter tiled reads associated with Roche 454 [17], Illumina [16], ION-PGM and Ion-Torrent [15]. HLA's hyper-polymorphic nature can result in difficulty assigning phase and implementing sequencing analysis, due to highly variable or even unavailable reference sequences. A

majority of the underlying work on and with the IPD-IMGT/HLA Database is currently focused on addressing the challenge of lacking detailed referential data.

As the number of known alleles now exceeds 17,000, obtaining source material for all of these in order to procure a full-length genomic sequence may be impossible. Rather, the likely increase of NGS data will start to populate areas of the database that were previously missing. Currently, the IPD-IMGT/HLA Database will only accept full-length genomic sequences where all tiled reads phasing can be proven. Without accurate phasing, the accuracy of full-length sequences assembled from fragments becomes a concern: fragment assemblies based upon regions with missing information or low coverage in the reference database may be vulnerable to low accuracy or incorrect phasing. Given the hyper-polymorphic nature of key genes sequenced for HSCT, SNPs that may be crucial to full-length sequence assembly may not be available from the database, likely implying depth of coverage is as essential to accurate assembly as coverage alone. Despite the IPD-IMGT/HLA Database containing genomic sequences from all the main HLA class I serological groups (HLA-A, -B and -C), coverage for HLA class II genes is far lower and this may impact the accuracy of any assembled sequence.

5 Alternative Descriptions for HLA Allele Variation

Allelic variation for each gene in the IPD-IMGT/HLA Database is displayed against a reference sequence. These differences can be visualized using the sequence alignment tool. The Human Genome Variation Society (HGVS) reporting system [19, 20] is another method for reporting alleles. HGVS descriptions compare each allele to a reference sequence and describe the change(s) instead of a separate sequence for both alleles. HGVS reporting for all alleles was introduced to the IPD-IMGT/HLA Database in 2014, as part of the main website's allele report page (Figs. 4 and 5). This report lists each allele's observed variation against both the GRC Reference Sequence (GRCh38/hg38), and the WHO Nomenclature Committee for Factors of the HLA System approved Reference Sequence [21]. It should be noted that these references often refer to various sequences, and some genes within the GRC reference sequences do not match existing sequences within the IPD-IMGT/HLA Database. The hyper-polymorphism of the HLA system is emphasized when using HGVS reporting methods. Currently, 3500 HLA-B variant sequences, which differ by at least an SNP within the gDNA sequence, are known. As such, the HGVS report for HLA-B must list over 113,629 descriptors in order to cover this polymorphism at merely the CDS level. Using B*15:01:01:01 as an alternative reference sequence to



Allele Report for A*01:02

A*01:02			View HGVS Report
Pre-2010 Nomenclature:	A*0102	Local Names:	None
IMGT/HLA Acc No:	HLA00002	OMIM Entry:	*142800
Assigned:	1994-02-24	Sequence Last Modified:	1998-12-16
Source Entries:	EF156371 , HG965147 , KP336672 , U07161		
Alignment of CDS Source Entries	A*01:01:01:01 A*01:02 Submission 1 Submission 2 Submission 3 Locus Reference Virtual entry for A*01:02 Individual submission(s) *This feature is not currently available for some submissions of this allele.*		
Cells Sequence Derived From:	211908 , DAUDI , MDA0000169 , NT00610		
Known Ethnic Origin Of Source:	Black		
CWD v2.0 Status:	Common		
ISO OID:	2.16.840.1.113883.13.252 (HL7 Registered Organisational OID) 2.16.840.1.113883.13.252.1.1.2 (Allele OID)		

Publications

Browning MJ *et al. Tissue Antigens 45:177-87 1995.* (View abstract)

Confirmatory Data

Sequence Status:	Complete CDS	Confirmed:	Confirmed
Cells:	4	Laboratories:	4

HLA Dictionary Data

Expert Assigned Type:	A1	WHO Assigned Type:	A1
Tested by:		No. Cells Tested:	Assigned Type:
International Cell Exchange UCLA (2008)		-	-
NMDP (2008)		129	A1 [95%]
13 IHWC (2002)		-	-
Neural Network Assigned Type:		Training - A1	
Comments:			

Protein Sequence (365 aa)

```
MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFSTSVSRPGSGEPRFIAVGYVDDTQFVRFDSDAASQKMEPRAPWIEQEG
PEYWDQETRNMKAHSQTDRANLGTLRGYYNQSEDGSHTIQIMYGCDVGPDGRFLRGYRQDAYDGKDYIALNEDLRSWTAA
DMAAQITKRKWEAVHAAEQRRVYLEGRCVDGLRRYLENGKETLQRTDPPKTHMTHHPISDHEATLRCWALGFYPAEITLT
WQRDGEDQTQDTELVETRPAGDGTFQKWAAVVVPSGEEQRYTCHVQHEGLPKPLTLRWELSSQPTIPIVGIIAGLVLLGA
VITGAVVAAVMWRRKSSDRKGGSYTQAASSDSAQGSDVSLTACKV
```

Nucleotide Sequence Data (1098 bps)

```
ATGGCCGTCATGGCGCCCCGAACCCTCCTCCTGCTACTCTCGGGGGCCCTGGCCCTGACCCAGACCTGGGCGGGCTCCCA
CTCCATGAGGTATTTCTCCACATCCGTGTCCCGGCCCGGCAGTGGAGAGCCCCGCTTCATCGCAGTGGGCTACGTGGACG
ACACGCAGTTCGTGCGGTTCGACAGCGACGCCGCGAGCCAGAAGATGGAGCCGCGGGCGCCGTGGATAGAGCAGGAGGGG
CCGGAGTATTGGGACCAGGAGACACGGAATATGAAGGCCCACTCACAGACTGACCGAGCGAACCTGGGGACCCTGCGCGG
CTACTACAACCAGAGCGAGGACGGTTCTCACACCATCCAGATAATGTATGGCTGCGACGTGGGGCCGGACGGGCGCTTCC
TCCGCGGGTACCGGCAGGACGCCTACGACGGCAAGGATTACATCGCCCTGAACGAGGACCTGCGCTCTTGGACCGCGGCG
GACATGGCAGCTCAGATCACCAAGCGCAAGTGGGAGGCGGTCCATGCGGCGGAGCAGCGGAGAGTCTACCTGGAGGGCCG
GTGCGTGGACGGGCTCCGCAGATACCTGGAGAACGGGAAGGAGACGCTGCAGCGCACGGACCCCCCCAAGACACATATGA
CCCACCACCCCATCTCTGACCATGAGGCCACCCTGAGGTGCTGGGCCCTGGGCTTCTACCCTGCGGAGATCACACTGACC
TGGCAGCGGGATGGGGAGGACCAGACCCAGGACACGGAGCTCGTGGAGACCAGGCCTGCAGGGGATGGAACCTTCCAGAA
GTGGGCGGCTGTGGTGGTGCCTTCTGGAGAGGAGCAGAGATACACCTGCCATGTGCAGCATGAGGGTCTGCCCAAGCCCC
TCACCCTGAGATGGGAGCTGTCTTCCCAGCCCACCATCCCCATCGTGGGCATCATTGCTGGCCTGGTTCTCCTTGGAGCT
GTGATCACTGGAGCTGTGGTCGCTGCCGTGATGTGGAGGAGGAAGAGCTCAGATAGAAAAGGAGGGAGTTACACTCAGGC
TGCAAGCAGTGACAGTGCCCAGGGCTCTGATGTGTCTCTCACAGCTTGTAAAGTGTGA
```

Genomic Sequence Data (3374 bps)

```
CAGGAGCAGAGGGGTCAGGGCGAAGTCCCAGGGCCCCAGGCGTGGCTCTCAGGGTCTCAGGCCCCGAAGGCGGTGTATGG
ATTGGGGAGTCCCAGCCTTGGGGATTCCCCAACTCCGCAGTTTCTTTTCTCCCTCTCCCAACCTACGTAGGGTCCTTCAT
CCTGGATACTCACGACGCGGACCCAGTTCTCACTCCCATTGGGTGTCGGGTTTCCAGAGAAGCCAATCAGTGTCGTCGCG
GTCGCTGTTCTAAAGTCCGCACGCACCCACCGGGACTCAGATTCTCCCCAGACGCCGAGGATGGCCGTCATGGCGCCCCG
```

Fig. 4 IPD-IMGT/HLA Allele Report. The figure shows an example of an allele report, with vital information listed, including ENA source entry IDs, cell information, and various relevant sequences in full

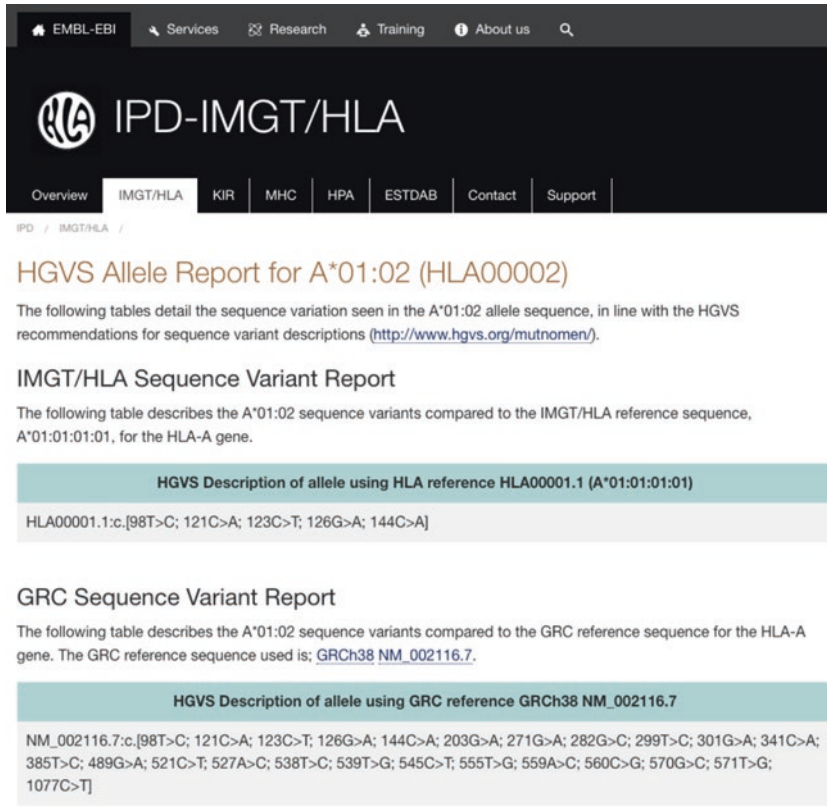


Fig. 5 IPD-IMGT/HLA HGVS Variant Report. The figure shows an example of an allele report which utilizes the HGVS variant reporting format to describe the allele rather than display the entire sequence. The variations are described in relation to the WHO HLA Reference Sequence and a GRC reference sequence

B*07:02:01:01 reduces the number of descriptors by 23%. This suggests the HGVS descriptors are less reliable indicators of variation levels as they are easily influenced by the choice of reference sequence. Continued development of these descriptors, their use, and publication will allow linking the descriptors catalogued by the IPD-IMGT/HLA Database to other reference databases. Particularly, this is of interest for NGS analysis—for example, the establishment of cross references between the IMGT/HLA HGVS descriptors and the rs# used in dbSNP [22] is an increasingly common request from new database users. However, until dbSNP contains all polymorphisms, this will not be completely possible.

6 Tools Available at IPD-IMGT/HLA

The IPD-IMGT/HLA Database provides a diverse array of tools for analysis of HLA sequences. These tools were either custom written for the IPD-IMGT/HLA Database, or incorporated

from the existing toolset described on the EBI website [23, 24]. The latter includes tools for creating user-defined sequence alignments at protein, cDNA, and gDNA level. Users may also perform queries for specific HLA alleles; outputs provide access to descriptive information on any HLA allele—ranging from the ethnicity of the source material donor and database cross-references to seminal publications. Such information is also available through integration with the European Bioinformatic Institute (EBI) EB-Eye search engine [25].

Both protein and nucleotide sequence data from IPD are incorporated into the EBI search toolset, including the FASTA suite of programmes [26] and BLAST [10] and are downloadable from the EBI's File Transfer Protocol (FTP) directory in a variety of commonly used formats such as FASTA, MSF, and PIR.

7 IPD-IMGT/HLA as a Model for Other Highly Polymorphic Gene Systems

The publication of the HLA Nomenclature through the IPD-IMGT/HLA Database has provided a model for numerous groups analyzing and curating MHC sequence variation. MHC sequences from various species have been reported [27–41] across different formats and nomenclature systems used in identifying and naming new genes and alleles in each species [42]. This disparate approach has led to numerous, often unrelated, individual studies, leading to potentially conflicting nomenclature.

Non-human MHC gene and allele nomenclature has historically been overseen either by formal nomenclature committees set up by the International Society for Animal Genetics (ISAG) [43] or by informal groups generating sequences. The Comparative MHC Nomenclature Committee now oversees this work, supported by ISAG and the International Union of Immunological Societies (IUIS) and Veterinary Immunology Committee (VIC) [27]. Given the high degree of similarity seen in MHC sequences across numerous species [44], a consistent curation methodology, naming and publication strategy is recommended. A central resource that facilitates further research of the MHC can be developed through bringing together the work from different nomenclature committees [45]. The IPD-MHC database in its earliest incarnation involved work from groups focusing on non-human primates (NHP) [30], canids (DLA) [34, 35] and felids (FLA) [46] as well as all data previously available in the IMGT/MHC Database [47]. Since that point, the IPD-MHC Database was expanded to accommodate bovin (BoLA) [32], equid (ELA) [48], salmonid [37], murid (RT1) [31], ovid (OLA) [28], and suid (SLA) [33] sequences.

In 2015, a Biotechnology and Biology Sciences Research Council (BBRSC) Bioinformatics and Biological Resource (BBR)

grant was awarded with the aim of updating and expanding the IPD-MHC database to include even more taxonomic groups of economic and scientific interest. This led to the 2016 release of the IPD-MHC Database version 2.0 [49]. The database was completely reorganized in order to provide a centralized resource, allowing a convenient and highly accurate comparison of data. To this end, a new set of tools was developed for the intra- and inter-species analysis. In particular, the multi-locus alignment allows a real time comparison of loci from different species for the first time, and the download of aligned sequences for further study and analysis. Furthermore, novel online submission tools have allowed the database to sensibly grow since its release, with an increasing annual rate of >10%. In turn, this had led to frequent publications reporting updates or changes to the nomenclature.

The model set by the IPD-IMGT/HLA Database has also been applied beyond the MHC—such as within the IPD-KIR Database. KIR genes are members of the immunoglobulin super family (IgSF) and they are highly polymorphic at both allelic and haplotypic levels [50], and are composed of two or three Ig-domains: a transmembrane region and cytoplasmic tail, which may be short (activatory) or long (inhibitory). Given the complexity in KIR regions and sequences, the KIR Nomenclature Committee was established in 2002, in order to undertake the naming of human KIR allele sequences. 2003 saw the publication of the first KIR Nomenclature report [51], coinciding with the first IPD-KIR database release. The initial release saw 89 human KIR alleles named officially, and as of January 2014, there are now over 600 alleles, coding for over 320 unique KIR protein sequences. Further information on the content of the differing IPD projects can be found in Table 1.

Table 1
Composition of the five databases that comprise the IPD project

Project	Description	Species	Genes	Sequences
IPD-IMGT/HLA	Human major histocompatibility complex and related genes	1	37	17,166
IPD-MHC	Non-human major histocompatibility complex	75	576	8488
IPD-KIR	Human killer-cell immunoglobulin-like receptors	1	16	907
IPD-HPA	Human platelet antigens	1	6	22
IPD-ESTDAB	The European searchable tumour line database (ESTDAB) database and cell Bank contains 211 cells characterized for 240 markers	1	NA	NA

8 IPD-HPA and IPD-ESTAB

The IPD-HPA and IPD-ESTDAB projects do not possess the same structure nor tools that other IPD projects retain. The IPD-HPA Database provides a centralized repository for the relevant data, which define the human platelet antigens (HPA). Alloantibodies against human platelet antigens are involved in neonatal alloimmune thrombocytopenia, post-transfusion purpura, and refractoriness to random donor platelets. In 1990, the HPA nomenclature system was constructed [52] to attend to, and ultimately solve, issues present within the previous nomenclature. From this point, further antigens have had descriptions added, with the molecular basis of a number of them resolved and in 2003, the nomenclature was revised [53]. The European Searchable Tumour Line Database (ESTDAB) Database and Cell Bank [54, 55] enable online searches for HLA typed immunologically characterized tumor cells, as a part of the European Commission Fifth Framework Infrastructures Program.

9 Future Developments

The database developers and curators face the challenge of keeping pace with an ever-increasing number of submitted allele sequences—exemplified by an average annual increase of database-held sequences of 29% in recent years. Therefore, new sequence visualization tools must be in constant development—all while maintaining the standards set for the quality of the HLA sequences and nomenclature and their presentation to the research community. NGS development techniques may allow the potential to phase polymorphisms across genes, rather than within the individual genes. Thus, consideration needs to be made as to how this data is made available by the database, and possible nomenclature implementation for these haplotypes. Alternatively, it is possible existing reporting formats [1] could also be reutilized to present this data, with the new variant reports using HLA Nomenclature and allele designations as core components. Continual development of tools, as well as refinement of existing tools, is the consistent target of the database, and all IPD projects.

10 Conclusions

The IPD-IMGT/HLA Database provides a centralized resource for those interested, clinically or scientifically, in the MHC system. The Database and accompanying tools allow the study of these alleles from a single site on the World Wide Web. It aids in the

management and development of nomenclature, providing a continuing and updated resource for the Nomenclature Committees. The challenges for the Database are to keep up with the increase in submitted sequences, keep pace with the increasing difficulties in performing analyses on larger datasets, and develop new tools for the visualization of sequences while maintaining the high standards set in the presentation and quality of the sequences and nomenclature published to the research community.

11 Licensing

The IPD-IMGT/HLA Database is covered by the Creative Commons Attribution-NoDerivs Licence, which is applicable to all copyrightable parts of the database, which includes the sequence alignments. This means that users are free to copy, distribute, display, and make commercial use of the databases in all jurisdictions provided they give the appropriate credit.

Support for the database is requested from commercial users of the database resources. If users intend to distribute a modified version of the data in any form, then they must ask for permission; this can be done by contacting hla@alleles.org, for further details of how modified data can be reproduced.

12 Availability

IPD Homepage: http://www.ebi.ac.uk/ipd/.

IPD FTP Site: ftp://ftp.ebi.ac.uk/pub/databases/ipd/.

IPD-IMGT/HLA Homepage: http://www.ebi.ac.uk/ipd/imgt/hla/.

IPD-IMGT/HLA FTP Site: ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/.

Contact: hla@alleles.org.

Acknowledgments

We would like to acknowledge the work of all the individual nomenclature committees. We would also like to acknowledge the support provided by the European Molecular Biology Laboratory's European Bioinformatics Institute—in particular, Paul Flicek—which allows the IPD project to be hosted within the EBI infrastructure.

We also recognize the work of Libby Guethlein and Peter Parham at Stanford University Medical School on IPD-KIR, IPD-MHC and IPD-KIR, IPD-HLA respectively and Jeff Miller and

Sarah Cooley at the University of Minnesota on IPD-KIR, as well as the individual IPD-MHC Nomenclature committees and curators for their work in collaboration with the IPD-MHC Database.

The authors would like to thank Todd Peterson of the Be The Match Foundation, for his work in securing on going funding for the database. We would like to thank all of the individuals and organizations that support our work financially.

Funding

European Commission within the Fifth Framework Infrastructures program [QLRI-CT-2001-01325 to IPD projects for IPD-ESTDAB]; National Institutes of Health [NIH/NCI P01 111412 to IPD projects for IPD-ESTDAB]. International Union of Immunological Societies (IUIS) for KIR nomenclature through the IUIS KIR Nomenclature Committee and MHC Nomenclature by the International Society for Animal Genetics (ISAG) and the Veterinary Immunology Committee (VIC) [to IPD databases]. One Lamda Inc.; Histogenetics; DKMS; American Society for Histocompatibility and Immunogenetics; FujireBio; European Federation for Immunogenetics; Olerup SSP; LabCorp; Zentrales Knochenmarkspender-Register Deutschland; Lifecodes + ImmucorGamma; Illumina; Omixon Biocomputing; Roche; Anthony Nolan; Asia-Pacific Histocompatibility and Immunogenetics Association; BAG Healthcare; Be the Match Foundation; Linkage Biosciences; National Marrow Donor Program; GenDx; Imperial Cancer Research Fund (now Cancer Research UK); EU Biotech [BIO4CT960037; all to the IPD-IMGT/HLA Database project].

References

1. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG (2015) The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 43(Database issue):D423–D431. https://doi.org/10.1093/nar/gku1161
2. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK et al (2004) Gene map of the extended human MHC. Nat Rev Genet 5(12):889–899. https://doi.org/10.1038/nrg1489
3. Robinson J, Guethlein LA, Cereb N, Yang SY, Norman PJ, Marsh SGE et al (2017) Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. PLoS Genet 13(6):e1006862. https://doi.org/10.1371/journal.pgen.1006862
4. Robinson J, Bodmer JG, Malik A, Marsh SGE (1998) Development of the international immunogenetics HLA database. Hum Immunol 59(S1):S17
5. Erlich HA, Opelz G, Hansen J (2001) HLA DNA typing and transplantation. Immunity 14(4):347–356
6. Opelz G, Wujciak T (1994) The influence of HLA compatibility on graft survival after heart transplantation. The collaborative transplant study. N Engl J Med 330(12):816–819. https://doi.org/10.1056/NEJM199403243301203
7. Flomenberg N, Baxter-Lowe LA, Confer D, Fernandez-Vina M, Filipovich A, Horowitz M et al (2004) Impact of HLA class I and class II high-resolution matching on outcomes of unrelated donor bone marrow transplantation: HLA-C mismatching is associated with a strong adverse effect on transplantation outcome. Blood 104(7):1923–1930. https://doi.org/10.1182/blood-2004-03-0803

8. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M et al (2007) High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood 110(13):4576–4583. https://doi.org/10.1182/blood-2007-06-097386
9. Shaw BE, Mayor NP, Russell NH, Apperley JF, Clark RE, Cornish J et al (2010) Diverging effects of HLA-DPB1 matching status on outcome following unrelated donor transplantation depending on disease stage and the degree of matching for other HLA alleles. Leuk Off J Leuk Soc Am Leuk Res Fund UK 24(1):58–65. https://doi.org/10.1038/leu.2009.239
10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
11. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947–2948. https://doi.org/10.1093/bioinformatics/btm404
12. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323(5910):133–138. https://doi.org/10.1126/science.1162986
13. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ et al (2010) Real-time DNA sequencing from single polymerase molecules. Methods Enzymol 472:431–455. https://doi.org/10.1016/S0076-6879(10)72001-2
14. Mayor NP, Robinson J, McWhinnie AJ, Ranade S, Eng K, Midwinter W et al (2015) HLA typing for the next generation. PLoS One 10(5):e0127153. https://doi.org/10.1371/journal.pone.0127153
15. De Santis D, Dinauer D, Duke J, Erlich HA, Holcomb CL, Lind C et al (2013) 16(th) IHIW: review of HLA typing by NGS. Int J Immunogenet 40(1):72–76. https://doi.org/10.1111/iji.12024
16. Lange V, Bohme I, Hofmann J, Lang K, Sauter J, Schone B et al (2014) Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics 15:63–73. https://doi.org/10.1186/1471-2164-15-63
17. Moonsamy PV, Williams T, Bonella P, Holcomb CL, Hoglund BN, Hillman G et al (2013) High throughput HLA genotyping using 454 sequencing and the Fluidigm access Array system for simplified amplicon library preparation. Tissue Antigens 81(3):141–149. https://doi.org/10.1111/tan.12071
18. Marsh SGE, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA et al (2010) Nomenclature for factors of the HLA system, 2010. Tissue Antigens 75(4):291–455. https://doi.org/10.1111/j.1399-0039.2010.01466.x
19. den Dunnen JT, Antonarakis SE (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15(1):7–12. https://doi.org/10.1002/(SICI)1098-1004(200001)15:1<7::AID-HUMU4>3.0.CO;2-N
20. Taschner PE, den Dunnen JT (2011) Describing structural changes by extending HGVS sequence variation nomenclature. Hum Mutat 32(5):507–511. https://doi.org/10.1002/humu.21427
21. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N et al (2011) Modernizing reference genome assemblies. PLoS Biol 9(7):e1001091. https://doi.org/10.1371/journal.pbio.1001091
22. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308–311
23. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J et al (2010) A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res 38(Web Server issue):W695–W699. https://doi.org/10.1093/nar/gkq313
24. McWilliam H, Valentin F, Goujon M, Li W, Narayanasamy M, Martin J et al (2009) Web services at the European bioinformatics Institute-2009. Nucleic Acids Res 37(Web Server issue):W6–W10. https://doi.org/10.1093/nar/gkp302
25. Valentin F, Squizzato S, Goujon M, McWilliam H, Paern J, Lopez R (2010) Fast and efficient searching of biological data resources--using EB-eye. Brief Bioinform 11(4):375–384. https://doi.org/10.1093/bib/bbp065
26. Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85(8):2444–2448
27. Ballingall KT (2012) Progress of the comparative MHC Committee and a summary of the comparative MHC workshops held at the 32nd ISAG, Edinburgh and the 9th IVIS, Tokyo, 2010. Vet Immunol Immunopathol 148(1–2):202–208. https://doi.org/10.1016/j.vetimm.2011.05.012
28. Ballingall KT, Herrmann-Hoesing L, Robinson J, Marsh SGE, Stear MJ (2011) A single nomenclature and associated database for alleles at the major histocompatibility complex class II DRB1 locus of sheep.

Tissue Antigens 77(6):546–553. https://doi.org/10.1111/j.1399-0039.2011.01637.x

29. Briles WE, Bumstead N, Ewert DL, Gilmour DG, Gogusev J, Hala K et al (1982) Nomenclature for chicken major histocompatibility (B) complex. Immunogenetics 15(5):441–447
30. de Groot NG, Otting N, Robinson J, Blancher A, Lafont BA, Marsh SGE et al (2012) Nomenclature report on the major histocompatibility complex genes and alleles of great ape, old and new world monkey species. Immunogenetics 64(8):615–631. https://doi.org/10.1007/s00251-012-0617-1
31. Fujii H, Kakinuma M, Yoshiki T, Natori T (1991) Polymorphism of the class II gene of rat major histocompatibility complex, RT1: partial sequence comparison of the first domain of the RT1.B beta 1 alleles. Immunogenetics 33(5–6):399–403
32. Hammond JA, Marsh SGE, Robinson J, Davies CJ, Stear MJ, Ellis SA (2012) Cattle MHC nomenclature: is it possible to assign sequences to discrete class I genes? Immunogenetics 64(6):475–480. https://doi.org/10.1007/s00251-012-0611-7
33. Ho CS, Lunney JK, Ando A, Rogel-Gaillard C, Lee JH, Schook LB et al (2009) Nomenclature for factors of the SLA system, update 2008. Tissue Antigens 73(4):307–315. https://doi.org/10.1111/j.1399-0039.2009.01213.x
34. Kennedy LJ, Altet L, Angles JM, Barnes A, Carter SD, Francino O et al (2000) Nomenclature for factors of the dog major histocompatibility system (DLA), 1998: first report of the ISAG DLA nomenclature committee. Anim Genet 31(1):52–61
35. Kennedy LJ, Angles JM, Barnes A, Carter SD, Francino O, Gerlach JA et al (2001) Nomenclature for factors of the dog major histocompatibility system (DLA), 2000: second report of the ISAG DLA nomenclature committee. Anim Genet 32(4):193–199
36. Longenecker BM, Mosmann TR (1981) Nomenclature for chicken MHC (B) antigens defined by monoclonal antibodies. Immunogenetics 13(1–2):25–28
37. Lukacs MF, Harstad H, Bakke HG, Beetz-Sargent M, McKinnel L, Lubieniecki KP et al (2010) Comprehensive analysis of MHC class I genes from the U-, S-, and Z-lineages in Atlantic salmon. BMC Genomics 11:154–171. https://doi.org/10.1186/1471-2164-11-154
38. Naessens J (1993) Leukocyte antigens of cattle and sheep. Nomenclature. Vet Immunol Immunopathol 39(1–3):11–12
39. Rodgers JR, Levitt JM, Cresswell P, Lindahl KF, Mathis D, Monaco JT et al (1999) A nomenclature solution to mouse MHC confusion. J Immunol 162(10):6294
40. Smith DM, Lunney JK, Ho CS, Martens GW, Ando A, Lee JH et al (2005) Nomenclature for factors of the swine leukocyte antigen class II system, 2005. Tissue Antigens 66(6):623–639. https://doi.org/10.1111/j.1399-0039.2005.00492.x
41. Symposium RSIV (1991) Leukocyte antigens in cattle, sheep and goats. Nomenclature. Vet Immunol Immunopathol 27(1–3):15–16
42. Klein J, Bontrop RE, Dawkins RL, Erlich HA, Gyllensten UB, Heise ER et al (1990) Nomenclature for the major histocompatibility complexes of different species: a proposal. Immunogenetics 31(4):217–219
43. Ellis SA, Bontrop RE, Antczak DF, Ballingall K, Davies CJ, Kaufman J et al (2006) ISAG/IUIS-VIC comparative MHC nomenclature committee report, 2005. Immunogenetics 57(12):953–958. https://doi.org/10.1007/s00251-005-0071-4
44. Parham P (1999) Virtual reality in the MHC. Immunol Rev 167:5–15
45. Robinson J, Mistry K, McWilliam H, Lopez R, Marsh SGE (2010) IPD--the Immuno polymorphism database. Nucleic Acids Res 38(Database issue):D863–D869. https://doi.org/10.1093/nar/gkp879
46. Drake GJ, Kennedy LJ, Auty HK, Ryvar R, Ollier WE, Kitchener AC et al (2004) The use of reference strand-mediated conformational analysis for the study of cheetah (Acinonyx jubatus) feline leucocyte antigen class II DRB polymorphisms. Mol Ecol 13(1):221–229
47. Robinson J, Waller MJ, Parham P, de Groot N, Bontrop R, Kennedy LJ et al (2003) IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res 31(1):311–314
48. Tseng CT, Miller D, Cassano J, Bailey E, Antczak DF (2010) Identification of equine major histocompatibility complex haplotypes using polymorphic microsatellites. Anim Genet 41(Suppl 2):150–153. https://doi.org/10.1111/j.1365-2052.2010.02125.x
49. Maccari G, Robinson J, Ballingall K, Guethlein LA, Grimholt U, Kaufman J et al (2017) IPD-MHC 2.0: an improved inter-species database for the study of the major histocompatibility complex. Nucleic Acids Res 45(D1):D860–D864. https://doi.org/10.1093/nar/gkw1050
50. Garcia CA, Robinson J, Guethlein LA, Parham P, Madrigal JA, Marsh SG (2003) Human KIR sequences 2003. Immunogenetics 55(4):227–239. https://doi.org/10.1007/s00251-003-0572-y

51. Marsh SGE, Parham P, Dupont B, Geraghty DE, Trowsdale J, Middleton D et al (2003) Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. Tissue Antigens 62(1):79–86
52. von dem Borne AE, Decary F (1990) Nomenclature of platelet-specific antigens. Hum Immunol 29(1):1–2
53. Metcalfe P, Watkins NA, Ouwehand WH, Kaplan C, Newman P, Kekomaki R et al (2003) Nomenclature of human platelet antigens. Vox Sang 85(3):240–245
54. Pawelec G, Marsh SG (2006) ESTDAB: a collection of immunologically characterised melanoma cell lines and searchable databank. Cancer Immunol Immunother 55(6):623–627. https://doi.org/10.1007/s00262-005-0117-3
55. Robinson J, Roberts CH, Dodi IA, Madrigal JA, Pawelec G, Wedel L et al (2009) The European searchable tumour line database. Cancer Immunol Immunother 58(9):1501–1506. https://doi.org/10.1007/s00262-008-0656-5

Chapter 4

Allele Frequency Net Database

Faviel F. Gonzalez-Galarza, Antony McCabe, Eduardo J. Melo dos Santos, Louise Takeshita, Gurpreet Ghattaoraya, Andrew R. Jones, and Derek Middleton

Abstract

The allele frequency net database (AFND, http://www.allelefrequencies.net) is an online web-based repository that contains information on the frequencies of immune-related genes and their corresponding alleles in worldwide human populations. At present, the system contains data from 1505 populations in more than ten million individuals on the frequency of genes from different polymorphic regions including data for the human leukocyte antigens (HLA) system. This resource has been widely used in a variety of contexts such as histocompatibility, immunology, epidemiology, pharmacogenetics, and population genetics, among many others. In this chapter, we present some of the more commonly used searching mechanisms and some of the most recent developments included in AFND.

Key words HLA, Polymorphisms, Frequencies, Immunogenetics, Population genetics

1 Introduction

The allele frequency net database (AFND, http://www.allelefrequencies.net) is a publicly available web-based resource dedicated to the storage of allele, haplotype, and genotype frequencies of several immune-related genes [1–3]. Since its inception, the goal of AFND has been to serve as a warehouse of frequency data sets and provide an online repository with a set of querying tools for the examination of frequencies in different worldwide human populations.

The information available on this database consists of genes principally related to the major histocompatibility complex (MHC). As described in earlier chapters, one of the main components of the MHC is the human leukocyte antigen (HLA) system, which is considered to be the most polymorphic region in the human genome [4, 5]. These genes play key roles in the immune system response, as well as being important in clinical applications such as solid organ and bone marrow transplantation [6–8]. The HLA system is also recognized for its importance in other different disciplines including

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_4,

associations in infectious diseases [9], autoimmune diseases [10], and studies of diversity in populations [11]. Additionally, there is also a growing field of study identifying associations between particular HLA polymorphisms and increased risk for adverse drug reactions (ADRs) [12–14].

With more than 12,000 allelic variants described in the IMGT/HLA database [15–17] as of Release 3.29.0 July 2017, the AFND constitutes an up-to-date data source for the examination of frequencies, confirmation of presence of HLA alleles, and, more recently, serving also as an HLA genotype raw data warehouse for further analyses. At present, individuals can also submit their raw data to the *Human Immunology* journal as a short population report [18]. This enables the data to be subjected to quality control aspects such as Hardy-Weinberg equilibrium testing and enabling meta-analysis to be run. In addition, existing data sets can be compared with other data sets at different time periods with regards to degree of resolution, new alleles being found, etc.

Although the HLA system comprises more than 20 genes, only six loci are routinely typed by laboratories, i.e., HLA-A, -B, and -C for Class I and HLA-DRB1, – DQB1 and -DPB1 for Class II. Hence, most of the data sets in AFND cover principally these genes, also known as classical HLA loci.

In addition to the searching mechanisms, the website also provides individuals with an online submission system, allowing data to be contributed by the wider research community and HLA typing laboratories.

One of the main features that distinguishes AFND from other repositories is that data sets stored have been manually curated, through a process of data validation to provide researchers with more accurate results [19]. Additionally, recent efforts have been focused on the organization of the HLA population data sets according to the quality of the data (*see* **Note 1**).

In the last year (August 2016–July 2017), AFND has been accessed by over 33 thousand users, from more than 150 countries, illustrating the relevance of AFND to the scientific community and HLA typing laboratories.

In this chapter, we described some of the most commonly used searching mechanisms and provided also a summary on the different tools that have been recently added, including the HLA-EpiDB and HLA-ADR databases (Table 1).

2 Materials

2.1 Hardware Requirements

The AFND is an online repository, as such, users are able to browse data without the need to install a particular software package.

2.2 Software Dependencies

The use of a web browser as a front-end gives the facility to users to access data from computers and/or mobile devices. Web pages

Table 1
Overview of the available tools for examining HLA populations in AFND

Name	Description	Website address	Reference
Allele frequency search	Querying tool to explore allele and phenotype frequencies in one or many populations from the given criteria	http://allelefrequencies.net/hla6006a.asp	[1]
Haplotype frequency search	Tool to consult a particular haplotype in a set of populations at two or more loci.	http://allelefrequencies.net/hla6003a.asp	[1]
Rare HLA alleles search	Section to ascertain the rarity of HLA alleles based on the confirmation from different databases and individual laboratories	http://allelefrequencies.net/hla6001a.asp	[20]
HLA-EpiDB	Database for the analysis of epitope frequencies	http://allelefrequencies.net/hlaepitopes/hlaepitopes.asp	-
HLA-ADR	Database for the analysis of HLA and adverse drug reactions	http://allelefrequencies.net/hla-adr/	[22]
Online submissions	Online submission tool for a new population	http://allelefrequencies.net/submit/Frequency.aspx	[19]

were implemented using the active server pages scripting environment, with the assistance of the JavaScript language and the asynchronous JavaScript and XML (AJAX) technology to allow simpler user interaction and improved visualizations. The database can be accessed utilizing any of the most common web browsers (i.e., Internet Explorer®, Microsoft Edge®, Mozilla Firefox®, Safari®, Opera®, and Google Chrome®).

Note: Please ensure you have the Javascript functionality enabled in your browser to guarantee the searches and tools work correctly. To check if have your Javascript enabled follow the instructions shown on the website (http://allelefrequencies.net/sysreq.asp).

2.3 Data Sets

As of August 2017, the collection of populations available on the AFND consists of 1505 population samples from 10,466,980 healthy unrelated individuals. These populations are divided into 1081 HLA populations, 240 KIR populations, 122 Cytokines populations, and 62 MIC populations. AFND receives data from three main sources: (1) data from peer-reviewed publications, (2) from populations that are analyzed at International HLA and Immunogenetics Workshops (IHWSs), and (3) submissions from individual laboratories across the world. However, the majority of the data (~80%) come from data extraction and curation by the AFND team from peer-reviewed publications. The literature review comprises not only histocompatibility and immunogenetics-related journals, but also, semi-automated methods have been established using regular structured queries of literature databases to verify other journals that may contain suitable data for inclusion. Additionally, due to the constant increase in the number of alleles identified by molecular methods, the database is periodically updated according to the official nomenclature from latest releases available on the IMGT/HLA database. At present, alleles on the website have been updated containing the most recent nomenclature guidelines allele designations.

3 Methods

3.1 Website Organization

AFND is divided into four main sections: HLA, KIR, MIC, and cytokine frequencies. Each section consists of different querying tools depending on the availability of data in each polymorphic region (Fig. 1). In this chapter, we have structured four main sections: (1) frequency searches, for example, the HLA allele frequency search and the HLA haplotype frequency search, (2) rare HLA alleles, (3) HLA-EpiDB for the analysis of epitope population coverage, and (4) HLA-ADR for the analysis of adverse drug reaction associations to some HLA alleles.

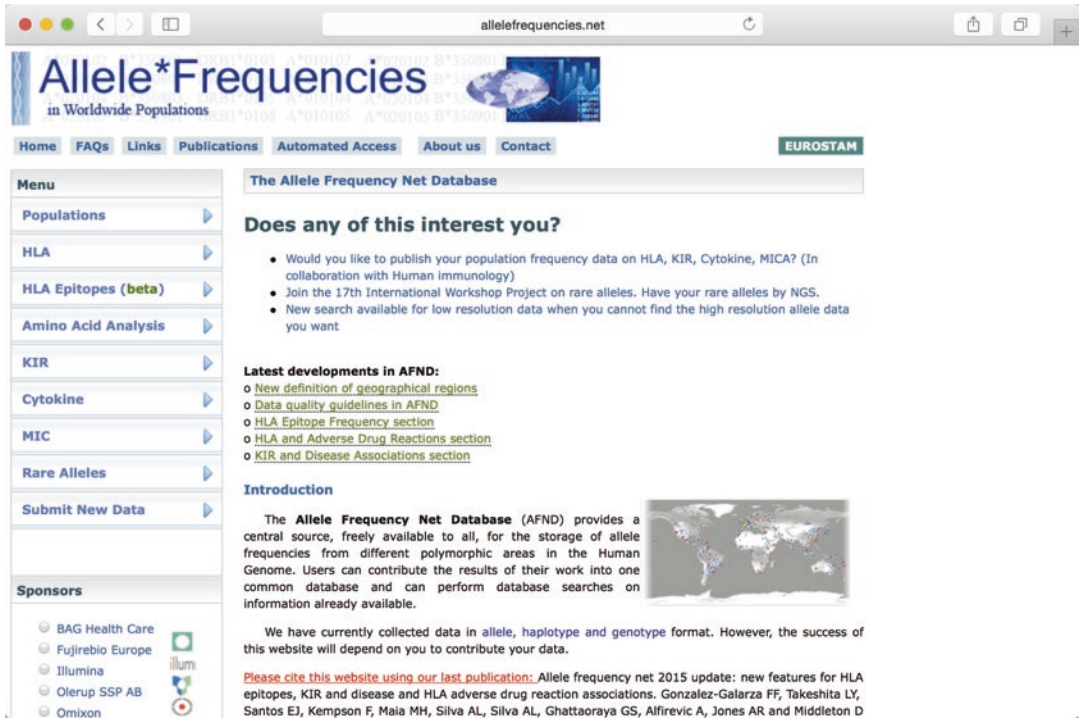


Fig. 1 Screenshot of the AFND website homepage

3.2 HLA Allele Frequency Search

The most commonly used tool within AFND is the allele frequency search (here abbreviated as AFS), with which users can examine the frequency of a particular allele in the existing population data sets, by filtering results with a set of criteria. The AFS is also available for all polymorphisms on the website. The following workflow shows a typical AFS for the HLA-DRB1*15:03 allele:

1. Go to the www.allelefrequencies.net website. Then, on the main menu, choose **HLA→HLA Allele Freq (Classical)**.
2. After this, users usually start with the selection of a locus and a particular allele to identify which populations are more likely to present the allele. For this example, choose **Locus=DRB1**, **Starting Allele=DRB1*15:03** and **Ending Allele=DRB1*15:03**.
3. To extend the searching criteria, users can select one, several, or all populations, a set or range of alleles, country, geographical region, ethnicity, and/or the year in which data was submitted. For this example, choose **Region=Sub-Saharan Africa**.
4. Then, select the option **Sort by=Allele, Highest to Lowest Frequency** to allow output records to be sorted by allele and the corresponding frequency from highest to lowest value.
5. To start the query, click on **Search** button.
   Additional options

6. Users are also able to optimize their queries to further refine data sets by selecting populations with a sample size from a range of values and/or a specific level of resolution. Moreover, for HLA, alleles can be typed at different levels of resolution (i.e., allele group, specific HLA protein, synonymous allele with a substitution within the coding region and differences in a noncoding region in that order, e.g., HLA-A*01:01:01:01). In addition, the search uses parsing methods to display all information that may be relevant to the user to ensure that high-resolution data can be retrieved when a low level resolution allele is selected. For instance, a search for the HLA-DRB1*15:03 allele will also display incidences of alleles at high resolution that start with HLA-DRB1*15:03. Furthermore, other additions include filters to search information on a specific source of data set and type of study, for example populations available in the literature oriented to anthropology studies.

Output results

7. As shown in Fig. 2a, results displayed in the search include: the allele name, name of the population, allele and/or phenotype frequency and the sample size of the population to estimate the number of individuals who carry the allele. Also, haplotype associations and graphical distribution overlaid on world maps are some of the recent options added for each record (Fig. 2b).
8. By clicking on the **Population Name** hyperlink users can access demographic details of the population in which the allele is present (Fig. 2c).
9. Finally, for export options, *see* **Note 2**.

3.3 HLA Haplotype Frequency Search

The AFND repository also includes a tool for querying haplotype frequencies (HFS) from 100,000 HLA haplotypes from around seven million individuals. At present, the collection of haplotypes consists of 456 globally distributed populations in 90 countries. The following workflow shows an example of a HFS:

1. Go to the www.allelefrequencies.net website. Then, on the main menu, choose **HLA➔HLA Haplotype Frequency Search**.
2. After this, users can select two or more alleles from two to eight routinely typed HLA loci (HLA-A, -B, -C, -DRB1, -DPA1, -DPB1, –DQA1, and -DQB1). In this example, choose **DRB1=DRB1*15:03** and **DQB1=Any DQB1**. Then, click **Search** to run the query.

Additional options

3. The program also permits the user to customise a frequency search by a particular population, country, source of data, geo-

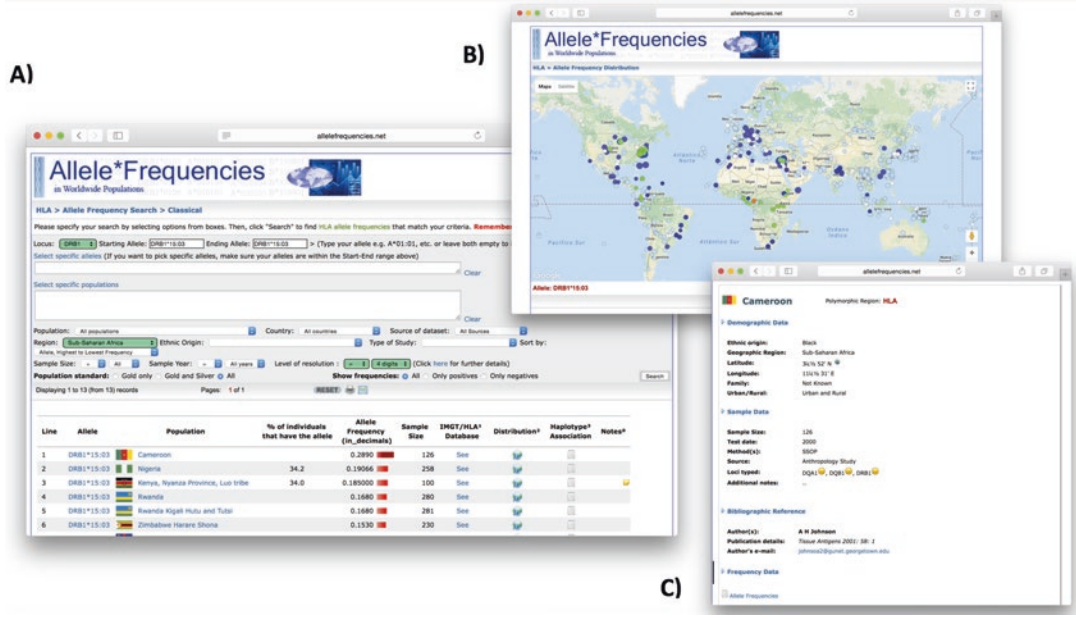


Fig. 2 Example of an allele frequency search of the DRB1*15:03 allele in AFND

graphical region, ethnicity of the individual, and number of loci tested for the haplotype.

The haplotypic information can be more useful than information only on the allele, especially in clinical applications. Therefore, this search can be used as a complement of haplotype searches performed in bone marrow and solid organ transplant registries in which, on some occasions, the information about the ethnicity of the individual is unknown.

Output results.

4. Results displayed in the search include the haplotype, name of the population, frequency, and the sample size of the population, along with an option to display frequency distribution on maps.

3.4 Rare HLA Alleles

Following the continuation of a project of the 15th and 16th IHWSs related to the rarity of specific HLA alleles, AFND has a utility that allows users to search for a particular allele and display the number of confirmations (after the initial submission of the sequence of the new allele to IMGT/HLA) submitted by different data sources [AFND, IMGT/HLA, national marrow donor program (NMDP) in the United States and individual laboratories] [20]. In this search, users are also invited to confirm an allele, which has been seen in their laboratories by providing basic information concerning the rare allele. Although a default mechanism has been set to classify the rarity of the alleles into "rare," "very rare," or "frequent," according to the number of reports (see more

about this classification in [20]), the tool also allows individuals to decide whether an allele is considered to be rare by selecting their own criteria. The following workflow shows a typical search for the HLA-DRB1*15:03 allele:

1. Go to the www.allelefrequencies.net website. Then, on the main menu, choose **Rare Alleles ➔ HLA Rare Alleles**.
2. Then, type **DRB1*15:03** in the **Search** allele field.
3. Then, click on **Search** to perform the query.

Additional options

4. Users can also filter results by locus, typing method, group identical alleles over exons 2 and/or 3, filter by the number of times an allele has been reported in a particular database (i.e., AFND, IMGT/HLA, NMDP, or individual laboratories) or the whether the allele has been reported as "common and/or well documented" (C/WD) by the American Society for Histocompatibility and Immunogenetics (HLA) [21].

Output results

5. As shown in (Fig. 3), each allele is shown with information on whether the initial sequence submitted to IMGT/HLA has been confirmed and, if so, in how many individuals, whether this allele has been found in individuals typed in NMDP or reported by individual laboratories. Most of the data from individual laboratories comes from projects conducted under

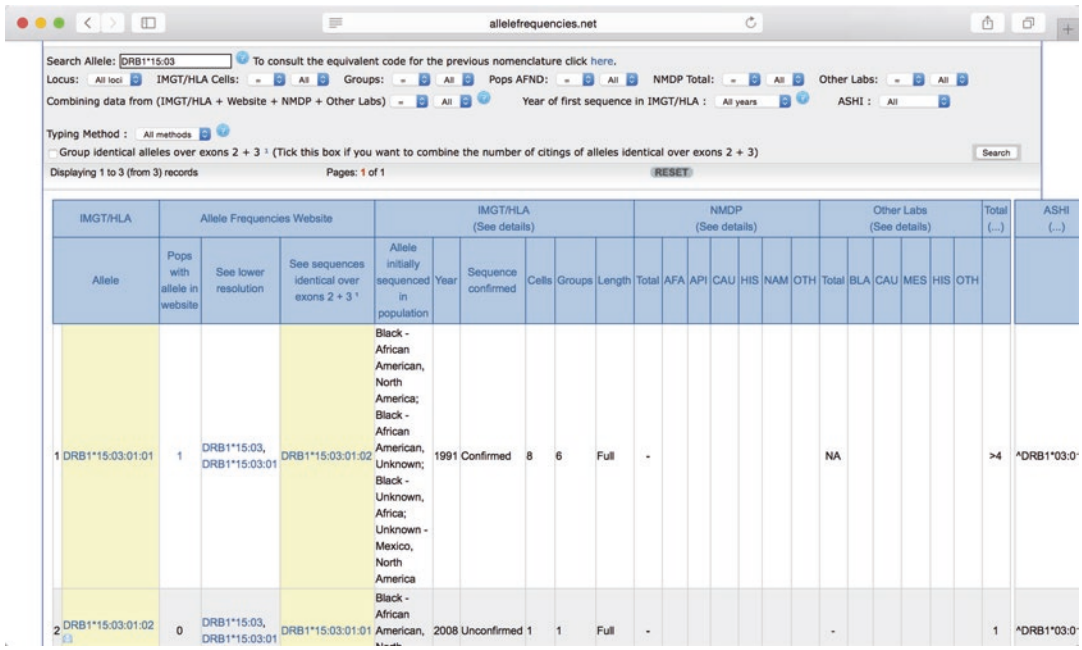


Fig. 3 Example of a rare allele search in AFND

auspices of the 15th and 16th IHWSs. Previous reports during the International Workshops have shown that around 40% of HLA alleles have never been reported in another individual, after the report of the initially sequenced sample. Thus, laboratories can use this information in estimating what allele is present when they are faced with an ambiguous combination in the HLA type. At present, AFND provides information on the country and ethnicity in which the allele is found, but in the future the intention is to be able to show rare alleles filtered by each country, by each geographical region and by ethnicity.

3.5 HLA-EpiDB

The HLA-EpiDB database is a recent development stimulated by the collection of data from the EUROSTAM project (http://eurostam.eu/). In transplantation, HLA epitope data is starting to change the current view of HLA matching, from allele matching to structural matching where epitopes are patches of polymorphic residues that can stimulate production of specific anti-HLA antibodies, a concept especially important in preventing high sensitization of transplant patients. The HLA-EpiDB section uses the nomenclature released through the HLA Epitope Registry (http://www.epregistry.ufpi.br/), which indicates the mapping from HLA allele-level nomenclature to confirmed or predicted epitopes. The following workflow may be used for a typical search in the HLA-EpiDB section:

1. Go to the www.allelefrequencies.net website. Then, on the main menu, choose **HLA➔HLA Epitopes ➔ HLA Epitopes ABC**.
2. After this, select **Locus=A+B+C** and choose for example **Ireland Northern** in the **Select specific populations** option (Fig. 4a)**.**
3. To perform the query, click on **Search**.

Additional options

4. Other options are available for the search including filtering by epitope name, position on the amino acid sequence, starting and ending position, country, geographical region, ethnic origin and sample size.

Output results

5. Figure 4b displays an example of the results including the population name, the amino acid position, the different loci selected in the search, the epitope and epitope frequency (calculated using raw data or from haplotype frequencies), the sample size and the option to visualize the graphical distribution overlaid on world maps.

3.6 HLA-ADR Database

In the field of pharmacogenomics, there are two main study approaches that are implemented when trying to determine the genetic components of HLA-induced ADRs. These are genome-wide association studies (GWAS) and case–control candidate gene

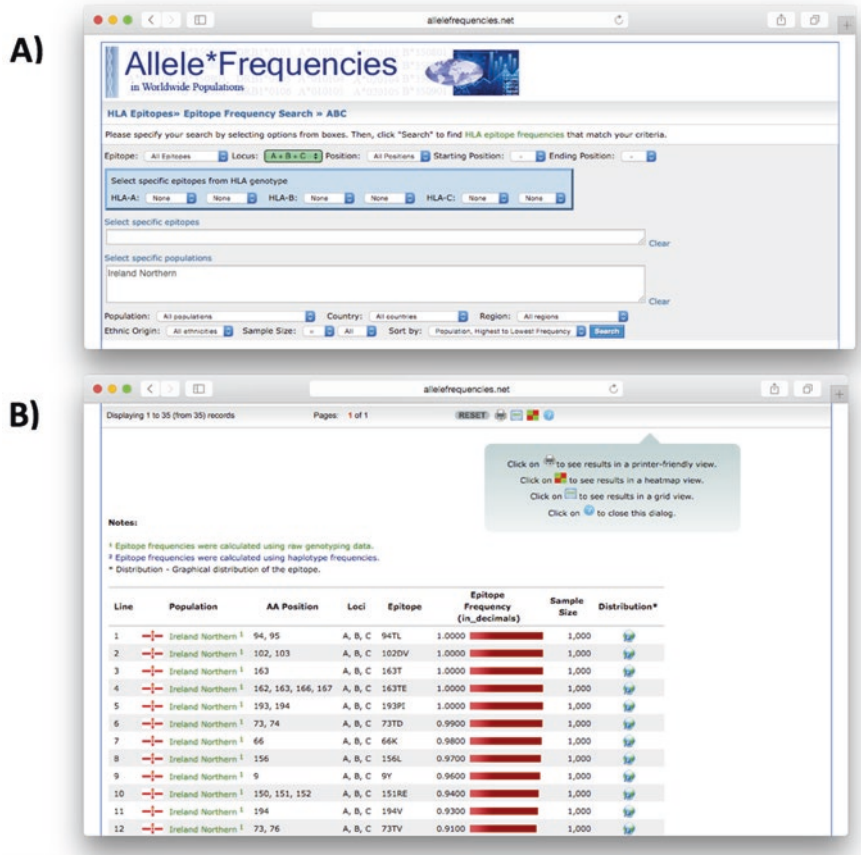


Fig. 4 Example of a search in the HLA epitopes database

studies. Both approaches have helped to identify HLA alleles associated with increased risk of developing ADRs. The following workflow shows an example of a search in the HLA-ADR database:

1. Go to the www.allelefrequencies.net website. Then, on the main menu, choose **HLA→HLA and Adverse Drug Reactions**.
2. In that page, click on the **HLA Adverse Drug Reaction Database** link.
3. The query page allows users to retrieve data via the use of dropdown filters where users may select associations with certain conditions (Fig. 5). The dropdowns are divided into three sets; with the first set, users may choose the HLA gene, a specific allele or a nonstandard allele (e.g., a serotype/antigen). In this example, select **Gene=All HLA-B**.

Note: The options within this set are mutually exclusive, meaning the user may only apply a filter from one of these although the user may use an option from this set in combination with filters from the other sets.

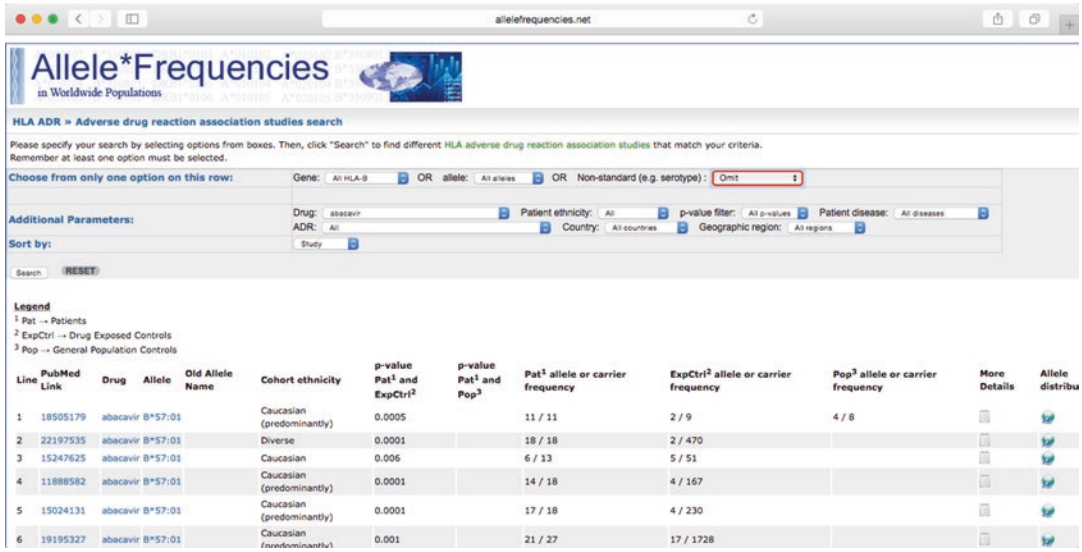


Fig. 5 Example of an adverse drug reaction association studies search

4. The second set of options allow the user to specify additional parameters, specifically: a drug, patient ethnicity, strength of the association (P values), the country/region where the study was conducted or the condition for which the patients are being treated for. The final set allows the users to choose which order they wish the data to be presented. The filters from this set can be applied in combination with each other. For this example, choose **Drug=abacavir**.

Additional options

5. In addition to the query page, an HLA-ADR report page is also provided. Here, the webpage allows the user to select a particular drug and returns all database records pertaining to that drug, which are statistically significant (the user may select the significance threshold).
6. Finally, an optional filter also enables filtering by the patient group. The returned entries are initially provided as a summary table, indicating alleles that have been reported, statistical significance values and whether the association implies that the allele is a risk or protective marker.

Output results

7. For simplicity of display, summary information about each record is provided: a link to the PubMed/Medline abstract for the original study, the drug, tested allele, the patient/control cohort ethnicity, the strength of the association and the number of patients and controls in the study cohort carrying the allele. A link is also provided ("More Details") whereby the complete data are shown for that specific association. A second

link ("Allele Distribution") connects to the main AFND site showing the worldwide distribution of the allele on a map of the world.

4 Notes

1. Recently, HLA data sets stored in AFND have been classified based on their quality (gold, silver, and bronze, abbreviated as GSB) to assist users in identifying the most suitable populations for their tasks. The "gold standard" has resulted from manual curation to identify data sets reliable in terms of sample size, summation of the allele frequencies and resolution. The gold standard includes >500 populations covering over 3 million individuals from >100 countries at one or more of the following loci: HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1 and -DRB1, with good coverage for all loci, except DPA1 (Fig. 6). (See more information about GSB in [19]).
2. At present, users can print results from all searches using the printer-friendly version available for each search, which can be used to export data sets into a tabular format. To complement frequency data in searches for further analyses, the printer-friendly option includes information of latitude and longitude if users wish to plot frequencies on maps. Other download options such as tab or comma-separated value are attended on request.

Fig. 6 HLA populations in AFND by data quality

References

1. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR (2011) Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. Nucleic Acids Res 39(Database issue):D913–D919
2. Gonzalez-Galarza FF, Takeshita LY, Santos EJ, Kempson F, Maia MH, da Silva AL, Teles e Silva AL, Ghattaoraya GS, Alfirevic A, Jones AR, Middleton D (2015) Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Res 43(Database issue):D784–D788
3. Middleton D, Menchaca L, Rood H, Komerofsky R (2003) New allele frequency database: http://www.allelefrequencies.net/. Tissue Antigens 61(5):403–407
4. The MHC sequencing consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. Nature 401(6756):921–923
5. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC Jr, Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S (2004) Gene map of the extended human MHC. Nat Rev Genet 5(12):889–899
6. Brand A, Doxiadis IN, Roelen DL (2013) On the role of HLA antibodies in hematopoietic stem cell transplantation. Tissue Antigens 81(1):1–11
7. Park M, Seo JJ (2012) Role of HLA in Hematopoietic Stem Cell Transplantation. Bone Marrow Res 2012:680841
8. Susal C, Opelz G (2013) Current role of human leukocyte antigen matching in kidney transplantation. Curr Opin Organ Transplant 18(4):438–444
9. Blackwell JM, Jamieson SE, Burgner D (2009) HLA and infectious diseases. Clin Microbiol Rev 22(2):370–385
10. Bluestone JA, Herold K, Eisenbarth G (2010) Genetics, pathogenesis and clinical interventions in type 1 diabetes. Nature 464(7293):1293–1300
11. Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA, Kimani J, Carrington M, Middleton D, Rajalingam R, Beksac M, Marsh SG, Maiers M, Guethlein LA, Tavoularis S, Little AM, Green RE, Norman PJ, Parham P (2011) The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334(6052):89–94
12. Alfirevic A, Pirmohamed M (2010) Drug-induced hypersensitivity reactions and pharmacogenomics: past, present and future. Pharmacogenomics 11(4):497–499
13. Ghattaoraya GS, Middleton D, Santos EJ, Dickson R, Jones AR, Alfirevic A (2017) Human leucocyte antigen-adverse drug reaction associations: from a perspective of ethnicity. Int J Immunogenet 44(1):7–26
14. Yip VL, Marson AG, Jorgensen AL, Pirmohamed M, Alfirevic A (2012) HLA genotype and carbamazepine-induced cutaneous adverse drug reactions: a systematic review. Clin Pharmacol Ther 92(6):757–765
15. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG (2015) The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 43(Database issue):D423–D431
16. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SG (2013) The IMGT/HLA database. Nucleic Acids Res 41(Database issue):D1222–D1227
17. Robinson J, Soormally AR, Hayhurst JD, Marsh SG (2016) The IPD-IMGT/HLA Database - New developments in reporting HLA variation. Hum Immunol 77(3):233–237
18. Mack SJ, Middleton D (2015) Introducing a new manuscript format: Enabling access to immunogenomic population data with short population reports. Hum Immunol 76(6):393–394
19. Dos Santos EJ, McCabe A, Gonzalez-Galarza FF, Jones AR, Middleton D (2016) Allele Frequencies Net Database: Improvements for storage of individual genotypes and analysis of existing data. Hum Immunol 77(3):238–248
20. Gonzalez-Galarza FF, Mack SJ, Hollenbach J, Fernandez-Vina M, Setterholm M, Kempenich J, Marsh SG, Jones AR, Middleton D, Consortium HLARA (2013) 16(th) IHIW: extending the number of resources and bioinformatics analysis for the investigation of HLA rare alleles. Int J Immunogenet 40(1):60–65
21. Mack SJ, Cano P, Hollenbach JA, He J, Hurley CK, Middleton D, Moraes ME, Pereira SE, Kempenich JH, Reed EF, Setterholm M, Smith AG, Tilanus MG, Torres M, Varney MD, Voorter CE, Fischer GF, Fleischhauer K, Goodridge D, Klitz W, Little AM, Maiers M, Marsh SG, Muller CR, Noreen H, Rozemuller EH, Sanchez-Mazas A, Senitzer D, Trachtenberg E, Fernandez-Vina M (2013)

Common and well-documented HLA alleles: 2012 update to the CWD catalogue. Tissue Antigens 81(4):194–203

22. Ghattaoraya GS, Dundar Y, Gonzalez-Galarza FF, Maia MH, Santos EJ, da Silva AL, McCabe A, Middleton D, Alfirevic A, Dickson R, Jones AR (2016) A web resource for mining HLA associations with adverse drug reactions: HLA-ADR. Database (Oxford) 2016:pii: baw069

Chapter 5

High-Resolution HLA-Typing by Next-Generation Sequencing of Randomly Fragmented Target DNA

Michael Wittig, Simonas Juzenas, Melanie Vollstedt, and Andre Franke

Abstract

PCR- or probe-based targeted capturing enables the enrichment of specific genomic loci prior to Next-Generation Sequencing (NGS). Here, we describe a probe-based protocol, which allows for high-resolution HLA typing of DNA samples by NGS. We also describe existing software tools that can be used for the subsequent HLA data analysis. Key prerequisites that warrant an accurate HLA calling are specific mappings of the sequencing reads, phasing of the mapped reads, and the possibility to perform a manual inspection/curation of the read mapping.

Key words Next-Generation Sequencing, NGS, Targeted enrichment, In-solution capture, Sequencing, HLA typing, HLA analysis, RNA baits

1 Introduction

The human leukocyte antigen (HLA) complex contains the most polymorphic genes in the human genome. Classically, the gold standard for HLA typing was Sanger sequencing of a limited amplicon repertoire. High-quality NGS-based methods also rely on the targeted enrichment of the relevant HLA loci, either by PCR- or bait-based capturing [1]. The here-described targeted HLA typing approach (Fig. 1) consists of three main steps. First, the enrichment of the target DNA, which is then followed by the sequencing of randomly fragmented DNA and finally the data analysis. The enrichment of the target can be performed by bait-based capturing, long-range PCR, or even shorter PCR amplicons that span the region of interest (overlapping and/or nonoverlapping). The bait-based method, which we describe here, comprises a library of 120 bp long biotinylated RNA sequences (also referred to as baits) that reverse complement to the target sequence. By hybridization

Michael Wittig and Simonas Juzenas contributed equally to this work.

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_5,

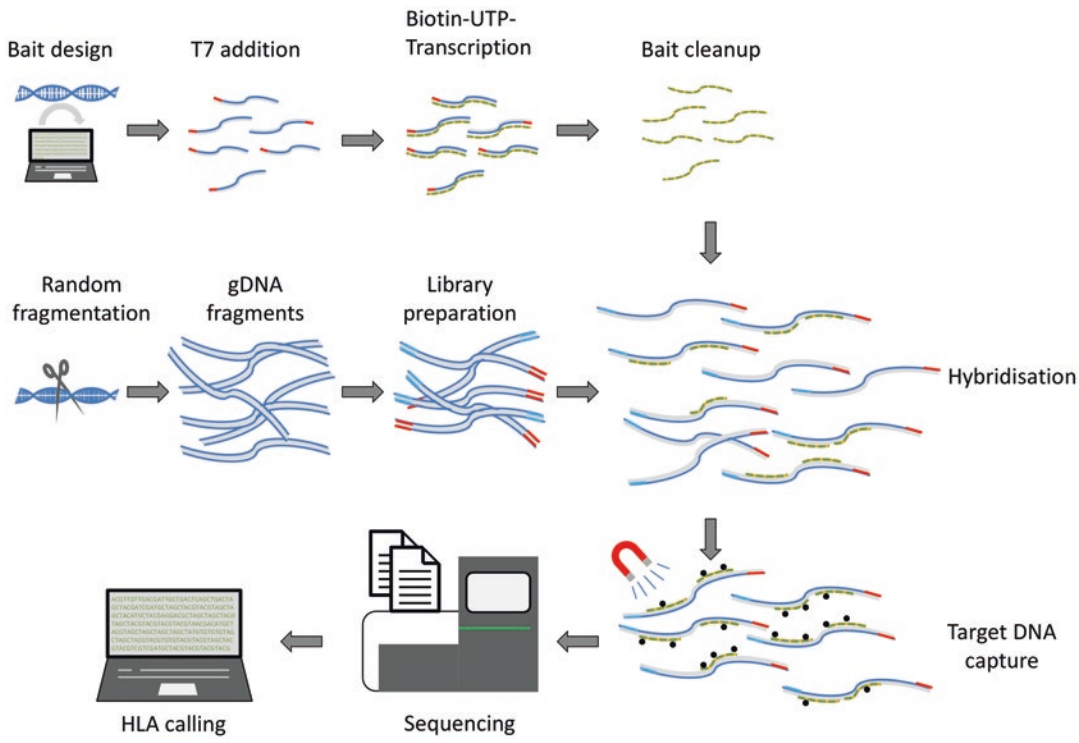


Fig. 1 *Workflow.* The workflow starts with two independent parts, the bait synthesis and the library preparation. For the bait synthesis a tiling of the target sequence is performed, which means all target sequence is fragmented into 120 bp parts. This can be performed in overlapping or nonoverlapping manner. For the HLA design nonoverlapping tiling was performed as the template sequence was a collection of all known HLA allele sequences, which is already a very redundant template. Order an oligo pool based on the design and attach a T7-promotor-binding site to the oligo pool. Perform a Biotin-UTP-transcription with these T7-added oligos and clean up the RNAs which will be used for the targeted capturing later. For the gDNA perform random fragmentation followed by a library preparation of the fragmented DNA. The library preparation adds sequences to the fragment flanks, which are needed during the sequencing reaction. These sequences consist of primer-binding sites, inserts, and indices (molecular barcodes) and the so manipulated gDNA is ready for sequencing. Perform a hybridization of the RNA baits and ready-to-sequence DNA to bind the target DNA to the biotynilated RNA baits. Using Streptavidin coated magnetic beads it is possible to capture the RNA-DNA-hybrids and to get rid of the majority of unspecific DNA. The captured DNA goes on a 2nd generation sequencer which generates a pair of fastq files (sequence data plus quality data). These fastq files can be analyzed with appropriate HLA calling software

of baits and DNA, followed by purification and clean-up with Streptavidin-coated magnetic beats, it is possible to efficiently enrich the target DNA [2]. Before this capturing is performed, the DNA has to be randomly fragmented by a tagmentation enzyme [3] or mechanically. For the PCR-based method, the fragmentation is performed on the PCR products. Both approaches, PCR- or bait-based, have their advantages and disadvantages. The PCR is very specific with regards to the targeted enrichment, i.e., little off-target sequence is produced through NGS and a high coverage of

the target is warranted. However, the PCR is prone to allelic drop outs caused by, for example, SNPs at the primer locations or just by poor performing primer sets or other amplification biases [4, 5]. Many companies sell their "secret" primer pools for different HLA loci and/or allele groups and the strategy ranges from exon-wise amplicon generation up to long-range PCR of a whole gene or at least spanning as many exons as possible. The bait-based method overcomes the problem of allelic drop outs, because a few mismatches between bait and target DNA still allow a hybridization with the target [2]. Furthermore, the typical fragment length of 300–800 bp allows for the hybridization of multiple of the predesigned baits and therefore may always bind even if there is a new/unreported mutation somewhere in the fragment. Hence, a comprehensive design, based on the currently available exhaustive information of the HLA variation [6], is more robust against allelic drop outs. But this advantage comes at the price of a higher off-target DNA capture and hence off-target sequencing information since the baits also tend to capture other regions in the genome with a lower specificity [7]. Increasing the template DNA and reducing bait concentration tend to increase the capturing specificity. The randomly fragmented DNA can be sequenced on a second generation sequencer like the Illumina® HiSeq2500 or any other comparable NGS instrument. With the bait-based method and an efficient targeted enrichment (the on-target rate should be >15%, i.e., <85% of the reads are discarded), it is good practice to sequence 96 samples in parallel on a single HiSeq2500 lane employing the Illumina® v4 chemistry and a read length between 2 × 100 bp–2 × 150 bp [8]. This protocol also instructs the user how to generate biotin-labeled RNA from a DNA oligo pool. The oligo pool is a collection of sequences designed to capture the genomic target of interest that can be ordered from several existing companies. Our protocol allows for bait synthesis for ten-thousands of samples, providing nearly endless amounts of "fishing baits." Users with small and moderately sized sample sizes should consider ordering ready-to-use biotin-labeled baits from any existing provider. While this is more cost-intensive per sample, this shortens the protocol tremendously and Subheadings 2.1–2.3 and 3.1–3.3 can be skipped.

2 Materials

2.1 Custom Oligo Library Amplification (Optional) and T7 Promotor Addition (Mandatory) (See Note 1)

The listed materials are needed for the protocol described in Subheading 3. Shown companies are exemplary providers, materials may also be available from other vendors.

1. Bait design in fasta format, downloadable at GitHub: (https://github.com/MiWitt/HLAbaits/blob/master/design/probes_T7_IKMB_custom_v0.1.fasta).

2. Custom oligo pool of the bait design (see above) (CustomArray).
3. Phusion High-Fidelity PCR Kit (NEB).
4. Ethanol 80% and 70%.
5. Elution buffer or TE 0.1%.
6. AMPure® XP Beads (Beckman Coulter, Inc.).
7. TE buffer pH 8.0 (500 mL).
8. D1000 Screen Tape, Reagents, and Tape Station from Agilent (or equivalent).
9. Forward primer: 5'-CTGGGATCGCACCAGCGTGT-3'.
10. Reverse primer: 5'-CGTGGATGAGGAGCCGCAGTG-3′.
11. T7 Forward primer: 5'-GGATTCTAATACGACTCACTATAGGGATCGCACCAGCGTGT-3'.

2.2 RNA Bait Synthesis from DNA Template (Biotin-UTP Transcription)

1. The DNA oligos with T7 promotor added (from previous step).
2. HiScribe™ T7 High Yield RNA Synthesis Kit (NEB).
3. Biotin RNA Labeling Mix (Roche).
4. TURBO™ DNase (2 U/μL) (Ambion).
5. RNeasy MinElute Cleanup Kit (Qiagen).
6. SUPERaseIn RNase inhibitor (Invitrogen).

2.3 Quality Control of RNA Bait Library

1. SuperScript III First-Strand Synthesis Kit (Invitrogen).
2. PowerUp™ SYBR Green Master Mix (Applied Biosystems).
3. RNA ScreenTape.
4. RNA Screen Reagents.
5. Bait QC 3 forward: 5'-AGCGACGTGGGGGAGTAC-3'.
6. Bait QC 3 reverse: 5'-GCTGTTCCAGTACTCGGCA-5'.

2.4 gDNA Fragmentation (See Note 2)

This is the 50 μL 500 bp fragmentation protocol of the Covaris® manual.

1. Covaris S2 or E210 Focused-ultrasonicator.
2. MicroTUBE AFA Fiber Snap-Cap.
   (a) With S-Series Holder microTUBE if running S2 setup.
   (b) With Rack 24 Place microTUBE Snap-Cap if running E210 setup.
3. Or 96 microTUBE Plate and E210 setup.
4. IE-DNA intensifier (for E210).
5. Centrifuge adapter—"Fit microTUBEs in bench top micro centrifuges."

6. 100 ng–1000 ng purified gDNA (>10 kb) in Tris EDTA, pH 8.0.
7. Ethanol 100%.

2.5 Library Preparation

1. NEBNext® Ultra™ DNA Library Prep Kit for Illumina®.
2. 80% Ethanol (freshly prepared).
3. Nuclease-free Water.
4. DNA LoBind Tubes (Eppendorf).
5. AMPure® XP Beads (Beckman Coulter, Inc.).
6. NEBNext Singleplex or Multiplex Oligos for Illumina.
7. Magnetic-stand-96, Life Technologies/Thermo Fischer.
8. Lit heated thermocycler with minimal evaporation over the hybridization time of 72 h.
9. Nuclease-free tubes compatible with the thermocycler.
10. D1000 Screen Tape, Reagents, and Tape Station from Agilent (or equivalent).
11. Water bath, 65 °C.
12. Vortex mixer and tube rotator.

2.6 Targeted Enrichment

This procedure follows the MYbaits® protocol v1.3.8 from MYcroarray.

1. Biotinylated RNA bait library (self-made from Subheading 2.2/3.2 or ready-to-use biotin-labeled baits ordered with this design: https://github.com/MiWitt/HLAbaits/blob/master/design/probes_IKMB_custom_v0.1.fasta).
2. UltraPure 20xSSPE (HYB#1).
3. 500 mM EDTA (HYB#2).
4. Denhardts Solution 50 × 100 mL (HYB#3).
5. UltraPure 1% SDS solution (HYB#4).
6. 1 M NaCl; 10 mM Tris–HCl, pH 7.5; 1 mM EDTA (Binding Buffer).
7. Tris Hydrochloride 1 M pH 7 (Neutralization Buffer).
8. 1× SSC, 0.1% SDS (Wash Buffer #1).
9. 0.1× SSC, 0.1% SDS (Wash Buffer #2).
10. Human cot-1 DNA 1 μg/μL (BLOCK #1).
11. Salmon Sperm DNA solution 1 μg/μL (BLOCK #2).
12. SuperaseIn 20 U/μL (RNase Block).
13. Nuclease-Free Water (not DEPC-Treated).
14. 100 μM TruSeq-dual-Ampl. (Illumina®).
15. 100 μM TruSeq-ind-Ampl. (Illumina®).

16. 100 µM TruSeq-dual-I8-block (BLOCK #3) (Illumina®).
17. 100 µM TruSeq-ind-I8-block (BLOCK #3) (Illumina®).
18. Dynabeads MyOne Streptavidin C1.
19. High Sensitivity D1000 Screen Tape.
20. High Sensitivity D1000 Reagents.
21. Herculase II Fusion DNA Polymerase (Agilent Technologies).
22. Post capture amplification primers:
    (a) TruSeq-dual 5′-AATGATACGGCGACCACCGAGATCTACAC-3'.
    (b) TruSeq-ind 5′-CAAGCAGAAGACGGCATACGAGAT-5'.

2.7 Sequencing

1. An Illumina® NGS sequencer like any of HiSeq, NextSeq, MiSeq, etc.
2. An appropriate sequencing kit which works with the chosen NGS machine and which generates at least 2 × 100 bp reads.

3 Methods

3.1 Custom Oligo Library Amplification (Optional) and T7 Promotor Addition (Mandatory) (See Note 1)

In this step the working aliquot, taken from the DNA oligo pool, is amplified and the T7 promotor is added via the PCR reaction to allow for RNA transcription in the next step.

3.1.1 Custom Oligo Library Amplification (Optional, "See Note 3")

1. Mix the following components listed in Table 1 in a sterile nuclease-free tube.
2. Run a PCR using the cycling conditions of Table 2.
3. To clean up, vortex AMPure XP beads to resuspend (*see* **Note 4**).
4. Add 90 µL of resuspended beads to the PCR reactions (50 µL). Mix by pipetting up and down at least 15 times.
5. Incubate for 5 min at room temperature.
6. Quickly spin the tube and place it on a magnetic stand to separate beads from the supernatant. Incubate at room temperature until the beads completely cleared from solution. Carefully remove and discard the supernatant. **Caution**: do not discard the beads.
7. Add 200 µL of 70% ethanol to the PCR plate while in the magnetic stand. Incubate at room temperature for 1 min, and then carefully remove and discard the supernatant.

Table 1
Components for custom oligo library amplification

Component	Volume in μL
Nuclease-free water	27.0
5× Phusion HF	10.0
10 mM dNTPs	1.0
10 μM forward primer	3.75
10 μM reverse primer	3.75
Custom oligo pool from CustomArray (1:100 dilution)	2.5
DMSO	1.5
Phusion DNA polymerase	0.5
Total	*50.0*

Table 2
PCR cycle conditions for custom oligo amplification

Cycle step	Cycles	Temp in °C	Time
Initial denaturation	1	98	2 min
Denaturation	18	98	10 s
Annealing		58	30 s
Extension		72	30 s
Final extension	1	72	5 min
Hold	1	4	∞

8. Repeat the last step once more.
9. Air dry the beads for 5 min while the PCR plate is on the magnetic stand with the lid open. Remove any residue liquid with a pipette.
10. Remove the tube from the magnet. Elute DNA target from beads into 42 μL Elution Buffer (EB) or 0.1× TE. Pipet up and down at least 15 times. Quickly spin the tube and incubate at room temperature for 2–3 min.
11. Place the sample on an appropriate magnetic stand to separate beads from the supernatant. After the solution is clear, carefully

transfer 40 µL supernatant to a new PCR tube. Samples can be stored at 2–8 °C for a few days or at −20 °C for long-term storage.

12. Check the size of product on Agilent Tape Station using D1000 Tape. The length of the product should be about 171 bp (Fig. 2).

3.1.2 T7 Promotor Addition (Optimized for NEB Phusion High-Fidelity PCR Kit)

1. Mix the components of Table 3 in a sterile nuclease-free tube.
2. Run a PCR using cycling conditions of Table 4.
3. To clean up and perform size selection of the PCR product vortex AMPure XP beads to resuspend (*see* **Note 5**).
4. Add 50 µL (1×) resuspended AMPure XP beads to 50 µL DNA solution. Mix well on a vortex mixer or by pipetting up and down at least 20 times.
5. Incubate for 5 min at room temperature.
6. Place the tube on a magnetic rack to separate the beads from the supernatant. After the solution is clear, carefully transfer the supernatant to a new tube (**Caution**: do not discard the supernatant). Discard beads that contain the large fragments.
7. Add 30 µL (0.6×) resuspended AMPure XP beads to the supernatant, mix well, and incubate for 5 min at room temperature.
8. Put the tube on a magnetic rack to separate beads from the supernatant. After the solution is clear, carefully remove and discard the supernatant. Be careful not to disturb the beads that contain DNA targets (**Caution**: do not discard beads).

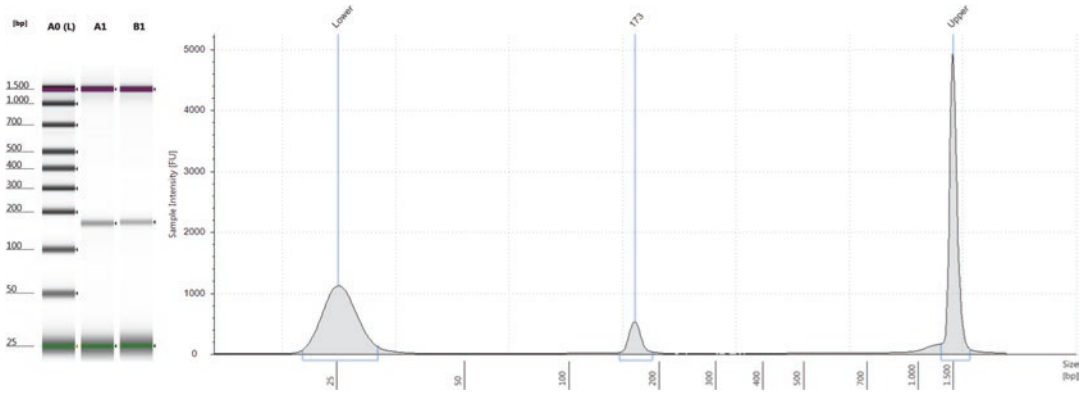


Fig. 2 *Examples of custom oligo libraries after first round PCR.* The left graph shows the ladder on the left and two lanes of oligonucleotide DNA. The right graph shows the electropherogram of lane A1. The band should be around the size of 171 bp as we added the primers to the pre-designed DNAs. No additional peak should be detectable

Table 3
Components for T7 promotor addition

Component	Volume in μL
Nuclease-free water	26.5
5× Phusion HF	10.0
10 mM dNTPs	1.0
10 μM T7 forward primer	3.75
10 μM reverse primer	3.75
First round PCR product	3.0
DMSO	1.5
Phusion DNA polymerase	0.5
Total	*50.0*

Table 4
T7 addition PCR cycle conditions

Cycle step	Cycles	Temp in °C	Time
Initial denaturation	1	98	2 min
Denaturation	20	98	10 s
Annealing		65	30 s
Extension		72	30 s
Final extension	1	72	5 min
Hold	1	4	∞

9. Add 200 μL of 80% freshly prepared ethanol to the tube while in the magnetic stand. Incubate at room temperature for 30 s, and then carefully remove and discard the supernatant.
10. Repeat **step 9** once.
11. Keeping the tube on the magnetic rack, with the cap open, air dry the beads for 5 min (*see* **Note 6**).
12. Remove the tube from the magnet. Elute DNA target from beads into 32 μL Elution Buffer (EB) or 0.1× TE. Mix well on a vortex mixer or by pipetting up and down, incubate for 2 min at room temperature.
13. Put the tube in a magnetic rack until the solution is clear, approximately 3 min. Transfer approximately 30 μL of the supernatant to a clean tube. Samples can be stored at 2–8 °C for a few days or at −20 °C for long-term storage.

14. Check the size of product on Agilent Tape Station using D1000 Tape. The length of product should be about 192 bp (Fig. 3).

3.2 RNA Bait Synthesis from DNA Template (Biotin-UTP Transcription)

In this step, the biotinylated RNA baits are synthesized using the DNA oligos produced in the last step (T7 added). The biotin RNA labeling mix incorporates biotin-16-UTP so that approximately every 20–25th nucleotide is carrying a biotin molecule. The biotin-UTP in vitro transcription is optimized using NEB HiScribe TM T7 High Yield RNA Synthesis Kit (#E2040S) and Biotin RNA Labeling Mix from Roche.

1. Thaw the necessary kit components, mix and pulse-spin in microfuge to collect solutions to bottom of tubes. Keep on ice.
2. Assemble the reaction at room temperature in the order of Table 5.
3. Incubate reaction at 37 °C for 14 h overnight.
4. To 20 μL reaction add 70 μL nuclease-free water, 10 μL of 10× Reaction Buffer, and 2 μL of RNase-free TURBO TM DNase, mix and incubate at 37 °C for 15 min.
5. Clean up RNA baits using Qiagen RNeasy MinElute Cleanup Kit according to the manufacturer's protocol. Split the reaction volume and use 50 μL per column. Elute twice in 15 μL and 10 μL of sterile nuclease-free water. Total elute volume is 25 μL.
6. Add 1 μL of SUPERaseIn RNase inhibitor into eluted RNA baits.

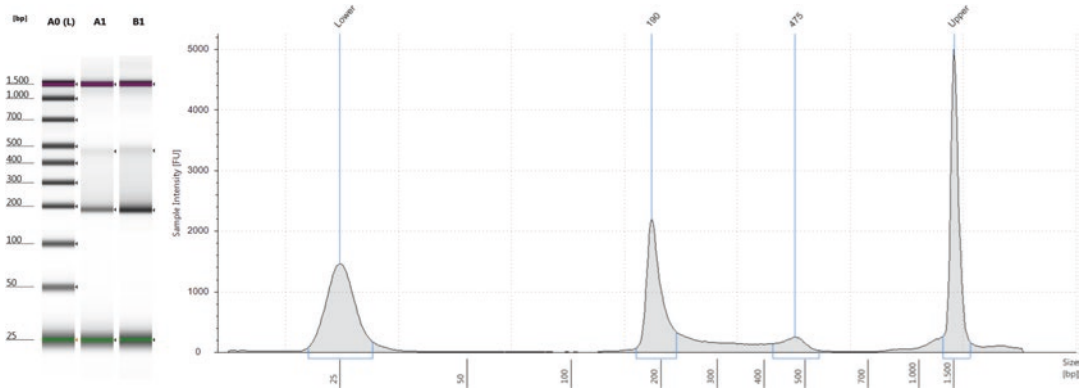


Fig. 3 *Examples of custom oligo libraries after T7 promotor addition.* The left graph shows the ladder on the left and two lanes of fragmented DNA to which the T7 promotor got attached. The right graph shows the electropherogram of the lane A1. The peak at 190 bp is the expected product and the lower and upper marker peaks are at 25 bp and 1500 bp respectively. The DNA from 200 bp up to 500 bp, with the small peak at 475 bp, is unexpected product

Table 5
RNA bait synthesis pipetting scheme

Component	Volume in μL
T7 added PCR product (Subheading 3.1.2)	9.0
10× reaction buffer	1.5
Biotin RNA labeling mix (Roche)	8.0
Enzyme mix	1.5
Total	*20.0*

7. Check the size distribution on Agilent Tape Station. The highest peak should be between 160 and 180 bp (Fig. 4). The concentration of produced baits should range from 100 to 240 ng/μL.
8. Adjust the concentration of baits to 100 ng/μL with very clean nuclease-free water (use separate tube of nuclease-free water only for this purpose). The concentration of 100 ng/μL is considered as stock solution of HLA RNA baits.

3.3 Quality Control of RNA Bait Library

To verify the success of the afore-described biotinylated RNA synthesis a quantification using real time PCR is performed. The primers used for that step are designed for the oligo-pool from Subheading 2.1 and target as many baits as possible. Eventually, only a small fraction of baits will be quality controlled, assuming that the transcription of the other biotinylated baits had the same performance. Quality assessment of custom RNA baits (Subheading 3.2) is performed by SYBR Green RT-qPCR with pre-designed primers (Subheading 2.3) for targeted sequences in predesigned HLA baits (Subheading 2.1 bait design). The protocol is optimized using Invitrogen SuperScript III First-Strand Synthesis Kit.

3.3.1 Reverse Transcription (cDNA Synthesis)

1. Mix and briefly centrifuge each component of kit before use and combine the components of Table 6 in a 0.2 mL tube.
2. Incubate the tube at 65 °C for 5 min, then place on ice for at least 1 min.
3. Prepare the cDNA Synthesis Mix as described in Table 7, adding each component in the indicated order.
4. Add 10 μL of cDNA Synthesis Mix to RNA baits/primer mixture, mix gently and collect by brief centrifugation. Incubate as described in Table 8.
5. Collect the reaction by brief centrifugation. Add 1 μL of RNase H to each tube and incubate the tubes for 20 min at 37 °C. cDNA synthesis reaction can be stored at −20 °C for long-term storage or used for qPCR immediately.

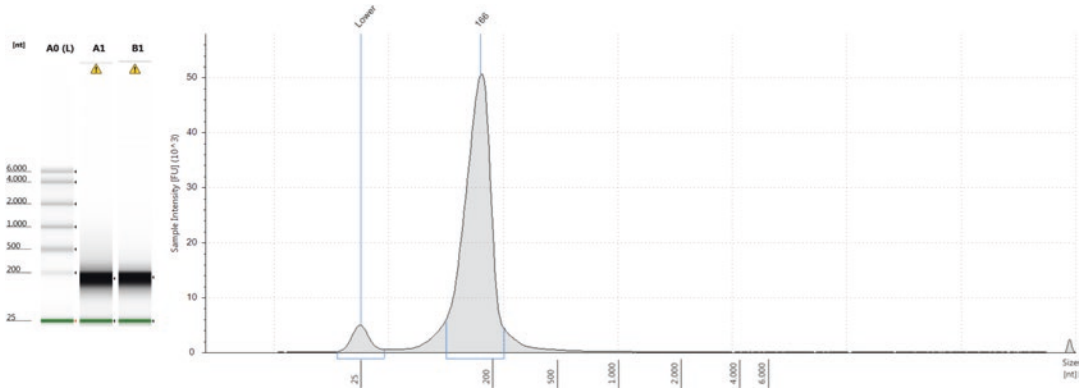


Fig. 4 *Example of RNA baits after biotin-UTP transcription.* The left graph shows the ladder on the left and two lanes of Biotin labeled RNA baits. The right graph shows the electropherogram of lane A1. There is a strong peak around 166 bp and this is what we expect here, a strong peak between 160 and 180 bp

Table 6
RNA synthesis QC, pipetting scheme

Component	Volume in µL
Custom RNA baits (100 ng/µL)	2.0
Random hexamers	1.0
10 mM dNTP mix	1.0
DEPC-treated water	6.0
Total	*10.0*

Table 7
cDNA Synthesis Mix pipetting scheme

Component	Volume in µL
10× RT buffer	2.0
25 mM $MgCl_2$	4.0
0.1 M DTT	2.0
RNaseOUT TM (40 U/µL)	1.0
SuperScript III RT (200 U/µL)	1.0
Total	*10.0*

3.3.2 SYBR Green qPCR

The protocol is optimized using PowerUp TM SYBR Green Master Mix. Please run at least two technical replicates of each qPCR reaction and negative control (NC) without cDNA.

1. Mix the components of Table 9 in the sterile optical 96-well plate for real time PCR.

Table 8
cDNA Synthesis conditions

Step	Temp in °C	Time in min
1	25	10
2	50	50
3	85	5
4	4	5

Table 9
qPCR pipetting scheme

Component	Volume in µL
2× SYBR green PCR master mix	6.25
Bait QC 3 forward primer (10 µM)	2.0
Bait QC 3 forward primer (10 µM)	2.0
cDNA of RNA baits	2.0
Total	*12.25*

Table 10
qPCR cycling conditions

Cycle Step	Cycles	Temp in °C	Time
Initial denaturation	1	95	10 min
Denaturation	40	95	15 s
Annealing/extension		65	1 min

2. Run the reactions using standard cycling parameters (Table 10).
3. If synthesis of RNA baits was successful, CT values should appear between 10 and 15 cycles. NC should not show up at all or at >30 cycles (Fig. 5).

3.4 gDNA Fragmentation (See Note 7)

Random fragmentation is the random shearing of genomic DNA. In this protocol, the shearing is done by ultrasonic sound with a Covaris Focused-ultrasonicator. Alternatives using DNA cutting enzymes (transposases) are also available and can replace the random fragmentation method described here. For the later sequencing and data generation the random nature of the fragmentation is crucial. Many software tools rely on the random

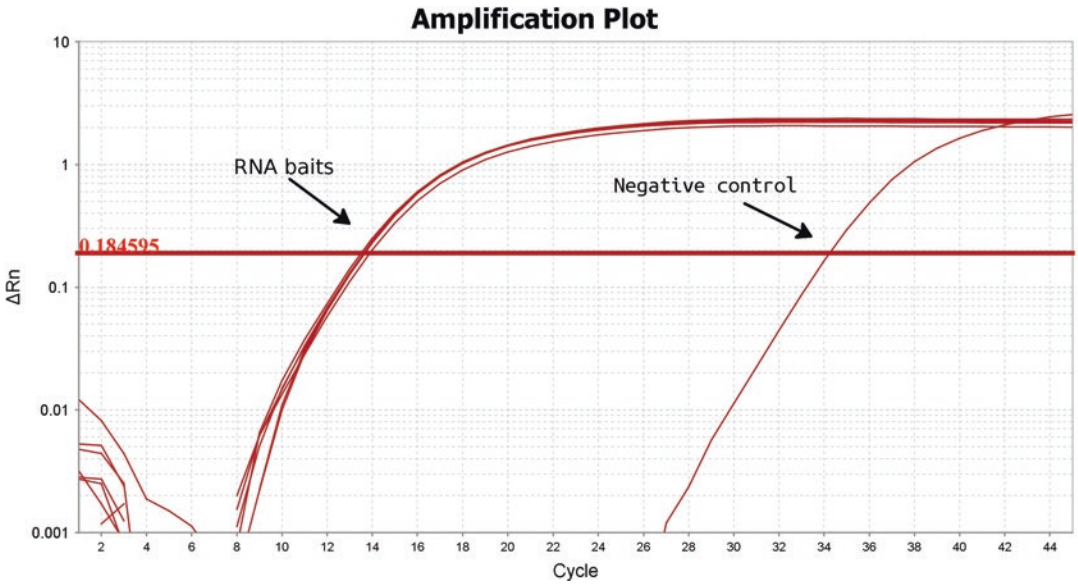


Fig. 5 *Amplification plot of RNA baits.* If the synthesis of the RNA baits was successful, CT values should be between 10 and 15 cycles. The negative control should not show any signal or if, at >30 cycles

fragmentation nature and NGS sequencers work best for random fragmented DNA (for NGS of Amplicons it is usually necessary to spike in at least 15% phiX DNA).

3.5 Library Preparation

After the random fragmentation, the sequencing libraries are prepared. This consists of adapter ligation and amplification of the adapter ligated fragments. The adapters are essential for the sequencer chemistry. They usually consist of a sequence that binds to its counterpart at the surface of the sequencer slide, followed by sequencing primer-binding sites, spacers and molecular barcodes for DNA indexing.

3.5.1 gDNA Fragmentation

Provide 50 µL of gDNA (100 ng-5 µg) and shear with Covaris using the settings of Table 11 to achieve a 500 bp fragment peak.

This protocol performs the library prep with the NEBNext Multiplex Oligos for Illumina (Dual Index Primers) kit (*see* **Note 8**).

1. To perform the NEBNext End Prep mix the components of Table 12 in a sterile nuclease-free tube.
2. Mix by pipetting followed by a quick spin to collect all liquid from the sides of the tube.
3. Place in a thermocycler, with the heated lid on, and run the program of Table 13.

Table 11
gDNA fragmentation pipetting scheme

Intensity	5
Duty cycle	5%
Cycles per burst	200
Treatment time (s)	35
Temperature in °C	7
Water level - S2	12
Water level - E210	6
Sample volume in µL	50
E210 - intensifier	Yes

Table 12
End prep pipetting scheme

Component	Volume in µL
(green) end prep enzyme mix	3.0
(green) end repair reaction buffer (10×)	6.5
1 µg fragmented DNA (Subheading 3.4.1)	55.5
Total	*65.0*

Table 13
End prepare thermocycler program

Step	Temp in °C	Time in min
1	20	30
2	65	30
3	4	∞

4. Adaptor ligation starts now—add the components of Table 14 directly to the End Prep reaction mixture and mix well (if DNA input is <100 ng, *see* **Note 9**).
5. Mix by pipetting followed by a quick spin to collect all liquid from the sides of the tube.
6. Incubate at 20 °C for 15 min in a thermal cycler.
7. Add 3 µL of USER™ Enzyme to the ligation mixture from last step. The USER™ Enzyme can be found in the NEBNext Multiplex kit.

Table 14
Adaptor ligation pipetting scheme

Component	Volume in µL
Blunt/TA ligase master mix	15.0
NEBNext adaptor for Illumina[a]	2.5
Ligation enhancer	1.0
Total	*83.5*

[a]*The NEBNext adaptor in NEBNext Multiplex Oligos for Illumina (Dual Index Primer)*

8. Mix well and incubate at 37 °C for 15 min.
9. Clean-up of adaptor-ligated DNA starts by vortexing AMPure XP Beads to resuspend.
10. Add 86.5 µL resuspended AMPure XP Beads to the ligation reaction. Mix well by pipetting up and down at least 10 times.
11. Incubate for 5 min at room temperature.
12. Quickly spin the tube and place it on an appropriate magnetic stand to separate beads from the supernatant. After the solution is clear (about 5 min), carefully remove and discard the supernatant. Be careful not to disturb the beads that contain DNA targets (**Caution**: do not discard beads).
13. Add 200 µL of 80% freshly prepared ethanol to the tube while in the magnetic stand. Incubate at room temperature for 30 s, and then carefully remove and discard the supernatant.
14. Repeat **step 13** once.
15. Air dry the beads for 5 min while the tube is on the magnetic stand with the lid open (*see* **Note 6**).
16. Remove the tube/plate from the magnet. Elute the DNA target from the beads by adding 17 µL of 10 mM Tris-HCl or 0.1× TE.
17. Mix well by pipetting up and down, or on a vortex mixer. Incubate for 2 min at room temperature.
18. Quickly spin the tube and place it on the magnetic stand.
19. After the solution is clear (about 5 min), transfer 15 µL to a new PCR tube for amplification.
20. Mix the components of Table 15 in sterile strip tubes to start PCR Amplification including indexing.
21. Run the PCR program described in Table 16.
22. Continue with PCR clean-up by vortexing AMPure XP Beads to resuspend.
23. Add 45 µL of resuspended AMPure XP Beads to the PCR reactions (~50 µL). Mix well by pipetting up and down at least 10 times.

Table 15
PCR amplification and indexing pipetting scheme

Component	Volume in μL
Adaptor ligated DNA fragments	15.0
NEBNext Q5 hot start HiFi PCR master mix	25.0
One i7 primer per reaction (*See* **Note 10**)	5.0
One i5 primer per reaction (*See* **Note 10**)	5.0
Total	*50.0*

Table 16
Thermocycler program for amplification with indexing

Cycle step	Cycles	Temp in °C	Time
Initial denaturation	1	98	2 min
Denaturation	4–12[a]	98	10 s
Annealing/extension		65	75 s
Final extension	1	65	5 min
Hold	1	4	∞

[a] *We suggest 6 PCR cycles for 1 μg DNA input. Slightly increase cycles if input DNA is below 50 ng and keep 12 cycles as an upper limit. Further optimization of PCR cycle number may be required*

24. Incubate for 5 min at room temperature.
25. Quickly spin the tube and place it on an appropriate magnetic stand to separate beads from the supernatant. After the solution is clear (about 5 min), carefully remove and discard the supernatant. Be careful not to disturb the beads that contain DNA targets. (**Caution**: do not discard beads).
26. Add 200 μL of 80% ethanol to the PCR plate while in the magnetic stand. Incubate at room temperature for 30 s, and then carefully remove and discard the supernatant.
27. Repeat **step 5** once.
28. Air dry the beads for 5 min while the PCR plate is on the magnetic stand with the lid open (*see* **Note 6**).
29. Remove the tube/plate from the magnet. Elute DNA target from beads into 33 μL 0.1× TE. Mix well by pipetting up and down at least 10 times. Quickly spin the tube and incubate at room temperature for 2 min.

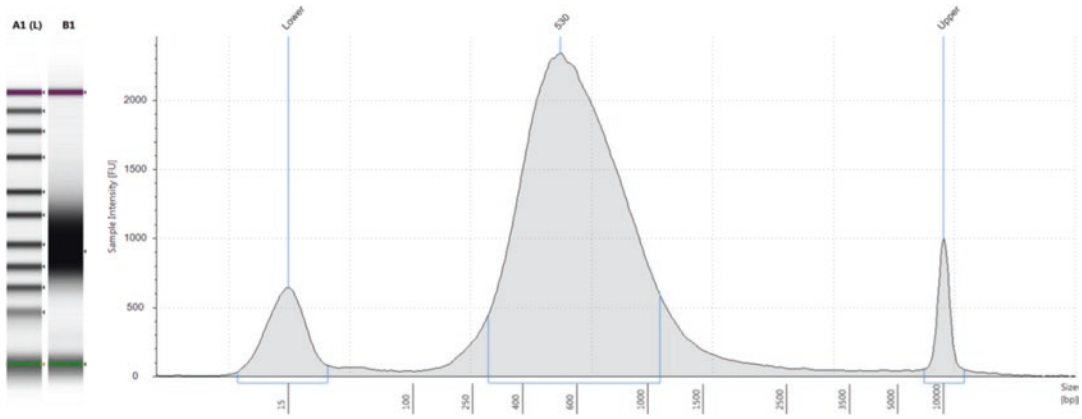


Fig. 6 *Example of a ready-to-sequence library.* The left graph shows the ladder on the left followed by a lane of DNA sequencer library. The right graph shows the electropherogram of lane B1. The peaks at 25 and 1000 bp are from the upper and lower markers and help to determine the fragment size of the DNA in between. The peak at 530 bp that ranges from 250 bp up to 1500 bp represents the ready-to-load library. This is an optimal result, as we expect only a minor fraction of overlapping paired end sequences here. A slightly higher fraction is 1000 bp and larger, which in some cases allows for phasing across intronic sequences. The majority of fragments have the correct size to expect a high amount of paired end sequences that map to adjacent sequence parts

30. Place the sample on an appropriate magnetic stand to separate beads from the supernatant. After the solution is clear (about 5 min), carefully transfer 28 µL supernatant to a new PCR tube. Libraries can be stored at −20 °C.
31. Check the size distribution with an aliquot of the library on a D1000 Tape (Agilent Tapestation®). You find an example in Fig. 6.

3.6 Targeted Enrichment (Bait-Based)

The main idea behind the targeted enrichment is the hybridization between the biotinylated RNA and the library produced in Subheading 3.5. By capturing the DNA-RNA hybrids it is possible to enrich the target of interest. To achieve that, Streptavidin coated magnetic beads are added to the hybridization. The Streptavidin binds the Biotin of the RNA baits and it is thus possible to immobilize the hybridized DNA by a magnetic separator. Washing steps remove the unbound/unspecific DNA fragments. The next step performs a degradation of the RNA baits followed by a PCR cleanup to catch the afore-bounded DNA fragments. Finally, a PCR with a low number of cycles is performed to amplify the captured DNA.

3.6.1 Hybridization (See Note 11)

1. For singleplex hybridization (500 ng of DNA input), dilute HLA RNA baits stock solution (100 ng/µL, *see* Subheading 3.2, **step 8**) 1:5 with nuclease-free water.

Table 17
Hybridization conditions, thermocycler program

Cycle step	Temp in °C	Time
Initial denaturation	95	5 min
Pre-warm hybridization mix	65	5 min
Pre-warm capture mix	65	5 min
Hybridization	65	∞

2. For multiplex hybridization:
   (a) 8 samples (62.5 ng of DNA input of each), dilute HLA RNA baits stock solution 1:5 with nuclease-free water.
   (b) 6 samples (31.25 ng of DNA input of each), dilute HLA RNA baits stock solution 1:5 with nuclease-free water.
3. Set up the program of Table 17 on a thermocycler. This program will help during the pre-warming of the components and eventually perform the 65 °C incubation for hybridization.
4. Prepare the library mix of Table 18 in a nuclease-free tube and vortex.
5. Prepare the hybridization mix of Table 19 in a nuclease-free tube and vortex.
6. Prepare the capture mix of Table 20 in a nuclease-free tube and vortex.
7. Place the library mix in a thermocycler and start the program of Subheading 3.5.1 **step 3** (Table 17). Once cycle **step 2** is reached, place the tube with the hybridization mix in the cycler to perform pre-warming. When **step 3** of the cycler got reached, add the capture mix tube to the thermocycler to perform pre-warming of this component as well. Cycler **step 4** represents the hybridization. When this step is reached add 7 μL of the pre-warmed library mix and 13 μL of the pre-warmed hybridization mix to the pre-warmed capture mix. Mix by gentle pipetting.
8. Incubate the hybridization reaction at 65 ° C for 36 h.

3.6.2 Recovery of Bait-DNA-Hybrid

The aim of this step is to get rid of non-hybridized DNA and to recover the bait-DNA-hybrids bound to the magnetic beats. It consists of various washing step and eventually we want to keep the beat-bait-DNA complex.

1. Transfer 50 μL of MyOne Streptavidin C1 magnetic beads to a new 1.5 mL tube.

Table 18
Library mix pipetting scheme

Component	Volume in µL
Block#1	2.5
Block#2	2.5
Block#3	0.6
Sequencing library (from 3.4.2.5.9)	3.4
Total	*9*

Table 19
Hybridization mix pipetting scheme

Component	Volume in µL
HYB#1	20
HYB#2	0.8
HYB#3	8
HYB#4	8
Total	*36.8*

Table 20
Capture mix pipetting scheme

Component	Volume in µL
Diluted RNA baits	5
RNAse block	1
Total	*6*

2. Pellet beads using a magnetic particle stand and discard the supernatant.
3. Add 200 µL Binding Buffer to beads to wash. Vortex the tube for 5–10 s, place on magnetic particle stand for 2 min to pellet the beads and remove and discard the supernatant.
4. Repeat Subheading 3.6.2, **step 3** twice for a total of three washes.
5. Resuspend the beads in 200 µL Binding Buffer.
6. Transfer the hybridization solution to the Binding Buffer/Beads and incubate for 30 min at room temperature on a rota-

tor. Pellet beads with magnetic particle stand for 2 min and remove the supernatant.

7. Add 500 µL Wash Buffer 1 to the beads and briefly vortex to resuspend. Incubate for 15 min at room temperature. Pellet beads with magnetic particle stand for 2 min and remove the supernatant.
8. Add 500 µL 65 °C Wash Buffer 2 to the beads and briefly vortex to mix. Incubate for 10 min at 65 °C. Pellet beads with magnetic particle stand for 2 min and remove the supernatant.
9. Repeat **step 8** twice for a total of three 65 °C washes. After third wash make sure all additional buffer is removed.

3.6.3 Elution of Target DNA

In this step, we recover the target DNA from the RNA-DNA-hybridization. We achieve that by degrading the RNA molecules by an alkaline treatment followed by a centrifugation where we pellet the magnetic beats. Eventually we keep the supernatant which contains the target DNA.

1. Add 50 µL freshly prepared Elution Buffer to beads from Subheading 3.6.2.
2. Vortex for 5–10 s to mix.
3. Incubate for 10 min at room temperature.
4. Pellet the beads and transfer the supernatant to a tube containing 70 µL Neutralization Buffer. **Caution**: we keep the supernatant here.

3.6.4 AMPure XP Bead Purification

1. Add 120 µL AMPure XP Beads and incubate 5 min at room temperature.
2. Incubate for 5 min at room temperature.
3. Quickly spin the tube and place it on an appropriate magnetic stand to separate beads from the supernatant. After the solution is clear (about 5 min), carefully remove and discard the supernatant. Be careful not to disturb the beads that contain DNA targets (**Caution**: do not discard beads).
4. Add 200 µL of 80% ethanol while tube is in the magnetic stand. Incubate at room temperature for 30 s, and then carefully remove and discard the supernatant. Repeat that step one.
5. Dry 5 min at room temperature (*see* **Note 6**).
6. Remove the tube from the magnet. Add 30 µL 0.1× TE. Mix well by pipetting up and down at least 10 times. Quickly spin the tube and incubate at room temperature for 2 min.
7. Place the sample on an appropriate magnetic stand to separate beads from the supernatant. After the solution is clear (about 5 min), carefully transfer 28 µL of the supernatant to a new PCR tube.

Table 21
Post capture amplification pipetting scheme

Component	Volume in µL
Nuclease-free water	4.5
5× Herculase II buffer	10.0
dNTP mix (25 mM each)	0.5
PCR primers mix (10 mM each)	1.0
Herculase II fusion DNA polymerase	1.0
Captured library[a]	28.0
Total	*45.0*

[a]In this step, all DNA library after the purification Subheading 3.6.4 is used. If less volume is used in this step, increase the volume of nuclease-free water

Table 22
Post capture amplification thermocycler program

Cycle Step	Cycles	Temp in °C	Time
Initial denaturation	1	95.0	30 s
Denaturation	14[a]	95.0	20 s
Annealing		52.4[b]	30 s
Extension		72.0	[c]
Extension	1	72.0	5 min
Hold	1	4.0	∞

[a]We run 14 cycles, but reduce the number of cycles if possible (if enough product for sequencing is already available earlier). This reduces PCR introduced bias
[b]The annealing temperature should be 5 °C below the lowest primer TM. The primer sequences have to fit to the sequencing adapters. In this protocol, we refer to the NEBnext library prep. The corresponding Amplification primers are TruSeq-dual and TruSeq-ind with a minimal TM of 57.4. Set the annealing temperature to 52.4 °C
[c]Extension time (**step 4**) will depend on the genomic library average fragment size. Use 30 s for fragments shorter than 500 bp, 45 s for fragments with size between 500 and 700 bp and 1 min for fragment sizes ranging from 700 bp to 1 Kb

3.6.5 Amplification of Enriched Target

This step amplifies the DNA that was captured and cleaned in the previous steps. This is necessary to get enough material for the sequencing. Keep the number of PCR cycles in this step as low as possible to avoid PCR created bias. In this protocol, we run 14 cycles which should be the upper limit.

1. Prepare PCR Mix as described in Table 21 on ice in a nuclease-free tube and mix by pipetting.

2. Place the tubes in a thermocycler and run the program of Table 22.
3. Purify the DNA with AMPure XP Beads (perform all steps of Subheading 3.6.4).
4. Validate and quantify an aliquot of the enriched library on a D1000 Tape (Agilent Tapestation®, *see* Fig. 6).

3.7 NGS Sequencing of Enriched Library

Paired end sequencing is the method of choice here. The library preparation from Subheading 3.5 generated fragments that can be sequenced from both sites. Eventually, the sequencers deliver reads from both sites of these fragments. After Subheading 3.6 a ready-to-sequence library that is enriched for the HLA target is available. Sequencing can be performed with any kind of NGS sequencer that is able to sequence the library produced in Subheading 3.5. A guiding value for the number of expected sequencing reads would be around 3 million paired 125 bp reads or longer. One-hundred base pair read length should be the lower limit and typical values are 2 × 100, 2 × 125, 2 × 150, or 2 × 175 bp.

1. Load the HiSeq2500 with the enriched library (product of Subheading 3.6) and perform a paired-end sequencing of 2*125 bp using HiSeq® SBS Kit v4 (250 cycles).
2. Provide the demultiplexed fastq files from the sequencing run for the data analysis in step 3.8.

3.8 HLA Calling

The HLA calling can be performed by using different software tools. Free software packages are available such as OptiType (https://github.com/FRED-2/OptiType) [9], xHLA (https://github.com/humanlongevity/HLA) [10], or our HLAssign software (http://www.hlassign.org; tutorial also available under this link) [8]. Commercial software solutions are available through the companies Omixon https://www.omixon.com) or GenDX (http://www.gendx.com/products/ngsengine). All these software tools are aligning the paired end sequencing reads to the reference of known HLA alleles and evaluate the resulting alignment.

4 Notes

1. This protocol assumes that scientists generate biotin-labeled RNA from a DNA oligo pool. The oligo pool is a collection of sequences designed to capture the genomic target of interest. This approach allows for bait synthesis for ten-thousands of samples. If only interested in minor sample sizes it may be useful to order biotin labeled baits from any providing company. This shortens the protocol tremendously and Subheadings

2.1–2.3 and 3.1–3.3 can be skipped. The required RNA design can be downloaded in fasta format from GitHub: https://github.com/MiWitt/HLAbaits/blob/master/design/probes_IKMB_custom_v0.1.fasta.

2. Fragmentation and library preparation was established and heavily used as described in this protocol. But, recently many labs shift to transposase based protocols like the Illumina Nextera® Sample Preparation Kit (together with Nextera Index Kit) and we also have good experience with this alternative that reduces costs and working time. We recommend replacing Subheading 3.4 using this alternative.
3. This is an optional step and it can be skipped and directly proceeded to Subheading 3.1.2. However, this step helps to select full-length synthetic oligonucleotides and simplifies the later frequent re-synthesis of RNA baits by the usage of aliquots derived from this step. Additionally, this step can reduce the costs of baits per sample if processing a large number of samples. **Caution:** This step can reduce the complexity of oligo pool by PCR bias.
4. Alternative methods such as ultra-filtration or silica column-based purification could be used to achieve the same outcome.
5. In this step, we additionally performed size selection because of constant nonspecific amplification around 500 bp (*see* Fig. 3). However, in case there is no nonspecific amplification, perform a standard clean-up as described in methods Subheading 3.1.1 (start at 3) or use an alternative clean-up method. To validate this, it is necessary to verify the PCR product by a Tape Station run before. If there is only a single peak at around 190 bp (like Fig. 2 but at 190 bp) and not much unspecific product like in Fig. 3, a simple PCR cleanup without size selection can be performed.
6. Do not over dry the beads. This may result in lower recovery rate. Over drying can be identified by multiple cracks or a single huge crack in the beat pellet.
7. In this protocol we recommend a fragment size of around 500 bp. In the Illumina NGS protocol, this will be referred to as insert size, and it is the DNA fragment which will be sequenced from both sides during the paired end sequencing. For the data analysis, a widespread fragment range is very helpful. Short fragments will nicely cover the important exonic sequences of the HLA genes and support intra-exon phasing. Longer fragments sometimes help to allow phasing between 5′ and 3′ parts of exons, if variations are only at these flanks and

not in between. In some cases, long fragments even allow for trans-exon phasing. On the other hand, the packing rate of modern NGS sequencer becomes very dense and an upper limit is given for the insert size. As we refer in this protocol to the HiSeq2500 using v4 chemistry, the upper limit is between 800 bp and 1000 bp. But other Sequencer versions can have other restrictions, have a look at the specifications and pay attention to that restriction.

8. There are alternative kits available for Singleplex and Multiplex Oligos. Some require a different PCR enrichment of adaptor ligated DNA. With respect to sample size and reagent costs it may be a good idea to choose an alternative indexing/adaptor ligation. In this protocol, we assume that eventually up to 96 samples can be pooled in a single sequencer lane. For that case, we chose the NEBNext Multiplex Oligos for Illumina (Dual Index Primers) kit.
9. If DNA input is <100 ng, dilute the NEBNext Adaptor for Illumina (provided at 15 μM) tenfold in 10 mM Tris-HCl or 10 mM Tris–HCl with 10 mM NaCl to a final concentration of 1.5 μM, use immediately. We recommend using the highest amount of DNA (1 μg fragmented/end-prepped DNA) because this reduces the number of PCR cycles in some steps of the downstream protocol. And it reduces the number of possibly introduced PCR artifacts.
10. i7/i5 primers are provided in the NEBNext Multiplex Oligos for Illumina (Dual Index Primers) kit. The combination of each i7/i5 primer sequence generates a unique molecular barcode that tags a specific sample during the sequencing. So, it is mandatory to have a unique i7/i5 primer combination for each sample that is pooled together for sequencing and/or targeted enrichment later. 12 i7 and 8 i5 primers allow for pooling up to 96 samples.
11. The targeted enrichment can be performed for each sample separate, or multiple samples can be pooled and be enriched in a single step. The reason is not only to reduce costs, this can increase the on-target rate from sequencing as well. The HLA target, as designed in Subheading 2.1, is around 220 kb. This is rather a small target and the competition between specific and unspecific hybridization is very low, as an excess of baits is provided. Pooling samples and/or increasing the target/bait ratio can increase specificity here. We also tried different salt and temperature conditions which had only a minor effect. Another approach for better on-target rates could be a second round of target capturing.

References

1. Hosomichi K, Shiina T, Tajima A, Inoue I (2015) The impact of next-generation sequencing technologies on HLA research. J Hum Genet 60(11):665–673. https://doi.org/10.1038/jhg.2015.102
2. Wetmur JG, Fresco J (1991) DNA probes: applications of the principles of nucleic acid hybridization. Crit Rev Biochem Mol Biol 26:227–259
3. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X et al (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol 11:R119
4. Walsh PS, Erlich HA, Higuchi R (1992) Preferential PCR amplification of alleles: mechanisms and solutions. Genome Res 1:241–250
5. Voorter CE, Kik MC, van den Berg-Loonen EM (1998) High-resolution HLA typing for the DQB1 gene by sequence-based typing. Tissue Antigens 51:80–87
6. Lefranc MP, Giudicelli V, Kaas Q, Duprat E, Jabado-Michaloud J, Scaviner D, Ginestoux C, Clément O, Chaume D, Lefranc G (2005) IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res 33:D593–D597
7. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C et al (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27:182–189
8. Wittig M, Anmarkrud JA, Kässens JC, Koch S, Forster M, Ellinghaus E, Hov JR, Sauer S, Schimmler M, Ziemann M, Görg S, Jacob F, Karlsen TH, Franke A (2015) Development of a high-resolution NGS-based HLA-typing and analysis pipeline. Nucleic Acids Res 43(11):e70. https://doi.org/10.1093/nar/gkv184
9. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O (2014) OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics 30(23):3310–3316. https://doi.org/10.1093/bioinformatics/btu54
10. Xie C, Yeo ZX, Wong M, Piper J, Long T, Kirkness EF, Biggs WH, Bloom K, Spellman S, Vierra-Green C, Brady C, Scheuermann RH, Telenti A, Howard S, Brewerton S, Turpaz Y, Venter JC (2017) Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. Proc Natl Acad Sci U S A 114(30):8059–8064. https://doi.org/10.1073/pnas.1707945114

Chapter 6

High-Throughput Contiguous Full-Length Next-Generation Sequencing of HLA Class I and II Genes from 96 Donors in a Single MiSeq Run

Philip K. Ehrenberg, Aviva Geretz, and Rasmi Thomas

Abstract

The human leukocyte antigen (HLA) genes regulate and drive the immune system, and are among the most polymorphic loci in the human genome. HLA diversity is known to play an important role in transplantation and disease association studies. There are multiple approaches to DNA-based HLA genotyping and recent advances in next-generation sequencing (NGS) technologies have facilitated the development of whole gene sequencing methods. We describe an accurate, efficient, scalable, and cost-effective approach to contiguously amplify and sequence full-length genes of six HLA class I and II loci from 96 individuals on a single Illumina MiSeq run.

Key words HLA, HLA genotyping, NGS, Next-generation sequencing, Full-length amplification, MIT-NGS, Illumina, MiSeq

1 Introduction

Next-generation sequencing (NGS) approaches for HLA typing offer several advantages over conventional Sanger sequence-based typing (SBT). Most notably, while SBT HLA typing calls are generated from sequences of one or two core exons that comprise the peptide-binding region (PBR), NGS approaches allow cost-effective sequencing of the entire HLA gene. Though many polymorphisms reside within the PBR sequences of HLA genes, excluding potential variation within other exons, as well as introns and untranslated regions (UTR), can result in ambiguous calls and preclude identification of novel alleles [1]. Variation within these regions can also be important for HLA protein structure, splice site variation, transcription rate, and mRNA stability [2–4]. NGS HLA typing utilizes long-range PCR strategies resulting in high-resolution unambiguous HLA typing that can facilitate identification of novel as well as rare alleles. These benefits of NGS-based

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_6,

HLA typing can be particularly relevant in clinical diagnostics, transplantation studies, population genetics, and disease associations [5–8].

We have previously developed an accurate and cost efficient Multi-locus Individual Tagging-Next-Generation Sequencing (MIT-NGS) method as a high-throughput multi-locus indexed sequencing approach to simultaneously sequence contiguous full-length HLA genes, from 5′UTR through 3′UTR regions, of four HLA loci (A, B, C, DRB1) from 96 individuals in a single MiSeq run [9]. The flexibility and versatility of this method has been further demonstrated with subsequent additions of DPA1, DPB1, DQA1, and DQB1 loci, enabling up to eight-locus sequencing runs [1]. Availability of additional indices and higher throughput platforms, such as Illumina NextSeq and HiSeq, can accommodate additional loci and samples as required using the same approach.

This chapter will describe in detail the wet-lab experimental design and execution of a six-locus MIT-NGS method for typing HLA-A, B, C, DPB1, DQB1, and DRB1 loci from 96 donors in a single MiSeq run [1].

2 Materials

All reagents should be of Molecular Biology Grade (MBG). Similarly all disposables should be sterile and DNase-/RNase-free. The item or vendor that we used is shown in parentheses, but in some cases the reagents, kits, and instruments used here may not specifically be required to execute this protocol. Total reagent volumes or numbers of specific items required for the whole protocol are noted when first applicable.

2.1 PCR Amplification

1. Good quality DNA, typically extracted and purified using standard column-based procedures (e.g., QIAamp DNA Blood Mini kit; Qiagen, Valencia, CA).
2. PCR primer mixes for long-range contiguous full-length amplifications of HLA-A, B, C, DPB1, DQB1, and DRB1 loci (*see* Table 1). Standard desalt primer purification is sufficient.
3. Platinum Taq HiFi polymerase kit (ThermoFisher Scientific, Waltham, MA).
4. PrimeSTAR GXL polymerase kit (Takara Bio, Japan).
5. dNTP mix (1.25 mM each) (~1 ml).
6. DMSO (~200 μl).
7. Water, MBG.
8. Microcentrifuge tubes, 1.6 ml (~13 total).

Table 1
Primers for long-range PCR amplification of six HLA loci

Locus	Primer mix	Sequence	Volume/ reaction (μl)[a]	Final concentration (nM)	Reference
A	A F1	AAC TCA GAG CTA AGG AAT GAT GGC AAA T	0.1	100	[10]
	A F2	AAC TCA GAG CTA TGG AAT GAT GGT AAA T	0.1	100	[10]
	A R1	ATA TAA CCA TCA TCG TGT CCC AAG GTT C	0.2	200	[10]
B	B F1	CCC GGT TGC AAT AGA CAG TAA CAA A	0.2	200	[10]
	B R1	GGG TCC AAT TTC ACA GAC AAA TGT	0.2	200	[10]
C	C F1	TGC TTA GAT GTG CAT AGT TCA CGA A	0.1	100	[10]
	C F2	TGC TTA GAT GTG CAT AGT TCC GGA A	0.1	100	[10]
	C R1	TGG ACC CAA TTT TAC AAA CAA ATA	0.2	200	[10]
DPB1	DPB1 F2	GCC TAG TGA GCA ATG ACT CAT A	0.2	200	[5]
	DPB1 R1	CCC AGT TTG GAT GGT CTC TCA GCT CTT	0.2	200	[11]
DQB1	DQB1 F6	TAT GAC AGC AAT TTT CTC TCC CCT G	0.2	200	[5]
	DQB1 R	TCA TGT GCT TCT CTT GAG CAG TCT GA	0.2	200	[11]
DRB1	DRB1 F1.1	GCA TCC ACA GAA TCA CAT TTT CTA GTG TT	0.05	50	[11]
	DRB1 F1.2.1	TCC ACA GAA TCA CAG CAT TTT CTA GTG TT	0.05	50	[9]
	DRB1 F1.3	GCA TCC ACA GAA TCA CAT TTT CCA GTA TT	0.05	50	[9]
	DRB1 F1.4.1	TCC ACA GAA TCA CAG CAT TTT CCA GTA TT	0.05	50	[9]
	DRB1 R2.1	TGA TTG ACT TGC TGG CTG GTT TCT CAT C	0.2	200	[11]

[a]Input of 10 μM primer stocks

9. Matrix tubes, 2D, 0.5 ml (ThermoFisher Scientific) (~100 total).
10. 96-Well PCR plates (e.g., Denville, Metuchen, NJ) (17 total).
11. Plate sealing film (e.g., 120 μm CyclerSeal film; Axygen, Tewksbury, MA) (~55 total).
12. Plate sealer tool.
13. PCR plate coolers, 96-well.
14. Microplate centrifuge.
15. Single- and multi-channel manual and electronic pipets.
16. Programmable PCR thermal cyclers with heated lids (e.g., Veriti 96-well thermal cyclers; ThermoFisher Scientific).

2.2 Agarose Gel Electrophoresis

1. Agarose, electrophoresis-grade (e.g., LE agarose; GeneMate).
2. TBE Buffer (10×), to make 1× working stocks (~9.5 L).
3. Nucleic acid stain (e.g., GelRed; Phenix Research, Chandler, NC) (~45 μl).
4. DNA ladder (to accommodate 4–20 kb amplicons, e.g., λ/HindIII/φx174/HaeIII mix) (~48 μl).
5. Gel loading buffer (2×) [e.g., Orange G (Sigma Aldrich, St Louis, MO): 0.5% Orange G, 30% glycerol, 0.5X TBE] (~1.8 ml).
6. Gel electrophoresis system (e.g., Owl D3-14; ThermoFisher).
7. High-voltage power supply.
8. UV gel documentation imager.

2.3 PCR Product Purification

1. Agencourt AMPure XP beads (Beckman Coulter, Irving, TX) (~9.5 ml).
2. Ethanol, anhydrous (200 proof), to make 70% working stocks (~300 ml). Make fresh and use within 8 h.
3. Liquid handling system (e.g., NXp instrument; Beckman Coulter).

2.4 Amplicon Quantitation

1. DNA quantification system, fluorescence-based (e.g., Quant-iT dsDNA High Sensitivity Assay kit; Life Technologies, Carlsbad, CA).
2. Plate reader (e.g., FilterMax F3; Molecular Devices, Sunnyvale, CA).
3. Plate reader appropriate 96-well plates (e.g., OptiPlate-96F, black; Denville) (6 total).

2.5 Multiplex Library Preparation and MiSeq Run

1. Nextera XT DNA Sample Preparation Kit (Illumina, San Diego, CA).
2. Nextera XT Index Kit, 96 indices (Illumina).

3. PhiX Library (10 nM) (Illumina): to make 10 pM working stocks, add 2 µl of the PhiX library to 3 µl 10 mM Tris–HCl, pH 8.5/0.1% Tween 20, vortex, and spin briefly. Add 5 µl of 0.2 N NaOH, vortex, spin briefly, and incubate for 5 m at room temperature (RT). Immediately add 990 µl of pre-chilled HT-1 (Illumina) for 20 pM, vortex, and spin briefly. Finally dilute this 1:1 with pre-chilled HT-1. Store both concentrations at −20 °C for up to 3 weeks.
4. Ethanol, anhydrous (200 proof), to make 80% working stocks (~50 ml).
5. Tris–HCl, pH 7.5 (1 M) (~5 µl).
6. Tris–HCl, pH 8.5 (1 M), to make 10 mM solutions (~10 ml).
7. Tween 20, to make 0.1% solutions (~10 ml).
8. NaOH (10 N), to make both 0.2 N (~20 µl) and 0.1 N (~4 ml) stocks. Make fresh and use within 8 h.
9. Magnetic stand, for 96-well plates (e.g., 96S Super Magnet; Alpaqua, Beverly, MA).
10. Plate shaker, high speed (e.g., DMS-2500; VWR, Bridgeport, NJ).
11. Heat block.
12. MiSeq Reagent Kit v2 (500 cycles) (Illumina) (*see* **Note 1**).
13. MiSeq instrument (Illumina) (*see* **Note 2**).

2.6 Software for Analyses

1. Omixon Target v1.9.3, or similar versions including HLA Explore (Omixon Biocomputing Kft, Budapest, HU).
2. NGSengine v2.4.0 (GenDX, Utrecht, The Netherlands).
3. Integrative Genomics Viewer (IGV) [12].

3 Methods

This setup is based on 96 samples arranged in corresponding wells of six different HLA locus-specific 96-well PCR plates. Prepare all reagents and samples, and store all plates and tubes, on ice or in plate coolers unless otherwise noted. Prior to all plate vortex and centrifugation steps, which can be carried out at RT, be sure to seal plates with sealing film.

3.1 PCR Preparation

1. Dilute DNA samples to 50 ng/µl working stock concentrations (*see* **Note 3**).
2. Arrange and record the 96 sample layout (*see* **Note 4**).
3. Label six 96-well PCR plates as "A", "B", "C", "DPB1", "DQB1", and "DRB1".
4. Label six 1.6 ml microcentrifuge tubes as above.

3.2 Long-Range PCR of HLA-A, B, C and DPB1

1. In four separate 1.6 ml microcentrifuge tubes prepare Platinum Taq HiFi PCR master mixes for each of these loci based on reagent volume inputs for one sample (10 μl total volume) as follows: 5.5 μl water, 1 μl 10× Platinum Taq HiFi Buffer (final PCR concentration, 1×), 0.4 μl 50 mM $MgSO_4$ (final PCR concentration, 2 mM), 1.6 μl 1.25 mM dNTPs (final PCR concentration, 200 μM), 0.5 μl DMSO (final PCR concentration, 5%), 0.4 μl appropriate primer mix (final PCR concentration of the forward and reverse primer components, 200 nM each) (*see* Table 1), 0.1 μl Platinum Taq HiFi polymerase (final PCR input, 0.5 U). Master mixes for each locus should include additional volumes to account for pipetting errors. Vortex to mix, and spin briefly to pool contents.
2. Using an electronic pipet dispense 9.5 μl of each master mix to the wells of its corresponding plate.
3. Using a multi-channel pipet dispense 0.5 μl (25 ng) of the DNA sample working stocks into the corresponding wells of each plate (*see* **Note 5**). Pipet up and down to rinse residual template from the tips.
4. Seal the four plates with sealing film (*see* **Note 6**), gently vortex, and centrifuge 325 × *g*, 30 s.
5. Perform the amplifications in programmable thermal cyclers with heated lids per Table 2. The amplified product plates can be used immediately, or stored at −20 °C until ready for the next step.

3.3 Long-Range PCR of HLA-DQB1 and DRB1

1. In separate tubes prepare PrimeSTAR GXL PCR master mixes for these loci based on reagent volume inputs for one sample (10 μl total volume) as follows: 4.8 μl water for DQB1 (or 5.3 μl water for DRB1), 2 μl of 5× GXL Buffer (final PCR concentration, 1×), 1.6 μl 1.25 mM dNTPs (final PCR concentration, 200 μM) (*see* **Note 7**), 0.4 μl appropriate primer mix (final PCR concentration of the forward and reverse primer components, 200 nM each) (*see* Table 1), 0.2 μl PrimeSTAR GXL polymerase (final PCR concentration, 0.25 U). Master mixes for each locus should include additional volumes to account for pipetting errors. Vortex to mix, and spin briefly to pool contents.
2. Using an electronic pipet dispense 9 μl of the DQB1 (or 9.5 μl of the DRB1) master mix into the wells of the corresponding plate.
3. Using a multi-channel pipet dispense 1 μl (50 ng) of the DNA sample working stocks into the corresponding wells of the DQB1 plate [or 0.5 μl (25 ng) to the DRB1 plate] (*see* **Note 5**). Pipet up and down to rinse residual template from the tips.
4. Seal the two plates with sealing film, gently vortex, and centrifuge 325 × *g*, 30 s.

Table 2
Long-range PCR cycling parameters for six HLA loci

Locus	Amplicon length (kb)	PCR cycling parameters[a]	Cycle number
A	5.4–5.5	94 °C, 30 s; 55 °C, 45 s; 68 °C, 5.5 m	35
B	4.6	94 °C, 30 s; 60 °C, 45 s; 68 °C, 5.5 m	35
C	4.8	94 °C, 30 s; 55 °C, 45 s; 68 °C, 5.5 m	35
DPB1	12.7–12.8	94 °C, 30 s; 58 °C, 45 s; 68 °C, 10 m	35
DQB1	7.1–7.5	98 °C, 10 s; 52 °C, 15 s; 68 °C, 10 m	30
DRB1	11–16.5	98 °C, 10 s; 58 °C, 15 s; 68 °C, 10 m	30

[a]All parameters are preceded by a 94 °C, 2 m denaturation step, and for HLA loci A, B, C, and DPB1 followed by a 68 °C, 10 m final extension step

5. Perform the amplifications in programmable thermal cyclers with heated lids per Table 2. The amplified product plates can be used immediately, or stored at −20 °C until ready for the next step.

3.4 Agarose Gel Confirmation

1. Prepare six 1× TBE 150 ml 1% agarose gels, each with two rows of 50 wells: in a microwave oven melt 9 g of agarose in 900 ml of 1× TBE buffer, swirl flask to cool slightly for ~3 min, add 45 μl GelRed, swirl to mix well, pour ~150 ml into each of six gel casts, place combs, and allow to set (~1 h).
2. Transfer the gels to RT 1× TBE buffer in up to six gel tanks (*see* **Note 8**).
3. Add 2 μl of a λ/HindIII/φx174/HaeIII DNA ladder mix to the first and last wells of each of the two rows of each gel.
4. Into 96-well plates (*see* **Note 9**) prepare sample mixes from the amplified products in 5 μl total volumes at RT as follows: 1.5 μl water, 1 μl amplified product, and 2.5 μl 2× Orange G loading buffer. Pipet up and down to mix.
5. Load the 5 μl sample mixes into the 48 available wells from each of the two rows (*see* **Note 10**).
6. Electrophorese gels at RT at ~100 V for ~1 h, depending upon the desired separation.
7. Place gels over a UV light, photograph, and confirm the presence of the appropriate amplicon sizes (Fig. 1 and Table 2).

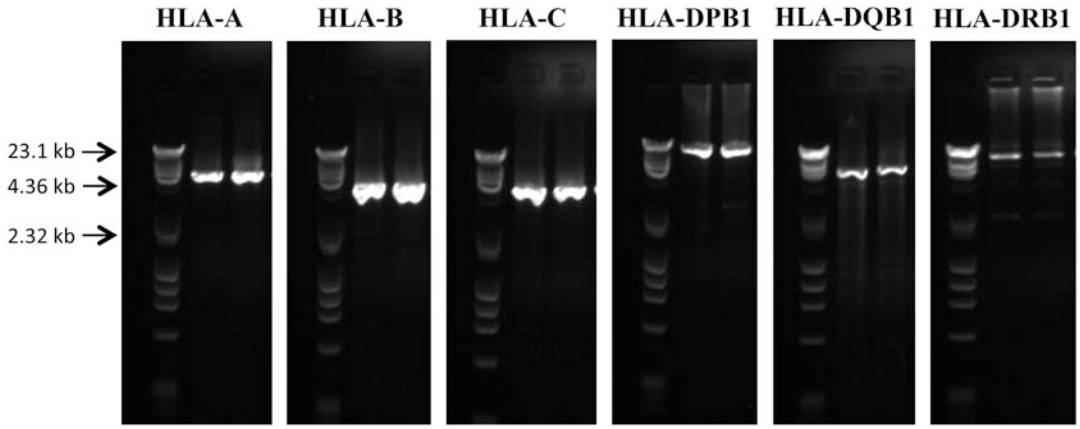


Fig. 1 Contiguous full-length amplification products from the six HLA loci. Sample mixes containing 1 μl of the long-range PCR products from HLA-A (5.4–5.5 kb), B (4.6 kb), C (4.8 kb), DPB1 (12.7–12.8 kb), DQB1 (7.1–7.5 kb), and DRB1 (11–16.5 kb) loci from two representative donor samples were loaded into wells of 1% agarose gels and electrophoresed at ~100 V, for ~1 h. Arrows indicate selected band sizes within the λ/HindIII/φx174/HaeIII DNA ladder

3.5 PCR Product Purification

1. Bring Agencourt AMPure XP beads to RT (~30 m), and centrifuge the six amplification plates at 325 × *g*, 30 s.
2. Place these amplification plates on the Biomek NXp instrument at RT, and set up per manufacturer's suggestions. Briefly, 16.2 μl of well-mixed AMPure XP beads are added to each well, followed by multiple washes with 70% ethanol. Purified product is eluted from the beads in 40 μl of MBG water.
3. Gently vortex the plates, centrifuge at 325 × *g*, 30 s, and immediately continue with the procedure or store at −20 °C indefinitely.

3.6 Amplicon Pooling

1. Label six plate reader appropriate locus-specific 96-well plates.
2. Prepare ~130 ml of Working Solution diluent at RT using the Quant-iT dsDNA High Sensitivity Assay Kit, and transfer 199 μl to each well of these plates using a multi-channel electronic pipet.
3. Transfer 1 μl of each sample from the AMPure-purified sample plates into the Working Solutions of the corresponding wells at RT using a multi-channel pipet, and pipet up and down to rinse the tips.
4. Gently vortex the plates, incubate at RT for 2 m, centrifuge at 325 × *g*, 30 s, remove plate seals, and determine sample stock concentrations at RT using the plate reader per manufacturer's suggestions.
5. Label a single new 96-well PCR plate for pooling into each well the six corresponding donor-specific purified full-length contiguous locus amplicons. Based on the concentrations of

the amplicons, transfer sufficient volumes of each to give equimolar donor-specific pools [4.8 ng of each HLA class I locus, 14.4 ng of DRB1, 12.2 ng of DPB1, and 7 ng of DQB1]. Adjust the final volumes to 30 μl with MBG water.

6. Gently vortex and centrifuge at 325 × *g*, 30 s.

3.7 Library Preparation

1. Label a single new 96-well PCR plate to be used in conjunction with the Nextera XT DNA Sample Preparation Kit (Illumina) executed per manufacturer's protocol, with the following adjustments:
   (a) Tagmentation step: the plates used throughout this protocol are all 96-well PCR plates (*see* **Note 11**).
   (b) Tagmentation step: input 5 μl (comprising 8 ng of total amplified product) from the above six-locus donor pool dilution plate.
   (c) PCR Clean-up step: prior to, and after, each plate shaker step throughout the protocol centrifuge the plate at 325 × *g* for 8 s (*see* **Note 12**).
   (d) Library Pooling step: the Diluted Amplicon Library (DAL) diluent is an HT1 buffer/5 mM Tris, pH 7.5 mix [13].
   (e) Library Pooling step: add 30 μl of the Pooled Amplicon Library (PAL) to 570 μl of the new DAL diluent (HT1 buffer/5 mM Tris, pH 7.5 mix) (*see* **Note 13**).
   (f) Library Pooling step: exchange 30 μl of the DAL mix with 30 μl of the 10 pM PhiX library for a final concentration of 5% PhiX.

3.8 MiSeq Instrument Setup and Run

1. To the "Load Samples" reservoir of the thawed MiSeq Reagent kit (500 cycles) (*see* **Note 2**), add the 600 μl total volume of the DAL, and sequence using the MiSeq Instrument (Illumina) per manufacturer's suggestions.

3.9 Sequence Analyses

Numerous approaches from commercial to custom in-house programs can be employed to align reads and generate HLA allele calls from the MiSeq Reporter software-generated FASTQ files (*see* **Note 14**). We separately analyze these files with two different commercially available HLA genotyping software: Omixon Target and NGSengine. Additionally, we use the Integrative Genomics Viewer (IGV) browser [12] to visualize individual donor HLA locus alignments generated by MiSeq Reporter. This redundant overlapping software approach can enhance the HLA allele typing accuracy as any call differences are flagged for in-depth follow-up evaluations (*see* **Notes 15** and **16**).

4 Notes

1. The MiSeq reagent kit is now available as a 600 cycle option.
2. The output capacity of the MiSeq is sufficient for six-locus HLA pools; however, increased sample multiplexing and/or depth of coverage may be accommodated with higher output instruments, including the NextSeq.
3. To facilitate downstream high-throughput pipetting steps these DNA stocks can be diluted and stored in rubber-capped Matrix tubes. Unlike most other tubes (i.e., microcentrifuge tubes), these are stored in racks having the same dimensions as the wells of a PCR plate, allowing direct template transfer using standard 8- or 12-channel pipets.
4. It is recommended that each PCR amplification plate include at least one positive control and one non-template control (MBG water).
5. Ensure that the sample template positioning is the same for all six HLA locus-specific plates. This will facilitate the pooling of the six HLA locus amplicons into corresponding sample-specific wells.
6. With the plate sealer tool ensure the sealing film tightly adheres to each raised well of the plate, and that the four edges of the film are tightly sealed to the top of the plate.
7. For consistency across the Platinum Taq and PrimeStar PCR master mix setups, we use a common 1.25 mM stock dNTP mix.
8. Fully polymerized gels can be moistened with 1× TBE buffer, wrapped well in plastic wrap, and stored overnight at 4 °C.
9. We use white round-bottom polystyrene plates to prepare the sample mixes. These offer easy visibility and can be rinsed, dried, and re-used.
10. Multi-channel pipets with adjustable spacing facilitate transfer of sample mixes from the plates to the closer-positioned wells in the gels.
11. The Nextera XT protocol employs both 96-well PCR plates and, for bead wash and normalization steps, 96-well (round well) 0.8 ml plates. We find that 96-well PCR plates work well for these bead steps in the context of the plate magnet, and for protocol consistency use them exclusively.
12. This centrifuge force/time combination is sufficient to bring residual liquid to the well bottoms, but will not pellet the beads.

13. This ratio of PAL and diluent can be adjusted, so that cluster densities are within the most efficient range of 850–1000/mm^2.
14. The current paucity of IMGT/HLA reference alleles with complete genomic sequences constrained our HLA typing calls to the exonic sequences. However, the contiguous full-length HLA alleles generated and processed through the MIT-NGS method can facilitate intronic SNP and indel studies in the context of specific alleles. Moreover, the MIT-NGS approach is well positioned to contribute to the expansion of complete HLA allele sequences.
15. The DRB1 primer mix weakly co-amplifies DRB4 and 5. This does not seem to interfere with DRB1 calls. There are various bioinformatics approaches to remove these contaminating sequences if desired [1].
16. Flag homozygous calls especially within the DRB1 and DQB1 loci, as these may signal allele dropout caused by allelic imbalance.

Acknowledgments

This work was supported by a cooperative agreement (W81XWH-07-2-0067) between the Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., and the U.S. Department of Defense (DOD). This research was funded, in part, by the U.S. National Institute of Allergy and Infectious Disease. The views expressed are those of the authors and should not be construed to represent the positions of the U.S. Army or the DOD.

References

1. Ehrenberg PK, Geretz A, Sindhu RK, Vayntrub T, Fernandez Vina MA, Apps R, Michael NL, Thomas R (2017) High-throughput next-generation sequencing to genotype six classical HLA loci from 96 donors in a single MiSeq run. HLA. https://doi.org/10.1111/tan.13133
2. Thomas R, Thio CL, Apps R, Qi Y, Gao X, Marti D, Stein JL, Soderberg KA, Moody MA, Goedert JJ, Kirk GD, Hoots WK, Wolinsky S, Carrington M (2012) A novel variant marking HLA-DP expression levels predicts recovery from hepatitis B virus infection. J Virol 86(12):6979–6985. https://doi.org/10.1128/JVI.00406-12
3. Pisapia L, Cicatiello V, Barba P, Malanga D, Maffei A, Hamilton RS, Del Pozzo G (2013) Co-regulated expression of alpha and beta mRNAs encoding HLA-DR surface heterodimers is mediated by the MHCII RNA operon. Nucleic Acids Res 41(6):3772–3786. https://doi.org/10.1093/nar/gkt059
4. Vince N, Li H, Ramsuran V, Naranbhai V, Duh FM, Fairfax BP, Saleh B, Knight JC, Anderson SK, Carrington M (2016) HLA-C level is regulated by a polymorphic Oct1 binding site in the HLA-C promoter region. Am J Hum Genet 99(6):1353–1358. https://doi.org/10.1016/j.ajhg.2016.09.023
5. Baldwin KM, Ehrenberg PK, Geretz A, Prentice HA, Nitayaphan S, Rerks-Ngarm S,

Kaewkungwal J, Pitisuttithum P, O'Connell RJ, Kim JH, Thomas R (2015) HLA class II diversity in HIV-1 uninfected individuals from the placebo arm of the RV144 Thai vaccine efficacy trial. Tissue Antigens 85(2):117–126. https://doi.org/10.1111/tan.12507

6. Duke JL, Lind C, Mackiewicz K, Ferriola D, Papazoglou A, Gasiewski A, Heron S, Huynh A, McLaughlin L, Rogers M, Slavich L, Walker R, Monos DS (2016) Determining performance characteristics of an NGS-based HLA typing method for clinical applications. HLA 87(3):141–152. https://doi.org/10.1111/tan.12736
7. Prentice HA, Tomaras GD, Geraghty DE, Apps R, Fong Y, Ehrenberg PK, Rolland M, Kijak GH, Krebs SJ, Nelson W, DeCamp A, Shen X, Yates NL, Zolla-Pazner S, Nitayaphan S, Rerks-Ngarm S, Kaewkungwal J, Pitisuttithum P, Ferrari G, McElrath MJ, Montefiori DC, Bailer RT, Koup RA, O'Connell RJ, Robb ML, Michael NL, Gilbert PB, Kim JH, Thomas R (2015) HLA class II genes modulate vaccine-induced antibody responses to affect HIV-1 acquisition. Sci Transl Med 7(296):296ra112. https://doi.org/10.1126/scitranslmed.aab4005
8. Profaizer T, Lazar-Molnar E, Close DW, Delgado JC, Kumanovics A (2016) HLA genotyping in the clinical laboratory: comparison of next-generation sequencing methods. HLA 88(1–2):14–24. https://doi.org/10.1111/tan.12850
9. Ehrenberg PK, Geretz A, Baldwin KM, Apps R, Polonis VR, Robb ML, Kim JH, Michael NL, Thomas R (2014) High-throughput multiplex HLA genotyping by next-generation sequencing using multi-locus individual tagging. BMC Genomics 15:864. https://doi.org/10.1186/1471-2164-15-864
10. Shiina T, Suzuki S, Ozaki Y, Taira H, Kikkawa E, Shigenari A, Oka A, Umemura T, Joshita S, Takahashi O, Hayashi Y, Paumen M, Katsuyama Y, Mitsunaga S, Ota M, Kulski JK, Inoko H (2012) Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. Tissue Antigens 80(4):305–316. https://doi.org/10.1111/j.1399-0039.2012.01941.x
11. Hosomichi K, Jinam TA, Mitsunaga S, Nakaoka H, Inoue I (2013) Phase-defined complete sequencing of the HLA genes by next-generation sequencing. BMC Genomics 14:355. https://doi.org/10.1186/1471-2164-14-355
12. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29(1):24–26. https://doi.org/10.1038/nbt.1754
13. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ (2008) A large genome center's improvements to the Illumina sequencing system. Nat Methods 5(12):1005–1010. https://doi.org/10.1038/nmeth.1270

Chapter 7

Application of High-Throughput Next-Generation Sequencing for HLA Typing on Buccal Extracted DNA

Yuxin Yin, James Lan, and Qiuheng Zhang

Abstract

Next-generation sequencing (NGS) is increasingly recognized for its ability to deliver high-resolution and high-throughput HLA genotyping. As a result, there is active interest in applying NGS technologies to perform high volume bone marrow donor recruitment typing. Currently, buccal-based DNA specimens are considered a noninvasive and cost-effective method for registry typing. Here, we describe the feasibility of using long-range PCR and clonal sequencing by Illumina MiSeq to deliver unambiguous HLA typing on buccal-based donor recruitment samples.

Key words Human leukocyte antigen, Next-generation sequencing, Illumina MiSeq, Donor recruitment samples, Buccal extracted DNA

1 Introduction

The HLA region contains the most highly polymorphic genes in the human genome. HLA genotyping is important in a number of clinical applications, including donor-recipient matching in hematopoietic stem cell transplantation (HSCT) [1, 2], and solid organ transplantation between unrelated individuals [3]. In addition, specific HLA alleles are associated with autoimmune diseases and drug hypersensitivity [4]. Over the years, conventional molecular HLA typing methods include sequence-specific oligonucleotide probes (SSOP), sequence-specific primers (SSP), and sequencing-based typing (SBT). Due to practical limitations of throughput and cost, these methods only focus on the antigen recognition site (ARS) of HLA genes and therefore cannot provide complete phasing of HLA genotypes. The lack of phasing results in cis (same chromosome) and trans (different chromosomes) ambiguities that are difficult, time-consuming, and expensive to resolve [5]. Recently, a number of groups have demonstrated the feasibility of HLA full gene sequencing using next-generation sequencing

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_7,

(NGS) to deliver unambiguous high-resolution typing at high throughput with relatively lower cost [6–9].

Illumina sequencing was introduced in 2006 and has become widely adopted due to its high accuracy and low cost. Similar to Sanger sequencing, this NGS method is built on the principle of sequencing-by-synthesis. A typical sequencing cycle begins with the passage of four *reversible* fluorescent-labeled terminator nucleotides on the flow cell. As one of the nucleotides becomes incorporated during polymerization, the sequencing reaction is briefly terminated to allow base identification through imaging. The fluorescent-labeled nucleotide is then cleaved and the next cycle starts again with the addition of another labeled terminator nucleotide [10]. The Illumina MiSeq platform is capable of producing 150–300 bp paired-end reads with a sequencing capacity of 300 Mb–15 Gb (Table 1). Due to its superior read length and high efficiency, Illumina MiSeq has become a popular choice for HLA genotyping in clinical laboratories.

Buccal swab samples are routinely collected for high volume registry typing. Despite its low cost and convenience, the buccal swab sample has many limitations. First, buccal-derived DNA yield is much lower than that obtained from peripheral blood. It has been reported that epithelial cells recovered from the mouth are usually superficial and about 25% of these cells are in the process of apoptosis [11]. Second, DNA isolated from cheek cells may contain exogenous bacterial DNA that interferes with the downstream PCR efficiency and data analysis [11]. Finally, buccal DNA is prone to nucleic acid degradation, which may limit the success of long-range

Table 1
MiSeq system performance parameters for HLA sequencing

Reagent kit	No. of Reads[a]	Output (max.)	2 × 150 bp Output/total time[b]	2 × 250 bp Output/total time[b]	2 × 300 bp Output/total time[b]
MiSeq Reagent Kit v2	30 M	7.5 Gb	4.5 Gb/~24 h	7.5 Gb/~39 h	N/A
MiSeq Reagent Micro Kit v2	8 M	1.2 Gb	1.2 Gb/~19 h	N/A	N/A
MiSeq Reagent Nano Kit v2	2 M	500 Mb	300 Mb/~17 h	500 Mb/~28 h	N/A
MiSeq Reagent Kit v3	50 M	15 Gb	N/A	N/A	15 Gb/~56 h

bp base pairs, *Mb* megabases, *Gb* gigabases, *M* million
[a]Paired-end reads passing filter
[b]Total times include cluster generation, sequencing, and base calling on a MiSeq system enabled with dual surface scanning

PCR. Here, we describe the development and validation of long-range HLA typing methods using MiSeq on HLA-A, -B, -C, -DRB1, -DQB1, and -DPB1 on buccal specimens for NMDP registry donors. Our primers amplify 5′ UTR to 3′ UTR of class I genes (HLA-A, -B, -C) and exon 2 through part of exon 4 for Class II (HLA-DRB1, -DQB1, -DPB1). Products generated using these primers are able to resolve all third field ambiguities for HLA class I and the majority of second field ambiguities for class II. Despite the challenges associated with buccal samples, robust amplification was achieved using the long-range PCR approach. By analyzing polymorphisms in the entire region of HLA genes, the importance of polymorphisms outside ARS, as well as introns, promoters, enhancers, and UTRs can be further revealed in transplantation, autoimmunity, diseases association, and pharmacogenomics.

2 Automation Equipments and Materials

Prepare all solution using Molecular Biology Grade (MBG) water, or Clinical Lab Reagent (CLR) water obtained from Millipore system (to attain a sensitivity of 18 MΩ-cm at 25 °C), and analytical grade reagents. Prepare and store all reagents at room temperature (unless indicated otherwise). Diligently follow all waste disposal regulations when disposing waste materials.

2.1 Automation Equipments

1. BioRobot Universal System (*see* **Note 1**). Qiagen, Cat. #9001094.
2. Apricot Pipette i-Pipette Pro Series Personal Pipettor™. Apricot Designs, Cat. #i-PP96-125.
3. Biomek NX Multichannel liquid handler. Beckman Coulter, Cat. #A31841.
4. Biomek FX laboratory automation workstation. Beckman Coulter, Cat. #A31844.

2.2 Reagents for Long-Range PCR

1. The Buccal swab DNA is suspended using 150 μL elution buffer (*see* **Note 1**).
2. High Fidelity Polymerase, PrimeSTAR GXL DNA Polymerase. Takara Clonetech. Store at −15 to −25 °C.
3. Multiplex primer sets for Class I (HLA-A, -B, -C) and Class II (HLA-DRB1, -DQB1, -DPB1). Store at −15 to −25 °C.
4. 1× Working TBE Buffer. Dilute 10× stock 1:10 with CLR Water. Store at room temperature.
5. 0.8% agarose gel (*see* **Note 2**).
6. Gel Loading Buffer: 0.05% bromophenol blue, 40% sucrose, 0.1 M EDTA (pH 8.0) and SDS. Sigma-Aldrich. Store at room temperature.

7. Quick-Load® 1 kb DNA Ladder. New England Biolabs. Store at 2–8 °C.

2.3 Reagents for Library Construction

1. 80% Ethanol: Make fresh (*see* **Note 3**). Store at room temperature up to 24 h from preparation.
2. Agencourt AMPure® XP. Beckman Coulter Store at 2–8 °C. Vortex the sample purification beads for at least 1 min or until they are well dispersed.
3. Qubit® dsDNA BR Assay Kit. Life Technologies (ThermoFisher). The range for this broad range kit is 2–1000 ng/μL (*see* **Note 4**).
4. TruSeq Nano DNA HT Library Prep Kit 96 samples. Illumina. Store contents as noted below up to the manufacturer's expiration date. Kit includes three boxes:
   (a) Core Reagent Box: RSB—Resuspension Buffer, ERP2—End Repair Mix, ATL—A-Tailing Mix, LIG2—Ligation Mix 2, STL—Stop Ligation Buffer, PPC—PCR Primer Cocktail. Store at −15 to −25 °C.
   (b) SP Beads Box: Sample Purification Beads. Store at 2–8 °C.
   (c) Adapter Plate Box (*see* **Note 5**): DNA Adapter Plate, 96plex. Store at −15 to −25 °C.
5. Diluted Bead Mixture for 550 bp insert size: Add 92 μL sample purification beads to 92 μL MBG water (Formula for one sample) (*see* **Note 6**).

2.4 Reagents for MiSeq Sequencing Setup and Post Wash

1. 0.2 N Sodium Hydroxide (NaOH): Make fresh (*see* **Note 7**). Store at room temperature.
2. MiSeq v2 reagent kit (500 cycles):
   (a) Box 1 of 2 (store at −15° to −25 °C):
      - Reagent cartridge.
      - Hybridization Buffer (HT1).
   (b) Box 2 of 2 (store at 2–8 °C):
      - Flow cell.
      - Incorporation Buffer (PR2).
3. 12.5 pM Phix Control (*see* **Note 8**). Store at −15 to −25 °C.
4. 0.5% Tween 20: Add 50 mL Tween 20 to 10 L CLR water. Invert several times to mix. Store at room temperature.
5. 10 mM Tris–HCl with 0.1% Tween 20, pH 8.5. Store at room temperature.
6. Absolute ethanol (100%). Store at room temperature.

3 Methods

There are multiple NGS platforms (Roche 454 [12], Ion Torrent [13], Illumina [9, 14], and Pacific Biosciences [15]) available on the market. However, regardless of the methodology, NGS HLA typing processes can be broken down into four major steps: (1) sample preparation; (2) library construction. This step begins with target DNA fragmentation (sonication, nebulization, or enzymatic shearing to fragment DNA into ~600–1000 bp fragments) followed by adaptor ligation and barcoding for multiplexing testing; (3) sequencing; and (4) data analysis [16]. Library construction is labor intensive, time-consuming, and prone to human errors. Therefore, we describe an automated NGS library construction platform to streamline the workflow, increase throughput, and reduce errors. HLA is highly polymorphic, therefore library construction is sensitive to very small amounts of DNA contamination. Thus, exercise extreme care to avoid cross-contamination when handling primers, assembling amplification reactions, during purifications and library construction. Maintain pre- and post-amplification reagents, supplies, and equipment separately. Each plate includes one positive control and one negative control. Carry out automation procedures for library construction (*see* **Note 9**) at room temperature unless otherwise specified.

3.1 Sample Preparation (Long-Range PCR and Gel Check)

1. Label two skirted PCR plates for a full typing: Class I and Class II.
2. Prepare reagent mix for Class I and Class II separately for pre-PCR primers. In brief, each PCR reaction consisted of: 5 μL of buccal swab gDNA (use the Apricot pipette to transfer 5 μL DNA to the designated well of the labeled plates), 4 μL polymerase buffer, 1.6 μL dNTP mixture, 5 μL primer mix, 0.8 μL PrimeSTAR GXL DNA polymerase (TaKaRa Bio Inc., Japan), and add MBG water to make a final volume of 20 μL.
3. Seal with adhesive plate seals.
4. GeneAmp PCR system 9700 (Life Technologies, Carlsbad, CA) thermal cycler conditions for class I genes: 94 °C for 2 min, followed by 35 cycles of (98 °C for 10 s, 70 °C for 3 min). For class II: 94 °C for 2 min, followed by 35 cycles of (98 °C for 10 s, 69 °C for 3 min).
5. Confirm PCR products on 0.8% agarose gel (Sigma–Aldrich, St. Louis, MO). Mix 3 μL of sample with 5 μL loading dye prior to loading samples onto the 0.8% gel.
6. Cover electrophoresis tank and electrophorese for at least 10–12 min at approximately 150 V.

7. Turn off the power and take a digital image of the gel with the Alphalmager HP system (ProteinSimple, San Jose, CA).
8. Mix Agencourt AMPure XP beads (Beckman Coulter, Brea, CA) in 0.6× PCR reaction volume to undergo purification using Biomek NX (Beckman coulter, Brea, CA).
9. Qubit fluorometer 2.0 (Life Technologies, Grand Island, NY) is used to quantitate amplicon concentrations. *This is a safe stopping point. If stopping, seal the plate and store at −15 to −25 °C for up to 7 days.*

3.2 Automated High Throughput Library Preparation

1. Equimolar pooling of class I and class II PCR products on Biomek FX (Beckman coulter, Brea, CA).
2. Amplicons of HLA class I and class II genes are sheared using Covaris M220 (Covaris, Woburn, MA) (*see* **Note 10**). Aim for an insert size of 500–600 bp to maximize the phasing of linked polymorphisms.
3. Amplicon fragments undergo further purification on Biomek FX (Beckman coulter, Brea, CA) with Agencourt AMPure XP beads in 1.6× reaction volume. Transfer 60 μL of clear supernatant from each well to a new plate.
4. Thaw End Repair Mix 2 (ERP2), and add 40 μL of ERP2 to each well. Mix up and down ten times.
5. Seal the plate with an adhesive seal.
6. Place sealed plate on pre-programmed thermocycler (30 °C for 30 min).
7. Dilute sample purification beads in a 15 mL conical tube according to the formula in **Note 6**.
8. Add 160 μL diluted bead mixture to each well of the plate containing 100 μL of end repaired sample. Mix up and down ten times (*see* **Note 11**).
9. Incubate at room temperature for 5 min, and then place the plate on the magnet plate at room temperature for 5 min or until the supernatant is clear.
10. Transfer the supernatant to a new plate which contains 30 μL undiluted sample purification beads to each well for removing of small fragment (*see* **Note 12**).
11. When done, transfer 17.5 μL of clear supernatant from each well to a new plate. *This is a safe stopping point. If stopping, seal the plate and store at −15 to −25 °C for up to 7 days.*
12. Add 12.5 μL thawed A-Tailing Mix to size selected sample. Mix up and down ten times.
13. Seal the plate with an adhesive seal.

14. Place sealed plate (containing 30 μL of each sample) on pre-programmed thermocycler. Close the lid and select the A-tailing program (37 ° C for 30 min).
15. When done, proceed immediately to ligating adapters (*see* **Note 13**).
16. Thaw the DNA Adapter Plate (DAP) for 10 min at room temperature (*see* **Note 14**).
17. Briefly centrifuge the DAP plate for 10 s to collect the adapter to the bottom of the well. Remove the plastic cover.
18. Immediately before use, thaw the Ligation Mix 2 tube.
19. Add 2.5 μL Resuspension Buffer, 2.5 μL Ligation Mix 2 to each well of the plate.
20. Transfer 2.5 μL DNA Adapter from the DAP well to the A-Tail plate. Mix up and down ten times.
21. Seal the plate and centrifuge briefly. Followed by incubation on 30 °C for 10 min.
22. When done, add 5 μL Stop Ligation Buffer to each well to inactivate the ligation.

Followed by ligation cleanup (*see* **Note 15**). Then, transfer 25 μL of clear supernatant to a new plate. *This is a safe stopping point. If stopping, seal the plate and store at −15 to −25 °C for up to 7 days.*

23. Add 5 μL PCR Primer Cocktail and 20 μL Enhanced PCR Mix to each well. Mix up and down ten times.
24. Seal the plate, and place the sealed plate (containing 50 μL of each sample) on a pre-programmed thermocycler to carry out target library enrichment with a limited-cycle PCR: 95 °C for 3 min, followed by 8 cycles of (98 °C for 20 s, 60 °C for 15 s, 72 °C for 30 s), and 72 °C for 5 min.
25. When done, enriched libraries are purified with sample purification beads in a 1:1 ratio. Transfer 30 μL of clear supernatant from each well to the corresponding well of a new PCR plate. *This is a safe stopping point. If stopping, seal the plate and store at −15 to −25 °C for up to 7 days.*

3.3 Loading for MiSeq Sequencing

1. Prepare sufficient Qubit working solution for a number of samples to be run (*see* **Note 4**). Quantify all the libraries.
2. Equimolar pooling of all samples on the Biomek FX (Beckman Coulter, Brea, CA) (*see* **Note 16**).
3. Thaw a tube of HT1 (Hybridization Buffer) at room temperature and then store at 2–8 °C until ready to use.
4. Prepare a fresh dilution of 0.2 N NaOH and use within 12 h (*see* **Note 7**).

5. Add 5 μL sample from pooled libraries to 5 μL 0.2 N NaOH. Vortex briefly and then quick spin. Incubate the tube for 5 min at room temperature to denature the DNA into single strands.
6. Add 750 μL cold HT1 into the tube. This step prepares 20 pM denatured library.
7. Add 360 μL from denatured libraries to 240 μL chilled HT1. Mix by pipetting up and down. Place on ice or at 2–8 °C until needed.
8. Add 5% Phix Control (*see* **Note 8**) to the denatured libraries.
9. Remove reagent cartridge and immerse in a room temperature water bath for approximately 1 h (*see* **Note 17**).
10. Load the denatured libraries into the sample well of the reagent cartridge.
11. Click Sequence on the main screen of the MiSeq software.
12. The BaseSpace (*see* **Note 18**) Options Screen opens to enable storage and analysis. Log into My Illumina account with unique user name and password.
13. Load the flow cell (*see* **Note 19**), reagent cartridge, as well as Incorporation Buffer (PR2) (*see* **Note 20**) into MiSeq.
14. Upload the TruSeq Nano Barcode template (*see* **Note 21**).
15. Click Start Run, when all items successfully pass the pre-check.
16. When the Illumina MiSeq sequencing is completed, post wash the instrument by using 0.5% Tween 20.

3.4 HLA Data Analysis

During Illumina sequencing, a real time base calling and quality scoring are performed. Sequencing *quality scores (Q scores)* measure the probability that a *base* is called incorrectly. For example, a Q score of 30 (Q30) to a base, this is equivalent to the probability of an incorrect base call 1 in 1000 times. At the end of each instrument run, sequence data is exported in a standard FASTQ format. When performing paired-end sequencing, two FASTQ files are generated, with each FASTQ file containing data from each end of the DNA fragment sequenced. Since HLA genes are highly polymorphic, even rare sequencing errors could lead to inaccurate genotyping. In the method described here, Omixon Twin employs two distinct analytical algorithms to ensure high-confidence allele calling. The first algorithm involves alignment of short sequencing reads to the IMGT HLA reference database (https://www.ebi.ac.uk/ipd/imgt/hla/). In contrast, the second algorithm utilizes de novo assembly—since no reference sequences are required, this algorithm is particularly useful in the analysis and detection of novel alleles. In NGS sequencing, coverage *breath* denotes the per-

centage of a target HLA gene that is sequenced, whereas coverage *depth* describes the number of times that a given nucleotide has been sequenced. Coverage is one of the most relevant technical variables in NGS and is typically limited by repetitive sequences, library complexity, and cost considerations. Sufficient coverage *depth* and *breadth* are crucial for accurate HLA genotyping calling.

1. Download the data from BaseSpace or generate the FASTQ data on MiSeq locally.
2. Raw sequence outputs are imported as unpacked and gzipped FASTQ files into HLA specific analysis software, e.g., Omixon Twin software (Omixon, Budapest, HU) (*see* **Note 22**) for read alignment and genotype calling (using IMGT/HLA Database as reference).
3. Click 'Next'. This screen will have default settings and then click 'Next' again.
4. Make sure 'illumina' is selected on this screen and automatic paired data is checked. Click on 'Next'.
5. Once the analysis is done, the results can be viewed by clicking view results (*see* **Note 23**).
6. Any data with green circle should be assigned by best matches from the top icons. Review any questionable results with the supervisor and/or director prior to release.
7. Once completed go to view results from top icons (can press back or forward arrow on the top left hand side of the window to get on the screen of choice).
8. Select the correct name of the batch and export it using excel XML (XLSX) or HML format (*see* **Note 24)**.

4 Notes

1. Buccal swab DNA is isolated using QIAamp 96 DNA Swab BioRobot Kit (Qiagen) on BioRobot Universal System according to the manufacturer's protocol.
2. For example to make a 50 mL gel of 0.8% agarose, use 0.4 g agarose, 50 mL of 1× Working TBE Buffer, and add 5 μL DNA SafeStain (Lamda Biotech).
3. Add 400 mL absolute alcohol to 100 mL MBG water. Store at room temperature.
4. Kit contents: Qubit dsDNA BR Reagent (Component A) 200× concentrate in DMSO. Store desiccated and protected from light at room temperature. Qubit dsDNA BR Buffer (Component B). Store at room temperature. Qubit dsDNA

BR Standard #1 (Component C) 0 ng/μL in TE buffer. Store at 2–8 °C. Qubit dsDNA BR Standard #2 (Component D) 100 ng/μL in TE buffer. Store at 2–8 °C. Qubit Instrument Calibration: Using the reagents from the kit—Broad Range (BR) in use, run the following standards once each day when instrument is in use:

(a) Standard 1: 190 μL Qubit working solution + 10 μL calibration standard #1.

(b) Standard 2: 190 μL Qubit working solution + 10 μL calibration standard #2.

5. No more than 4 freeze/thaw cycles allowed for the adapter plate box to maintain reagent integrity. Mark the date on the box each time thawed. Discard after the fourth thaw.
6. Determine the volumes using the formula, which includes 15% excess for multiple samples.
7. Add 10 μL 1 N Sodium Hydroxide (NaOH) to 40 μL Molecular Biology Grade (MBG) water.
8. Add 5 μL of 0.2 N NaOH (*see* **Note 7**) to 5 μL 4 nM PhiX Lib. Vortex briefly and then quick spin. Incubate the Denatured Control tube for 5 min at room temperature to denature DNA into single strands. Add 990 μL chilled HT1 to the Denatured Control tube. And then add 375 μL of Denatured tube to 225 μL chilled HT1. PhiX can be run as a control to troubleshoot run issues. It can be used to troubleshoot potential library preparation or MiSeq instrument issues and run either as a spike-in of the library run or as a library per se in the case of instrumentation issues.
9. Automated library construction can reduce hands-on time, and can be programmed for various needs of the lab, making it ideal for labs that require high flexibility. It requires extensive validation and optimization to ensure the programs work consistently and accurately. Any changes to library construction protocol requires a new validation.
10. Fragmentation (or shearing) of DNA is not as simple as it might sound and NGS quickly reveals the biases inherent in one protocol versus another. Choosing the shearing method depends on a few factors such as DNA size selection and how much DNA is available, as well as your budget! When we were looking for a system to robustly generate sheared DNA fragments, we evaluated several methods and settled on Covaris. It generates uniform fragments in reasonable time and its closed-tube mode of operation means there is virtually no risk of genomic DNA contamination. If you need to process hundreds of samples you could use enzymatic shearing methods.

11. Exercise extreme care during this step to avoid contamination as this volume is close to the maximum for each well.
12. There are mainly two steps for Dual Bead Size Selection: 1) Bead Selection to Remove Large Fragments—This step is used to bind the large, unwanted fragments to the beads. The supernatant will contain the desired fragments. 2) Bead Selection to Remove Small Fragments and to Bind DNA Target. You also can use Pippin prep for size selection. The Pippin Prep facilitates library construction for the most popular NGS platforms, from 100 bp to 1.5 kb, and cut size may need to be adjusted to target optimal size distribution of fragmented sample and is more and more recommended by Illumina and Ion Torrent for certain workflows. It saves time and effort, and will have a narrower size distribution than the beads-based method.
13. Illumina TruSeq HT kits for DNA support 96 libraries using dual indexing. If you want to pool more than 96 samples, you would use homegrown barcodes. If users choose to use "homegrown" barcodes, they do so at their own risk. Failure to have sufficient diversity in sequence in the barcodes has the strong potential to yield no data.
14. Visually inspect the wells to make sure that they are thawed.
15. Use sample purification beads for two rounds of washes to clean up ligated fragments. The volumes of the beads for the two rounds are 42.5 and 50 μL respectively.
16. Based on the values from the Qubit instrument, normalize the concentration of each sample to 4 nM, and pool equal volume of each normalized library into one Eppendorf tube.
17. Do not allow the water to exceed the fill line printed on the cartridge. Place paper towels on the bench, remove the cartridge from the bath, and gently tap on the towels to remove water from the base of the cartridge. Dry the base with Kimwipes. Invert the cartridge ten times to mix the reagents. Visually inspect for complete thawing and precipitates. Gently tap the cartridge to reduce bubbles.
18. BaseSpace is a cloud-based genomics analysis and storage platform that directly integrates with all Illumina sequencers. The user needs to set up runs from Illumina sequencers and monitor sequencing runs from the web. Stream data to the cloud directly from sequencers and share data instantaneously with anyone in the world.
19. Using plastic forceps, grip the flow cell by the base of the plastic cartridge to remove from storage buffer in the flow cell container. Gently rinse with MBG water. Completely remove excess salt from both the glass and plastic cartridge. Salt in the imaging area can affect imaging. Gently pat-dry the gasket area

and adjacent glass. Using an alcohol wipe, clean the flow cell glass. Make sure that the glass is free of streaks, fingerprints, and lint or tissue fibers. Avoid using the alcohol on the flow cell port gasket. Dry excess alcohol with a lint-free lens cleaning tissue. Visually inspect to make sure that the flow cell ports are free of obstructions and that the gasket is well seated around the flow cell ports.

20. Gently invert to mix and remove the lid. Open the reagent compartment door and raise the sipper handle until it locks into place. Remove the used bottles and empty the waste bottle. Return the waste bottle into position.
21. Use Illumina Experiment Manager program to edit the unique barcode ID for each sample.
22. There are several commercial software products for HLA analysis, such as TypeStream Visual (One Lambda), NGSengine (GenDx), MIA FOR A NGS FLEX (Immucor), Conexio Assign TruSight HLA Analysis Software (Illumina), as well as HLA Twin (Omixon). We used HLA Twin for the buccal swab extracted DNA HLA typing.
23. The Omixon Twin software uses two algorithms to assign genotype. When the two algorithms do not match, the result is red flagged. Common causes: low coverage, allele imbalance, and others.
24. Consensus requires a minimum of 10× coverage. If the coverage is less than 10×, no consensus will be displayed for the region.

Acknowledgment

This work was supported by NMDP for RFQ#C14-0040 and UCLA immunogenetics Center for HLA NGS-based development.

References

1. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, Fernandez-Vina M, Flomenberg N, Horowitz M, Hurley CK, Noreen H, Oudshoorn M, Petersdorf E, Setterholm M, Spellman S, Weisdorf D, Williams TM, Anasetti C (2007) High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood 110(13):4576–4583. https://doi.org/10.1182/blood-2007-06-097386
2. Flomenberg N, Baxter-Lowe LA, Confer D, Fernandez-Vina M, Filipovich A, Horowitz M, Hurley C, Kollman C, Anasetti C, Noreen H, Begovich A, Hildebrand W, Petersdorf E, Schmeckpeper B, Setterholm M, Trachtenberg E, Williams T, Yunis E, Weisdorf D (2004) Impact of HLA class I and class II high-resolution matching on outcomes of unrelated donor bone marrow transplantation: HLA-C mismatching is associated with a strong adverse effect on transplantation outcome. Blood

104(7):1923–1930. https://doi.org/10.1182/blood-2004-03-0803

3. Loupy A, Lefaucheur C, Vernerey D, Prugger C, Duong van Huyen JP, Mooney N, Suberbielle C, Fremeaux-Bacchi V, Mejean A, Desgrandchamps F, Anglicheau D, Nochy D, Charron D, Empana JP, Delahousse M, Legendre C, Glotz D, Hill GS, Zeevi A, Jouven X (2013) Complement-binding anti-HLA antibodies and kidney-allograft survival. N Engl J Med 369(13):1215–1226. https://doi.org/10.1056/NEJMoa1302506
4. Thorsby E, Lie BA (2005) HLA associated genetic predisposition to autoimmune diseases: genes involved and possible mechanisms. Transpl Immunol 14(3–4):175–182. https://doi.org/10.1016/j.trim.2005.03.021
5. Lan JH, Zhang Q (2015) Clinical applications of next-generation sequencing in histocompatibility and transplantation. Curr Opin Organ Transplant 20(4):461–467. https://doi.org/10.1097/MOT.0000000000000217
6. Shiina T, Suzuki S, Ozaki Y, Taira H, Kikkawa E, Shigenari A, Oka A, Umemura T, Joshita S, Takahashi O, Hayashi Y, Paumen M, Katsuyama Y, Mitsunaga S, Ota M, Kulski JK, Inoko H (2012) Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. Tissue Antigens 80(4):305–316. https://doi.org/10.1111/j.1399-0039.2012.01941.x
7. Lange V, Bohme I, Hofmann J, Lang K, Sauter J, Schone B, Paul P, Albrecht V, Andreas JM, Baier DM, Nething J, Ehninger U, Schwarzelt C, Pingel J, Ehninger G, Schmidt AH (2014) Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics 15:63. https://doi.org/10.1186/1471-2164-15-63
8. Wang C, Krishnakumar S, Wilhelmy J, Babrzadeh F, Stepanyan L, Su LF, Levinson D, Fernandez-Vina MA, Davis RW, Davis MM, Mindrinos M (2012) High-throughput, high-fidelity HLA genotyping with deep sequencing. Proc Natl Acad Sci U S A 109(22):8676–8681. https://doi.org/10.1073/pnas.1206614109
9. Yin Y, Lan JH, Nguyen D, Valenzuela N, Takemura P, Bolon YT, Springer B, Saito K, Zheng Y, Hague T, Pasztor A, Horvath G, Rigo K, Reed EF, Zhang Q (2016) Application of high-throughput next-generation sequencing for HLA typing on buccal extracted DNA: results from over 10,000 donor recruitment samples. PLoS One 11(10):e0165810. https://doi.org/10.1371/journal.pone.0165810
10. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17(6):333–351. https://doi.org/10.1038/nrg.2016.49
11. Rudney JD, Chen R (2006) The vital status of human buccal epithelial cells and the bacteria associated with them. Arch Oral Biol 51(4):291–298. https://doi.org/10.1016/j.archoralbio.2005.09.003
12. Bentley G, Higuchi R, Hoglund B, Goodridge D, Sayer D, Trachtenberg EA, Erlich HA (2009) High-resolution, high-throughput HLA genotyping by next-generation sequencing. Tissue Antigens 74(5):393–403. https://doi.org/10.1111/j.1399-0039.2009.01345.x
13. Barone JC, Saito K, Beutner K, Campo M, Dong W, Goswami CP, Johnson ES, Wang ZX, Hsu S (2015) HLA-genotyping of clinical specimens using Ion Torrent-based NGS. Hum Immunol 76(12):903–909. https://doi.org/10.1016/j.humimm.2015.09.014
14. Lan JH, Yin Y, Reed EF, Moua K, Thomas K, Zhang Q (2015) Impact of three Illumina library construction methods on GC bias and HLA genotype calling. Hum Immunol 76(2–3):166–175. https://doi.org/10.1016/j.humimm.2014.12.016
15. Albrecht V, Zweiniger C, Surendranath V, Lang K, Schofl G, Dahl A, Winkler S, Lange V, Bohme I, Schmidt AH (2017) Dual redundant sequencing strategy: full-length gene characterisation of 1056 novel and confirmatory HLA alleles. HLA 90(2):79–87. https://doi.org/10.1111/tan.13057
16. Carapito R, Radosavljevic M, Bahram S (2016) Next-generation sequencing of the HLA locus: methods and impacts on HLA typing, population genetics and disease association studies. Hum Immunol 77(11):1016–1023. https://doi.org/10.1016/j.humimm.2016.04.002

Chapter 8

Super High Resolution for Single Molecule-Sequence-Based Typing of Classical HLA Loci Using Ion Torrent PGM

Takashi Shiina, Shingo Suzuki, Jerzy K. Kulski, and Hidetoshi Inoko

Abstract

Super high resolution-single molecule-sequence-based typing (SS-SBT) is an HLA DNA typing method to the field 4 level of allelic resolution (formerly known as 8-digit typing) to efficiently detect novel and null alleles without phase ambiguity by combination of long ranged PCR amplification and next-generation sequencing (NGS) technologies. In this chapter, we describe three basic steps, long ranged PCR, NGS, and allele assignment.

Key words HLA, Next-generation sequencing (NGS), Ion Torrent PGM, long ranged polymerase chain reaction (PCR), Genotyping, Super high resolution single molecule—sequence-based typing (SS-SBT)

1 Introduction

The sequence-based typing (SBT) based on the Sanger method [1] and the PCR-based sequence-specific oligonucleotides (SSOs) methods [2] are the main HLA genotyping methods that are applied currently for clinical use such as hemopoietic stem cell and organ transplantation and disease association. However, both methods often detect more than one pair of unresolved HLA alleles because of chromosomal phase (*cis/trans*) ambiguity [3]. In contrast, next-generation sequencing (NGS) can determine a precise HLA allele sequence derived from a single DNA molecule with a high level of parallelism [4, 5]. Also, NGS techniques are expected to be more effective for high-throughput genotyping of HLA genes [6–9]. For example, the Ion Torrent Personal Genome Machine (PGM) system can produce 3–5 million sequencing reads with a read length of 400–500 bp per read that translates into genotyping up to eight different HLA genes per sequencing run per 27 individuals (*see* **Note 1** and Table 1).

Recently, we have developed the super high resolution-single molecule-sequence-based typing (SS-SBT) method for eight

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_8,

Table 1
Specification of the Ion PGM system

Ion Chip	Ion 314 Chip v2 BC	Ion 316 Chip v2 BC	Ion 318 Chip v2 BC
Number of wells	1.2 million	6.3 million	11.3 million
Number of reads	0.4–0.6 million	2–3 million	3–5 million
Throughput[a]	60–100 M	0.6–1 G	1.2–2 G
Run time	3.7 h	4.9 h	7.3 h
Analysis time[b]	4–5 h[a]	6–7 h[a]	9–10 h[a]

[a]A case of 400 base run
[b]Total time of run time plus analysis time

classical HLA loci, HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1, and HLA-DPB1 (Fig. 1) in combination with NGS platforms such as Ion PGM [10–12].

The SS-SBT method allows sequencing of the entire HLA gene region from the promoter-enhancer region to 3′ untranslated region (3′UTR) to solve the problem of phase ambiguity along with field 4 (8-digit) level typing that includes nucleotide differences in the coding regions as well as the noncoding regions resulting in the detection of novel and null HLA allele discovery. The SS-SBT method is largely divided into three basic steps, long ranged PCR, sequencing clonally amplified single DNA molecules, and allele assignment by using bioinformatics and computing for accurate phase alignments (Fig. 2).

In this chapter, we describe three basic steps, long ranged PCR, NGS, and allele assignment, of the SS-SBT method that we are using in our research as a non-commercial method, and as an example, we provide the long ranged primers that we designed and used in 2012 [10].

2 Materials

To avoid contamination by spurious DNA fragments from previous sample processing steps amplification setup should be done in a pre-PCR room. The pre-PCR kits/reagents should be stored separately and away from the post-PCR room. In our laboratory all pipettors, pipet tips with aerosol barriers, and widely used reagents and instruments are prepared and stored independently in both the pre- and post-PCR rooms. Prepare all solutions using ultrapure

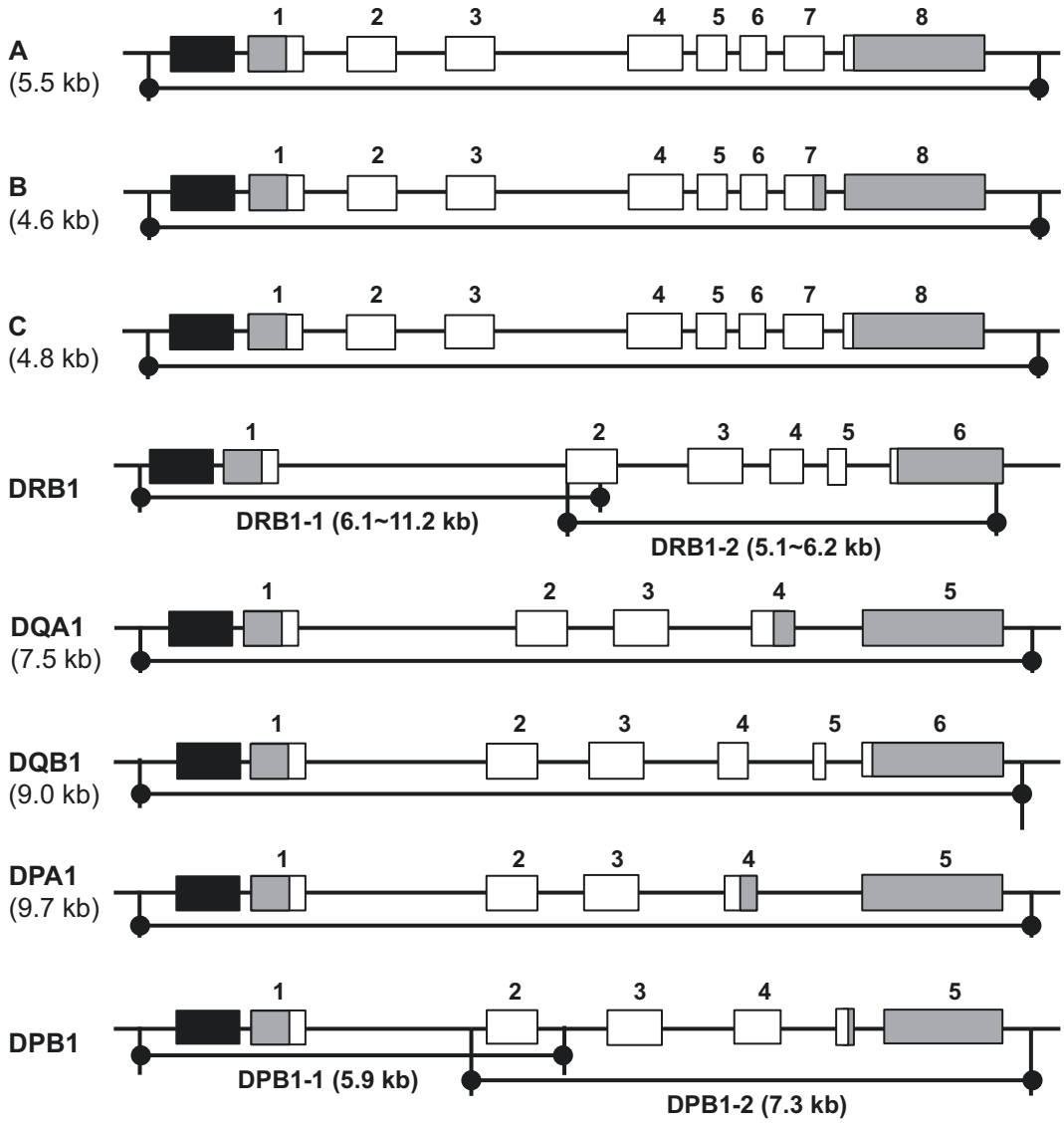


Fig. 1 Outline of the PCR regions in eight HLA loci. White, gray and black boxes indicate coding exons, 5′UTR and 3′UTR and enhancer-promoter regions, respectively. Numbers around the boxes are exon numbers

water (prepared by purifying deionized water to attain a sensitivity of 18 MΩ-cm at 25 °C) and analytical grade reagents. Prepare and store all reagents at room temperature (unless indicated otherwise). Our frequently used general and specialized instruments, reagents and kits are indicated in square bracket.

2.1 Commonly Used Materials

1. Thermal Cycler.
2. Vortexer with a rubber platform.
3. Microcentrifuge with rotor for 2 mL tubes (capable of >15,500 × *g*).
4. PCR tubes (0.2 mL) or 96 well 0.2 mL PCR plates.

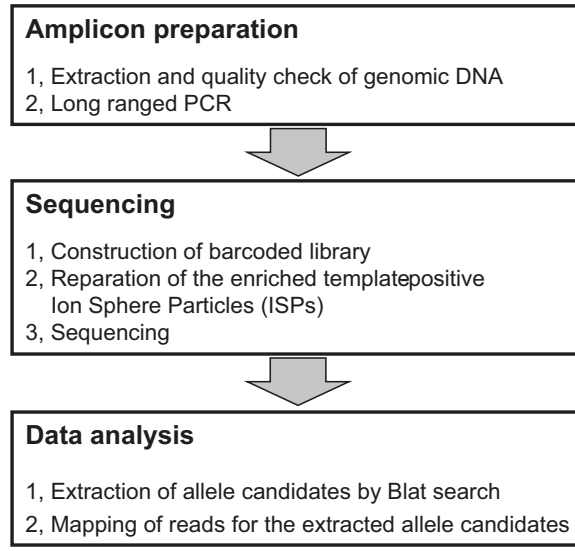


Fig. 2 A schematic workflow of the successive steps of the SS-SBT method

5. Nuclease-free Water.
6. Ethanol, 70%: Molecular biology grade (preparation just before the experiment).

2.2 Extraction and Quality Check of Genomic DNA

1. Genomic DNA extraction kit: [QIAamp DNA Blood Mini Kit (QIAGEN, Germany).] The Kit includes the following things: QIAamp Mini Spin Columns, Collection Tubes (2 mL), Buffer AL, Buffer ATL, Buffer AW1, Buffer AW2, Buffer AE, QIAGEN Protease and Protease Solvent.
2. 1.5 mL microcentrifuge tubes.
3. Water bath or heating block at 56 °C.
4. Spectrophotometer: [Nano Photometer (Implen)].
5. Submarine type electrophoresis device: [Mupid-2 (Mupid Co. Ltd.)].
6. TBE buffer (×10) to be used at 0.5×: 1 M Tris base, 1 M Boric acid, 0.02 M EDTA (disodium salt). Sterilize the solution with autoclave. Dilute ×10 buffer 1 in 20 to the working concentration of 0.5× when required.
7. Ethidium bromide (EtBr) solution: 10% Ethidium Bromide (w/v).
8. Agarose gel: 1% agarose, 100 mL 0.5× TBE buffer, 1 μL EtBr solution.
9. Gel loading buffer (×10): 0.25% Bromophenol blue (w/v), 0.25% Xylene cyanol FF (w/v), 5 mM EDTA, 30% Glycerol (v/v).
10. DNA size marker.

2.3 Long Ranged PCR

1. Long ranged primers (4 pmol/μL dissolved in TE buffer) described in Table 1 of the reported article [10]. The primer sequences shown here are one case, and any customized primers will be useful for this protocol.
2. Low Tris-EDTA (TE) buffer: 10 mM Tris pH 8.0, 0.1 mM EDTA.
3. 1.5 mL DNA LoBind Microcentrifuge Tubes (Eppendorf).
4. High fidelity DNA polymerase for long ranged PCR: [PrimeSTAR GXL DNA polymerase (TaKaRa Bio)] (*see* **Note 2**). 5× PrimeSTAR GXL Buffer (5 mM Mg^{2+}) and 2.5 mM of each dNTP are also attached. Store at −20 °C.
5. DNA purification reagent: [Agencourt AMPure XP (Beckman Coulter)]. Store at 4 °C.
6. Magnetic rack: [DynaMag-2 Magnet or 16–Position Magnetic Stand (Thermo Fisher Scientific)].
7. dsDNA quantitation assay: [Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific)]. The Kit includes 20× TE buffer and Lambda DNA standard. Store at 4 °C.
8. Microtiter plate, 96 well: [Microtiter Assembly Breakable Strip 1 × 8 (Thermo Fisher Scientific)].
9. Microplate reader: [Fluoroskan Ascent micro-plate fluorometer (Thermo Fisher Scientific)].

2.4 Construction of Barcoded Library

1. 1.5 mL DNA LoBind Microcentrifuge Tubes (Eppendorf).
2. DNA shearing machine: [Covaris M220 Focused-ultrasonicator (Covaris)].
3. DNA shearing tube: Covaris microTUBE Screw-Cap Tube (Covaris)].
4. Low TE buffer: 10 mM Tris pH 8.0, 0.1 mM EDTA.
5. Ion Plus Fragment Library kit (Thermo Fisher Scientific). The Kit includes the following reagents: 5× End Repair Buffer, End Repair Enzyme, 10× Ligase Buffer, DNA Ligase, Nick Repair Polymerase, dNTP Mix, Adapters, Platinum™ PCR SuperMix High Fidelity, Library Amplification Primer Mix and Low TE. Store at −20 °C.
6. DNA purification reagent: [Agencourt AMPure XP (Beckman Coulter)]. Store at 4 °C.
7. Magnetic rack: [DynaMag-2 Magnet or 16–Position Magnetic Stand (Thermo Fisher Scientific)].
8. Ion Xpress Barcode Adapter Kit (1 barcode adapter per library) (Thermo Fisher Scientific).
9. E-Gel iBase unit and E-Gel™ Safe Imager transilluminator combo kit (Thermo Fisher Scientific).

10. E-Gel SizeSelect 2% Agarose Gel (Thermo Fisher Scientific). Store at 4 °C.
11. 50-bp DNA Ladder (Thermo Fisher Scientific). Store at −20 °C.
12. Agilent 2100 Expert Bioanalyzer with IKA vortex (Agilent Technologies).
13. Agilent High Sensitivity DNA Kit (Agilent Technologies). Store at 4 °C.
14. Centrifuge with rotor for PCR Plate.

2.5 Preparation of Template-Positive Ion Sphere Particles (ISPs)

1. Ion PGM Hi-Q OT2 Kit (Thermo Fisher Scientific). The Kit is composed of three components: (1) Ion PGM OT2 Supplies, (2) Ion PGM Hi-Q View OT2 Reagents, and (3) Ion PGM Hi-Q OT2 Solutions. Store (1) and (3) at 15–30 °C and store (2) at −20 °C (2–8 °C after thawing).
2. Ion OneTouch 2 System: The system includes Ion OneTouch 2 Instrument and Ion OneTouch ES Instrument (Thermo Fisher Scientific).
3. Ion PGM Enrichment Beads (Thermo Fisher Scientific).
4. 1.5 mL DNA LoBind Microcentrifuge Tubes (Eppendorf).
5. Heat block for 1.5 mL tube at 50 °C.
6. Magnetic rack: [DynaMag-2 Magnet or 16–Position Magnetic Stand (Thermo Fisher Scientific)].
7. 1 M NaOH (10 N): Molecular biology grade.

2.6 Sequencing Run

1. Ion PGM Hi-Q View Sequencing Kit (Thermo Fisher Scientific). The Kit is composed of the following four components: (1) Ion PGM Sequencing Supplies, (2) Ion PGM Hi-Q View Sequencing Reagents, (3) Ion PGM™ Hi-Q View Sequencing Solutions, and (4) Ion PGM Hi-Q Sequencing dNTPs. Store (1) and (3) at 15–30 °C and store (2) and (4) at −20 °C.
2. Ion Chip kits, Ion 318, Chip v2 BC, Ion 316 Chip v2 BC or Ion 314 Chip v2 BC (Thermo Fisher Scientific).
3. Tank of compressed nitrogen.
4. NaOH (10 M): Molecular biology grade.
5. At least 4 L of fresh 18-MΩ water.
6. 0.22 or 0.45 μm vacuum filtration system and filters (nylon or PVDF filters, 1 L volume).
7. 15 mL conical tubes.
8. 1.5 mL DNA LoBind Microcentrifuge Tubes (Eppendorf).
9. Glass bottle (1 L).
10. Graduated cylinders (1 or 2 L volume).

2.7 Allele Assignment

1. Macintosh computer (recommended computer specifications = RAM 16 GB, Hard Drive 500 GB).
2. Linux OS (e.g., CentOS).
3. Blat software for Linux [13].
4. Reference Mapper (Roche).
5. fastq2fasta.pl file format conversion program [14].
6. HLA allele sequences downloaded [15].

3 Methods

3.1 Method for Extraction and Quality Check of Genomic DNA

1. Extract genomic DNA from 200 μL fresh peripheral blood cells in accordance with the protocol of the QIAamp DNA Blood Mini Kit [16].
2. At the final step elute the DNA from QIAamp Mini Spin Column with volumes of 200 μL Nuclease-free Water.
3. Measure the DNA concentration by absorbance at 260 nm and the purity by an A_{260}/A_{280} ratio using a spectrophotometer. Calibrate the spectrophotometer using Nuclease-free Water before measurement of the DNA sample (*see* **Note 3**).
4. Gently mix 1 μL Gel loading buffer and 1 μL extracted genomic DNA solution in a new tube or plate, and spin down the solution. Load the mixture into a well of 1% agarose gel, and run at 100 V for 30 min. Using a Mupid-2 submarine type electrophoresis device.
5. Check the length of genomic DNA by comparison with molecular weights of DNA size marker (*see* **Note 4**).

3.2 Long Ranged PCR

1. Make the primer mixture at the ratios described in Tables 2 and 3. Use the low TE buffer for dilution of the primers.
2. Adjust the DNA concentration for 20–30 ng/μL with Nuclease-free Water.
3. Add the diluted DNA solution, 1 U PrimeSTAR GXL DNA polymerase, 4.0 μL 5× PrimeSTAR GXL buffer (5 mM Mg^{2+}), 1.6 μL 2.5 mM of each dNTP, 2.0–7.0 μL (4 pmol/μL) of each primer mixture with the 20 μL PCR amplification-reaction-volume in a 0.2 mL PCR tube or a 96-well PCR plate on ice. Optionally, when the sample size is large, set up the appropriate volume of the amplification master mixture into a 1.5 mL tube.
4. Gently mix by pipetting (do not use a vortexer) and spin down the reaction mixture, and place it back on ice.
5. Turn on the thermal cycler to warm up the heating lid or block, and set the thermal cycler's "ramp speed" to the 9600

Table 2
Composition of primer mixture

HLA locus	PCR region described in Fig. 1	Volume of primer mixture per reaction	Sense primer name	Primer volume per reaction (4 pmol/μL)	Anti-sense primer name	Primer volume per reaction (4 pmol/μL)
HLA-A	A	2.0 μL	A_F1	0.5 μL	A_R1	1.0 μL
			A_F2	0.5 μL		
HLA-B	B	2.0 μL	B_F	1.0 μL	B_R	1.0 μL
HLA-C	C	2.0 μL	C_F1	1.0 μL	C_R1	1.0 μL
HLA-DRB1	DRB1–1	4.5 μL	DRB1_PE2-F1	0.5 μL	DRB1_PE2-R1	0.5 μL
			DRB1_PE2-F2	0.5 μL	DRB1_PE2-R2	0.5 μL
			DRB1_PE2-F3	0.5 μL	DRB1_PE2-R3	0.5 μL
					DRB1_PE2-R4	0.5 μL
					DRB1_PE2-R5	0.5 μL
					DRB1_PE2-R6	0.5 μL
	DRB1–2	7.0 μL	DRB1-E2-1.1-F	3.6 μL (*see* Table 2B)	DRB1-E2-12-R	3.4 μL (*see* Table 2B)
			DRB1-E2-1.1-F		DRB1-E2-3568-R	
			DRB1-E2-1.2-F		DRB1-E2-4-R	
			DRB1-E2-2-F		DRB1-E2-7-R2	
			DRB1-E2-3568-F		DRB1-E2-9-R	
			DRB1-E2-4-F		DRB1-E2-10-R	
			DRB1-E2-7-F4			
			DRB1-E2-9-F			
			DRB1-E2-10-F			
HLA-DQA1	DQA1	2.0 μL	DQA1_F	1.0 μL	DQA1_R	1.0 μL

(continued)

Table 2
(continued)

HLA locus	PCR region described in Fig. 1	Volume of primer mixture per reaction	Sense primer name	Primer volume per reaction (4 pmol/μL)	Anti-sense primer name	Primer volume per reaction (4 pmol/μL)
HLA-DQB1	DQB1	5.0 μL	DQB1-F3.1	1.0 μL	DQB1-R3.1	1.0 μL
			DQB1-F3.2	1.0 μL	DQB1-R3.2	1.0 μL
					DQB1-R3.3	1.0 μL
HLA-DPA1	DPA1	2.0 μL	DPA1_F	1.0 μL	DPA1_R	1.0 μL
HLA-DPB1	DPB1-1	2.0 μL	DPB1_F1	1.0 μL	DPB1_R1	1.0 μL
	DPB1-2	2.0 μL	DPB1_F2	1.0 μL	DPB1_R2	1.0 μL

Table 3
Composition of the DRB1-1 primer mixtures

Sense primer name	Primer volume (4 pmol/μL)	Anti-sense primer name	Primer volume (4 pmol/μL)
DRB1-E2-1.1-F	1 μL	DRB1-E2-12-R	2 μL
DRB1-E2-1.2-F	1 μL	DRB1-E2-3568-R	2 μL
DRB1-E2-2-F	4 μL	DRB1-E2-4-R	1 μL
DRB1-E2-3568-F	4 μL	DRB1-E2-7-R2	2 μL
DRB1-E2-4-F	2 μL	DRB1-E2-9-R	4 μL
DRB1-E2-7-F4	4 μL	DRB1-E2-10-R	2 μL
DRB1-E2-9-F	8 μL		
DRB1-E2-10-F	4 μL		
Total	28 μL		13 μL

Mix sense primers and anti-sense primers independently. Of them use 3.6 μL sense primer mixture and 3.4 μL anti-sense primer mixture per reaction

program when the 9700 Thermal Cycler GeneAmp PCR System 9700 (Thermo Fisher Scientific) is used.

6. Place the reaction tubes or plate into the thermal cycler, and run with the locus-specific cycling parameters as described in Table 4 (*see* **Note 5**).
7. After the PCR reaction, add 36 μL Agencourt AMPure XP Beads solution to 20 μL PCR product.

8. Purify the PCR product in accordance with the manufactural protocol of the Agencourt AMPure XP Beads [17].
9. At the final step (elution step) of the protocol elute the PCR product with 20 μL Nuclease-free Water.
10. Check that the presence of the PCR product is at expected molecular size by agarose gel electrophoresis shown in Subheading 3.1, **step 4** (*see* **Note 6**).
11. Dilute 1 μL of the purified PCR product with 99 μL 1× TE buffer.
12. Quantify the diluted PCR product in accordance with the manufacturer's protocol of the Picogreen assay [18].
13. Measure the fluorescence intensity of the samples using a Fluoroskan Ascent micro-plate fluorometer (excitation ~480 nm, emission ~520 nm). Prepare a standard curve from fluorescence emission intensity of the Lambda DNA standard, and calculate the DNA concentration (μg/uL) from fluorescence emission intensity of the purified PCR product by comparison with the standard curve.
14. Calculate the molar concentration from the quantified PCR product using a following formula:
    Molar concentration (pM) = DNA concentration (μg/uL)/ nucleotide length of PCR region (bp) × 1000 × 1.52.

Table 4
PCR conditions for long ranged PCR

		Denature	Annealing	Extension
PCR region	**The first denature**	**30 cycles**		
A	94 °C, 2 min	98 °C, 10 s	60 °C, 20 s	68 °C, 5 min
B			60 °C, 20 s	68 °C, 5 min
C			60 °C, 20 s	68 °C, 5 min
DRB1-1			70 °C, 5 min	
DRB1-2			70 °C, 5 min	
DQA1			68 °C, 7 min	
DQB1			70 °C, 9 min	
DPA1			70 °C, 9 min	
DPB1-1			70 °C, 5 min	
DPB1-2			70 °C, 5 min	

15. If several PCR gene products (e.g., A, B, C, DRB1-2) have been amplified separately from the same genomic DNA sample, pool the purified PCR products with the equimolar concentrations in a 1.5 mL DNA LoBind Microcentrifuge Tube (*see* **Note 7**).

3.3 Construction of Barcoded Library

1. Adjust 100 ng of the pooled PCR product to 50 μL with low TE in a 1.5 mL DNA LoBind Microcentrifuge Tube.
2. Transfer the pooled PCR product to a microTUBE, and place it into the Covaris M220 machine to fragment the DNA by shearing.
3. Select the "Ion_Torrent_400bp_50μl_ScrewCap_microTUBE" protocol in the SonoLab software, and click on "Run" in the SonoLab software to fragment the DNA.
4. After the fragmentation step, transfer the sheared DNA into a new 1.5 mL LoBind Microcentrifuge Tube, and add 29 μL Nuclease-free Water to the fragmented DNA (total volume: 79 μL).
5. Perform the following steps of end-repair and DNA purification, adapter ligation, nick-repair and purification of the ligated DNA, size-selection of the library with the E-Gel™ SizeSelect Agarose Gel, and amplification and purification the library according to the manufacturer's protocol for using the Ion Plus Fragment Library kit [19].
6. After completing the above procedures with the Ion Plus Fragment Library kit, 20 μL of a final barcoded library (in low TE buffer) is generated. Add 4 μL of low TE buffer to 1 μL of the library into a new 0.2 μL PCR tube to assess the quality and quantity of the library.
7. Characterize the size distribution of barcoded library, and determine the molar concentration in pmol/L of each barcoded library in accordance with the manufacturer's protocol for the use of the Agilent High Sensitivity DNA Kit [20] (*see* **Note 8**).
8. Review the run data. Usual integration area is 300–1000 bp. Record the concentration under the peak as this will be used to calculate the pool sizes.
9. Calculate the molar concentration (pM) within the usual integration area of 300–1000 bp using the Bioanalyzer software.
10. Determine the dilution factor that will give a concentration of 100 pM using the following formula:

    Dilution factor = (Library concentration in pM)/100 pM

    (*see* **Note 9**).
11. Mix each barcoded library into a single tube at equimolar concentrations if there are several libraries.

3.4 Preparation of the Enriched Template-Positive Ion Sphere Particles (ISPs)

1. Prepare 25 μL of the diluted library by mixing of 5 μL 100 pM barcoded library and 20 μL Nuclease-free Water on ice (1:5 dilution).
2. Proceed sequentially with the following steps: set up the Ion OneTouch 2 Instrument, prepare the reaction reagent, install the reaction filter with the reaction reagent for the OneTouch 2 Instrument, switch on and run the instrument to completion in order to enrich the diluted library sample with the template-positive ISPs according to the manufacturer's protocol using the Ion PGM Hi-Q OT2 Kit [21] for the Ion OneTouch 2 system.
3. After the sample of single-stranded DNA templates has been enriched with the ISPs, ensure that the enriched ISPs pellet is pipetted with >200 μL Melt-Off Solution to disperse the ISPs. Optionally, store the enriched ISPs at 4 °C for up to 3 days.

3.5 Sequencing

1. Perform the following presequencing and sequencing steps by using the Ion PGM Hi-Q Sequencing Kit [22] according to the manufacturer's instructions: prepare and generate a sequencing run by using Torrent Suite Software to clean and initialize the Ion PGM, and load the sample onto the Ion Chip v2 BC.
2. The appropriate Ion Chip v2 BC for sequencing (Table 1) is selected based on the number of prepared barcoded libraries and total PCR sizes of each library. The required throughput (Mb) per library can be calculated by using the following formula:
   Required throughput (Mb) = Total PCR length × 100–300 (Read depth per allele) × 2 (allele numbers)/1000.
3. Calculate the library numbers for loading the sequencing chip from the calculated required throughput (Mb) (*see* **Note 1**).

3.6 Allele Assignment

We developed a new allele sequence assignment program (Sequence Alignment Based Assigning Software; SeaBass) [12, 23] for NGS data analysis that includes and provides (1) output of sequence reads, (2) homology search using the Blat software with the "match" variable set to 100% to detect identical exons within the known HLA alleles released from the IPD-IMGT/HLA database, (3) selection of allele candidates, (4) mapping of the sequence reads to the selected allele candidates as references with the "match" set at 100% using Reference Mapper (Roche), (5) calculation of coverages, and (6) confirmation of the mapping data and allele assignment (Fig. 3). The operations from (2) to (5) are processed automatically.

3.6.1 Extraction of Allele Candidates by Blat Search

1. Install Linux OS, Blat program, Reference Mapper, and fastq-2fasta.pl into a Macintosh computer.
2. Make multi-fasta files for each exon of all the HLA loci from the IPD-IMGT/HLA database.

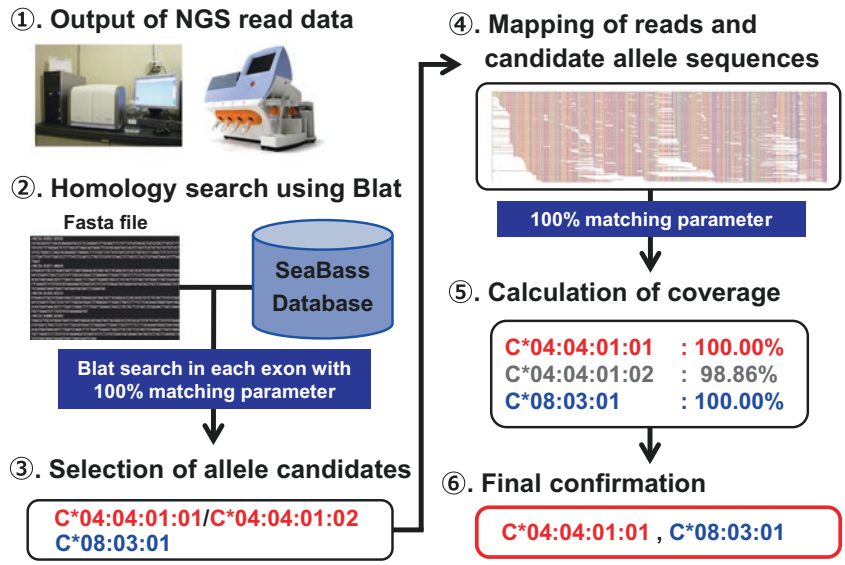


Fig. 3 Allele assignment method using the newly developed Sequence Alignment Based Assigning Software, SeaBass

3. Convert the filename extension ".fastq" outputted from Ion PGM to ".fas".
   Command line: *username$ ./fastq2fasta.pl -a IonPGM-out.fastq > IonPGM-out.fas*
4. Search for HLA allele candidates included in IonPGM-out.fas by comparing every exon (e.g., exon 1 to exon 7 in HLA-C) sequence with the allele sequence data using the Blat sequence alignment program with the parameter set to 100% matching.
   Command line: *./blat HLA-C_exon1.fas IonPGM-out.fas -noHead -minScore=73 -minIdentity=100 -out_HLA-C_exon1.psl* (*see* **Note 10**).
5. Sort each of the unique HLA allele names into descending order in column 14 of outputfilename.psl that shows the HLA allele name that perfectly matches with each read. Repeat this procedure for every exon (e.g., exon 1 to exon 7 in HLA-C).
   Command line: *username$ cut -f14 -out_HLA-C_exon1.psl | sort --u > -out_HLA-C_exon1.txt.*
6. Extract out the common HLA allele sequences in all of the exons (e.g., exon 1 to exon 7 in HLA-C) using the following strategy. The alleles called in all exons are the most common allele candidates rather than the actual allele (Fig. 4).

   (a) Extraction of common HLA alleles in exons 1 and 2.
   Command line: *username$ cat -out_HLA-C_exon1.txt -out_HLA-C_exon2.txt | sort | uniq -d > -out_HLA-C_exon1-2_common.txt*

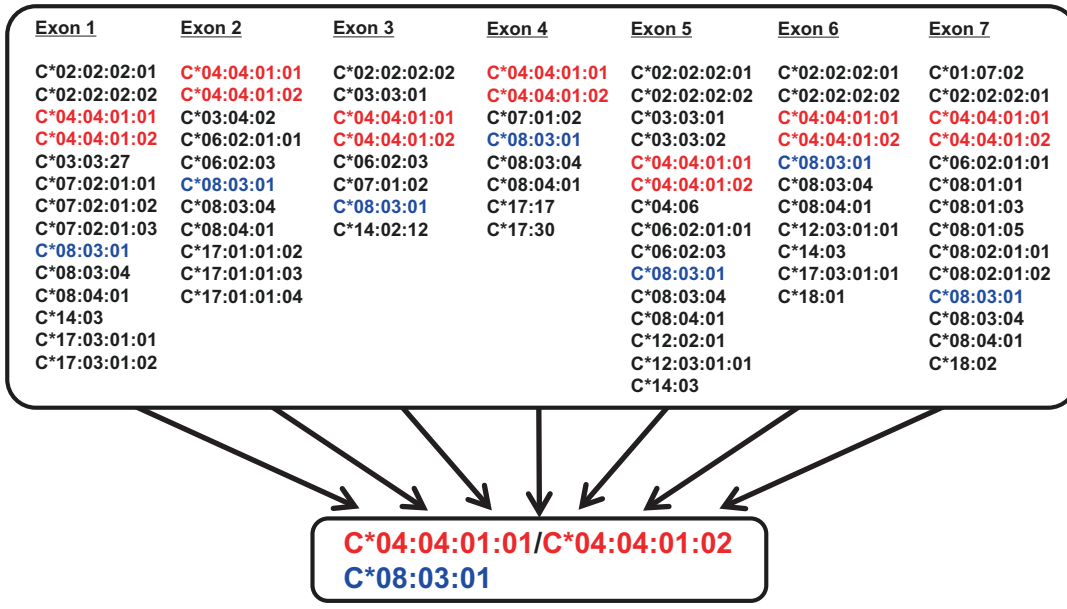


Fig. 4 Extraction of allele candidates by the Blat sequence alignment search. The allele candidates that are the same for each exon are shown either in red or blue letters

(b) Extraction of common HLA alleles in exons 3 and 4.
Command line: *username$ cat -out_HLA-C_exon3.txt -out_HLA-C_exon4.txt | sort | uniq -d > -out_HLA-C_exon3-4_common.txt*

(c) Extraction of common HLA alleles in exons 5 and 6.
Command line: *username$ cat -out_HLA-C_exon5.txt -out_HLA-C_exon6.txt | sort | uniq -d > -out_HLA-C_exon5-6_common.txt*

(d) Extraction of common HLA alleles from exons 1 to 4 using the common HLA alleles extracted in the processes 7-1 and 7-2.
Command line: *username$ cat -out_HLA-C_exon1-2_common.txt -out_HLA-C_exon3-4_common.txt | sort | uniq -d > -out_HLA-C_exon1-4_common.txt*

(e) Extraction of common HLA alleles from exons 1 to 6 using the common HLA alleles extracted in the processes 7-3 and 7-5.
Command line: *username$ cat -out_HLA-C_exon1-4_common.txt -out_HLA-C_exon5-6_common.txt | sort | uniq -d > -out_HLA-C_exon1-6_common.txt*

(f) Extraction of common HLA alleles from exons 1 to 7 using the common HLA alleles extracted in process 7-6 and from exon 7.

Table 5
New allele detection by Blat search

Exon	Allele 1	Allele 2
Exon 1	B*15:18:01	B*44:03:01
Exon 2	B*15:18:01	B*44:03:01
Exon 3	B*15:18:01	Not detected
Exon 4	B*15:18:01	B*44:03:01
Exon 5	B*15:18:01	B*44:03:01
Exon 6	B*15:18:01	B*44:03:01
Exon 7	B*15:18:01	B*44:03:01

Command line: *username$ cat -out_HLA-C_exon1-6_common.txt -out_HLA-C_exon7.txt | sort | uniq -d | grep '*' > -out_HLA-C_allele.txt*

If a new polymorphism is included in the exon list, we can detect its presence at the Blat search stage as shown in Table 5.

3.6.2 Mapping of Reads for the Extracted Allele Candidate Sequence

1. Boot Reference Mapper software on Linux OS.
2. Select a new project window and select the Reference sequence (extracted allele candidate sequence) and IonPGM-out.fastq (reads).
3. Set the mapping parameters in the parameters tab as follows: Minimum read length: 45, Minimum overlap length: 200, Minimum overlap identity: 100, Alignment identity score: 10, and Alignment difference score: 0.
4. Click the alignment start button.
5. After analyzing the reference sequence, open the Profile window and confirm Percent Coverage of Reference. If the coverage shows "100%," the allele candidate is considered to be correct.
6. Open the Alignment results window and visually inspect the sequence alignment for nucleotide mismatches. If the sequence reads represented by the consensus sequence are uniformly mapped to the reference sequence, as shown in Fig. 5a, we can assign the allele as correct. If a new polymorphism that has not been previously reported is included in the sequence reads, its presence will be detected during the calculation of the coverage and the final confirmation stages (Fig. 5b).

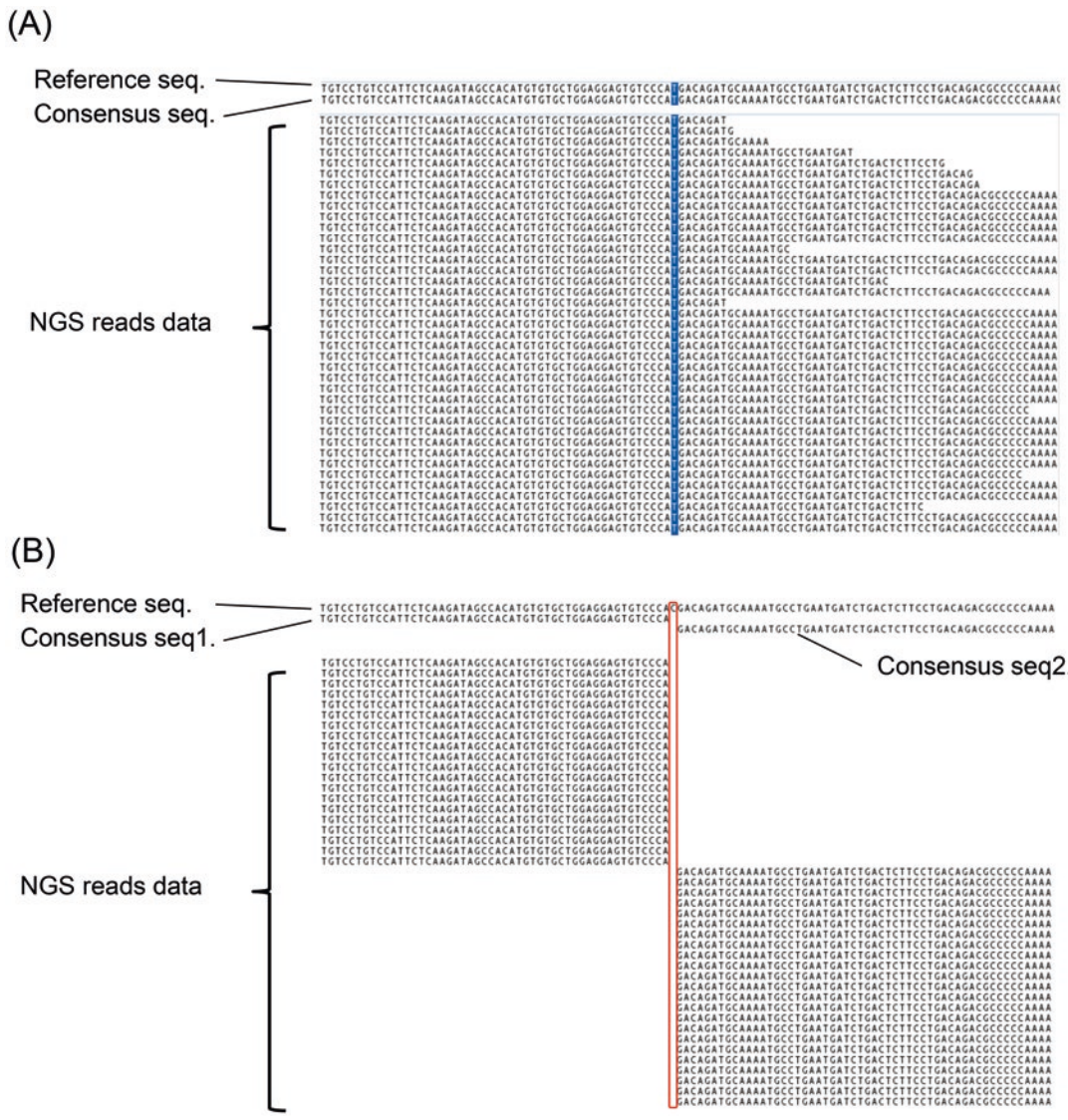


Fig. 5 Detection of a new allele during the calculation of the coverage and final confirmation stages in SeaBass. Two different examples of the mapping results of sequence reads using the GS Reference Mapper. (**a**) In this example, there is no mismatch between the reference and consensus sequences. (**b**) In this example, there is a mismatch between the reference and read sequences (the nucleotide C is in the reference but not the consensus sequences 1 and 2 as indicated by vertical red rectangle)

7. If a new polymorphism is detected as a mismatch (Fig. 5b), then the sequence should be confirmed by traditional methods such as Sanger sequencing and sub-cloning.

3.6.3 Evaluation of the SeaBass Program

To evaluate the SeaBass program, we used a total of 2414 HLA sequences from all the classical HLA loci that have frequent HLA alleles in Caucasians, African-Europeans and Japanese, and obtained an overall accuracy rate of >99.8%, and 100% for Japanese subjects (Table 6). The accuracy rate was not 100% for

Table 6
Evaluation of the SeaBass program

	Total	A	C	B	DRB345	DRB1	DQA1	DQB1	DPA1	DPB1
World-wide subject (1916 loci)										
Locus number	1916	250	250	242	186	239	140	234	140	235
Allele number	3832	500	500	484	372	478	280	468	280	470
Accuracy rate (%)	99.8	**100**	**100**	**100**	99.2	99.6	**100**	**100**	**100**	99.6
Japanese subject (498 loci)										
Locus number	498	86	80	77	50	68	4	65	4	64
Allele number	996	172	160	154	100	136	8	130	8	128
Accuracy rate (%)	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

HLA-DRB1/3/4/5 and HLA-DPB1 of the non-Japanese subjects because the complete coding sequences have not been determined as yet for some of their HLA-DRB and HLA-DPB1 alleles. Nevertheless, the allele assignment method that we developed for SeaBass appears to be the most accurate and efficient way to detect new and null alleles by NGS.

4 Notes

1. Although we evaluated 20 commercially available DNA polymerases that could be used for long ranged PCR, the PrimeSTAR GXL DNA polymerase showed the best performance for each of our PCR primer pairs.
2. In our experience 4–6 μg genomic DNA are obtained in 200 μL dissolved water (20–30 ng/μL) with an *A*260/*A*280 ratio of 1.7–1.9 when using healthy subjects.
3. Confirm that the bulk of the extracted DNA is >20 kb. If a large smear or a wide distribution of low molecular DNA fragments <15 kb is observed, we recommend the re-extraction of the DNA.
4. PCR time is approximately 3–4 h.
5. Confirm the molecular length of the PCR product, comparing with Fig. 1 of the previously reported article [10]. The DNA fragment sizes should be between 4.6 and 11.2 kb for optimal library preparation.
6. The PCR lengths of DRB1-1 and DRB1-2 are significantly different among DR-types (6.1–11.2 kb in DRB1-1 and 5.1–6.2 kb in DRB1-2). However, we calculate the length of DRB1-1 as 11.2 kb and DRB1-2 as 6.2 kb.
7. Expected concentrations should be 60–150 pg/μL. Make sure that sample concentrations are within 5–500 pg/μL sensitivity range for the DNA Kit. For accurate concentration estimations the peak size should be within 50–100 FU.

8. For example, the library concentration is 15,000 pM. Dilution factor = 15,000 pM/100 pM = 150. Thus, 1 μL of library pool mixed with 149 μL of low TE (1:150 dilution) yields approximately 100 pM. Use this library dilution for template preparation. Diluted libraries are stored at 2–8 °C and should be used within 48 h. Store undiluted libraries at −30 °C to −10 °C for long-term storage.
9. For example, the total PCR size shown in Fig. 1 for the eight gene loci is 72 kb. Required throughput (Mb) = 72 × 300 × 2 /1000 = 43.2. In contrast, the minimum throughput of Ion 314 Chip, Ion 316 Chip and Ion 318 Chip is 60 Mb, 0.6 Gb and 1.2 Gb, respectively (Table 1). Thus, the suitable library numbers are set for one library for Ion 314 Chip, 13 libraries for Ion 316 Chip and 27 libraries for Ion 318 Chip.
10. Please refer to an appropriate guidebook for how to use the command lines or ask a bioinformatician.

References

1. Santamaria P, Lindstrom AL, Boyce-Jacino MT, Myster SH, Barbosa JJ, Faras AJ, Rich SS (1993) HLA class I sequence-based typing. Hum Immunol 37(1):39–50
2. Saiki RK, Walsh PS, Levenson CH, Erlich HA (1989) Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. Proc Natl Acad Sci U S A 86(16):6230–6234
3. Adams SD, Barracchini KC, Chen D, Robbins F, Wang L, Larsen P, Luhm R, Stroncek DF (2004) Ambiguous allele combinations in HLA Class I and Class II sequence-based typing: when precise nucleotide sequencing leads to imprecise allele identification. J Transl Med 2(1):30. https://doi.org/10.1186/1479-5876-2-30
4. Wiseman RW, Karl JA, Bimber BN, O'Leary CE, Lank SM, Tuscher JJ, Detmer AM, Bouffard P, Levenkova N, Turcotte CL, Szekeres E Jr, Wright C, Harkins T, O'Connor DH (2009) Major histocompatibility complex genotyping with massively parallel pyrosequencing. Nat Med 15(11):1322–1326. https://doi.org/10.1038/nm.2038
5. Lind C, Ferriola D, Monos D (2010) Next generation sequencing: entering a new era in HLA sequence-based typing. ASHI 34(3):8–14
6. Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, Slavich L, Walker R, Hsiao T, McLaughlin L, D'Arcy M, Gai X, Goodridge D, Sayer D, Monos D (2010) Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum Immunol 71(10):1033–1042. https://doi.org/10.1016/j.humimm.2010.06.016
7. Gabriel C, Danzer M, Hackl C, Kopal G, Hufnagl P, Hofer K, Polin H, Stabentheiner S, Proll J (2009) Rapid high-throughput human leukocyte antigen typing by massively parallel pyrosequencing for high-resolution allele identification. Hum Immunol 70(11):960–964. https://doi.org/10.1016/j.humimm.2009.08.009
8. Lank SM, Wiseman RW, Dudley DM, O'Connor DH (2010) A novel single cDNA amplicon pyrosequencing method for high-throughput, cost-effective sequence-based HLA class I genotyping. Hum Immunol 71(10):1011–1017. https://doi.org/10.1016/j.humimm.2010.07.012
9. Holcomb CL, Hoglund B, Anderson MW, Blake LA, Bohme I, Egholm M, Ferriola D, Gabriel C, Gelber SE, Goodridge D, Hawbecker S, Klein R, Ladner M, Lind C, Monos D, Pando MJ, Proll J, Sayer DC, Schmitz-Agheguian G, Simen BB, Thiele B, Trachtenberg EA, Tyan DB, Wassmuth R, White S, Erlich HA (2011) A multi-site study using high-resolution HLA genotyping by next generation sequencing. Tissue Antigens 77(3):206–217. https://doi.org/10.1111/j.1399-0039.2010.01606.x
10. Shiina T, Suzuki S, Ozaki Y, Taira H, Kikkawa E, Shigenari A, Oka A, Umemura T, Joshita S, Takahashi O, Hayashi Y, Paumen M, Katsuyama Y, Mitsunaga S, Ota M, Kulski JK, Inoko H

(2012) Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. Tissue Antigens 80(4):305–316. https://doi.org/10.1111/j.1399-0039.2012.01941.x

11. Ozaki Y, Suzuki S, Shigenari A, Okudaira Y, Kikkawa E, Oka A, Ota M, Mitsunaga S, Kulski JK, Inoko H, Shiina T (2014) HLA-DRB1, -DRB3, -DRB4 and -DRB5 genotyping at a super-high resolution level by long range PCR and high-throughput sequencing. Tissue Antigens 83(1):10–16. https://doi.org/10.1111/tan.12258
12. Kulski JK, Suzuki S, Ozaki Y, Mitsunaga S, Inoko H, Shiina T (2014) In phase HLA genotyping by next generation sequencing – a comparison between two massively parallel sequencing bench-top systems, the Roche GS Junior and Ion Torrent PGM. HLA and associated important diseases. Intech, Croatia
13. University of California SCU (2017) Blat software. http://hgdownload.cse.ucsc.edu/admin/exe/
14. Knaus BJ (2014) fastq2fasta.pl file format conversion program
15. IPD-IMGT/HLA (2017) HLA allele sequences. ftp://ftp.ebi.ac.uk/pub/databases/imgt/mhc/hla/
16. QIAGEN (2016) Protocol of QIAamp DNA Blood Mini Kit. https://www.qiagen.com/us/resources/resourcedetail?id=62a200d6-faf4-469b-b50f-2b59cf738962&lang=en
17. Coulter B (2016) Protocol of Agencourt AMPure XP. https://beckman.jp/files/products/genomics/AMPureXP/IFU_AMPureXP.pdf
18. Scientific TF (2008) Protocol of Quant-iT PicoGreen dsDNA Assay Kit. https://www.thermofisher.com/order/catalog/product/P11496?SID=srch-srp-P11496
19. Scientific TF (2016) Protocol of Ion Plus Fragment Library kit. https://tools.thermofisher.com/content/sfs/manuals/MAN0009847_IonXpressPlus_gDNA_FragLibraryPrep_UG.pdf#search=%27Ion+Xpress%E2%84%A2+Plus+gDNA+Fragment+Library+Preparation%27
20. Agilent (2013) Protocol of Agilent High Sensitivity DNA Kit. http://www.agilent.com/cs/library/usermanuals/Public/G2938-90321_SensitivityDNA_KG_EN.pdf#search=%27Agilent+High+Sensitivity+DNA+Kit%27
21. Scientific TF (2017) Protocol of Ion PGM Hi-Q OT2 Kit
22. Scientific TF (2017) Protocol of Ion PGM Hi Q Sequencing Kit. https://tools.thermofisher.com/content/sfs/manuals/MAN0009816_Ion_PGM_HiQ_Sequencing_Kit_UG.pdf#search=%27HiQ+sequencing%27
23. Shiina T, Suzuki S, Kulski JK (2016) MHC genotyping in human and nonhuman species by PCR-based next-generation sequencing. Next generation sequencing. Intech, Croatia

Chapter 9

High-Resolution Full-Length HLA Typing Method Using Third Generation (Pac-Bio SMRT) Sequencing Technology

Sheetal Ambardar and Malali Gowda

Abstract

The human HLA genes are among the most polymorphic genes in the human genome. Therefore, it is very difficult to find two unrelated individuals with identical HLA molecules. As a result, HLA Class I and Class II genes are routinely sequenced or serotyped for organ transplantation, autoimmune disease-association studies, drug hypersensitivity research, and other applications. However, these methods were able to give two or four digit data, which was not sufficient enough to understand the completeness of haplotypes of HLA genes. To overcome these limitations, we here described end-to-end workflow for sequencing of HLA class I and class II genes using third generation sequencing, SMRT technology. This method produces fully-phased, unambiguous, allele-level information on the PacBio System.

Key words Single molecule, Real time, Sequencing, HLA typing, PacBio, High resolution, Full length

1 Introduction

The Human Leukocyte Antigen (HLA) proteins play a pivotal role in the immune response and are implicated in numerous human pathological conditions including autoimmune disease, infectious diseases, cancer, and drug reactions [1–3]. Clinically, HLA gene sequence information is widely used in organ transplantation to identify matching donor and recipient HLA alleles. Highly similar alleles improve the organ transplant outcome and reduce the risk of rejection [4].

The HLA region is highly polymorphic region on the short arm of chromosome 6 [5]. HLA region comprises over 200 genes, but six HLA genes (class I—A, B, C and class II—DR, DP, DQ) are crucial for self and non-self-recognition. Class I proteins expresses on the surface of all nucleated cells in the human body, but class II proteins can be found on the antigen presenting phagocytes such as dendritic cells, mononuclear phagocytes [2, 3]. A complex pattern of polymorphism is observed in antigen presenting

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_9,

regions of HLA genes including exon 2 and 3 of class I and exon 2 of class II. Currently over 10,000 HLA class I and II alleles are available IMGT (the international ImMunoGeneTics) database (http://www.ebi.ac.uk/imgt/hla/).

Many HLA typing laboratories across the globe have adopted SSO (sequence specific oligonucleotides), SSP (sequence specific primers), and Sanger sequencing methods. However, SSO or SSP can only detect known alleles with high accuracy, while Sanger sequencing is unable to identify phased heterozygous SNPs and it is also expensive and laborious [6]. Various NGS platforms including 454, Illumina and Ion Torrent have been explored for HLA typing [6–8]. 454 sequencing was one of the first NGS platforms tested for HLA typing [9, 10]. As the HLA genes (A, B, C, DRB1 and DQB1) are more than 5 Kb, the 454 or other short read sequencing are not able to resolve haplotypes of donors.

To overcome the phasing limitations across distant SNPs especially those found in class II genes [11], third generation sequencing technology has been utilized where full-length HLA genes are amplified using long-range PCR and sequenced using Single Molecule Real-time (SMRT) PacBio sequencing [10]. PacBio sequencing of HLA genes provide high-resolution (6–8 digit) allelic information for multiple HLA genes (six genes) with phased SNPs [12].

2 Materials

1. Qiagen, QIAamp® DNA Mini and Blood Mini kit (Cat. No. 51304).
2. Agarose (Sigma; Cat no: A2576).
3. TAE Buffer, Tris–Acetate–EDTA, 1× Solution, Electrophoresis, Fisher BioReagents (Cat no: BP24344).
4. DNA Gel Loading Dye (6×) (ThermoFisher Scientific, Cat no: R061).
5. Biotium GELRED 10000× IN DMSO 0.5ML (Biotium, Cat no: NC9524151).
6. Qubit® Quantitation Platform—Fluorometer, Invitrogen PN Q32857).
7. Qubit dsDNA BR assay kit (ThermoFisher Scientific, Cat no: Q32850).
8. NGSgo GenDx HLA primers (GenDx, Cat. no: 2341102 & 2341502).
9. Long range polymerase (Qiagen, Cat. no: 206402)
10. Agencourt AMPure Beads (Beckman Coulter, Cat No: A63880)

11. Bioanalyzer® Instrument Agilent Technologies PN 2100.
12. Agilent DNA 12000 kit (Agilent, Cat no: 5067-1508).
13. SMRTbell™ Template Prep Kit 1.0 (Pacific Biosciences, Part no: 100-259-100).
14. AMPure PB Beads (Pacific Biosciences, Part no: 100-265-900).
15. DNA/Polymerase Binding Kit P6 v2 (Pacific Biosciences, Part no: 100-372-700).
16. MagBead Kit v2 (Pacific Biosciences, Part no: 100-676-500).
17. PacBio RS II SMRT Cells 8Pac v3 (Pacific Biosciences, Part no: 100-171-800).
18. DNA Sequencing Reagent Kit 4.0 v2 (Pacific Biosciences, Part no: 100-612-400).
19. DNA Sequencing Bundle 4.0 v2 (Pacific Biosciences, Part no: 100-676-400).
20. PacBio RS II SMRT Cell Oil (Pacific Biosciences, Part no: 100-209-300).
21. PacBio RS II DNA Internal Control Complex (Pacific Biosciences, Part no: 100-356-500).
22. PacBio Disposables like, tube septa, sample plate septa, sequencing plate septa mixing plate, and pipette tips (https://www.pacb.com/products-and-services/consumables/pacbio-rs-ii-consumables/disposables/).
23. It is advisable to collect the human blood sample in EDTA coated tubes for efficient extraction of the human DNA.

3 Methods

The workflow of the HLA typing by PacBio SMRT sequencing can be broadly divided into four major steps (Fig. 1):

1. HLA amplicon generation (*see* Subheading 3.1).
2. HLA SMRTbell library preparation (*see* Subheading 3.2).
3. PacBio SMRT sequencing (*see* Subheading 3.3).
4. Data analysis using bioinformatics tools (*see* Subheading 3.4).

3.1 HLA Amplicon Generation

HLA amplicon generation can be further divided into various steps.

3.1.1 DNA Extraction

Extract the genomic DNA 200 µl of whole blood using Qiagen, QIAamp® DNA Mini and Blood Mini kit as per manufacturer's protocol [13].

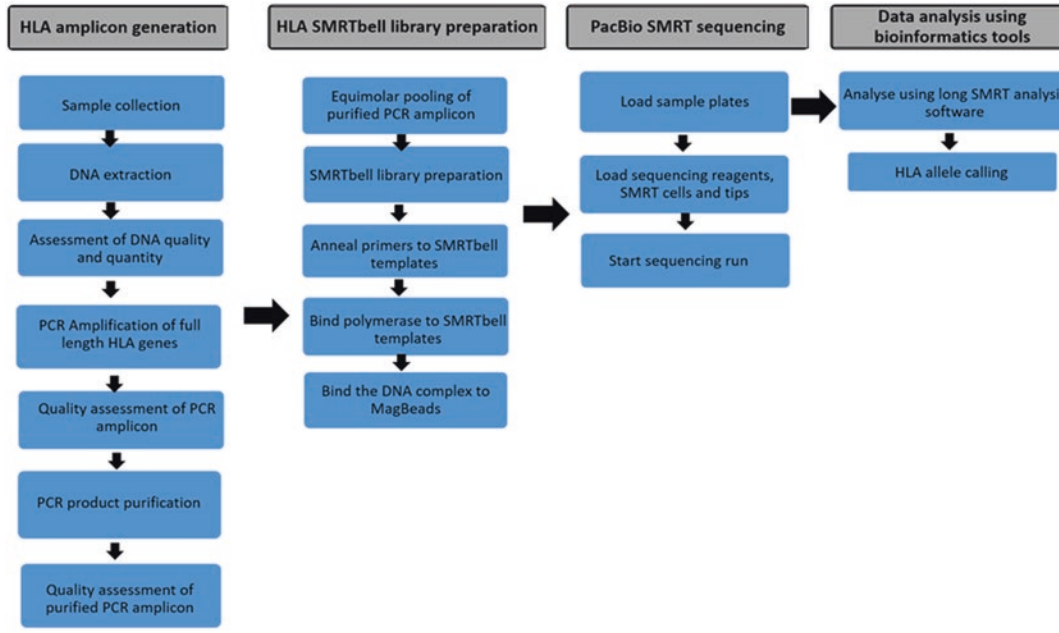


Fig. 1 Workflow of amplification of HLA gene and PacBio SMRT sequencing

1. Pipet 20 µl QIAGEN Protease (or proteinase K) into the bottom of a 1.5 ml micro centrifuge tube.
2. Add 200 µl blood sample followed by 200 µl Buffer AL to the microcentrifuge tube. Mix the sample by pulse-vortexing for 15 s (*see* **Note 1**).
3. Incubate at 56 °C for 10 min.
4. Spin down the contents and add 200 µl ethanol (96–100%) to the sample.
5. Mix again by pulse-vortexing for 15 s followed by spinning down to remove drops from the inside of the lid.
6. Keep the QIAamp Mini spin column in 2 ml collection tube and add the contents to it. Close the cap, and centrifuge at 6000 × *g* (8000 rpm) for 1 min.
7. Discard the collection tube containing the filtrate and place the QIAamp Mini spin column in a clean 2 ml collection tube.
8. Add 500 µl Buffer AW1 without wetting the rim and centrifuge at 6000 × *g* (8000 rpm) for 1 min.
9. Place the QIAamp Mini spin column in another clean 2 ml collection tube, and discard the collection tube containing the filtrate.
10. Add 500 µl Buffer AW2 to the column without wetting the rim. Close the cap and centrifuge at full speed (20,000 × *g*; 14,000 rpm) for 3 min.

11. Again place the QIAamp Mini spin column in a new 2 ml collection tube and discard the old collection tube with the filtrate.
12. It is recommended to centrifuge the column at full speed for 1 min. This step helps to eliminate the chance of possible Buffer AW2 carryover.
13. Place the QIAamp Mini spin column in a clean 1.5 ml microcentrifuge tube, and discard the collection tube containing the filtrate.
14. Add 200 μl Buffer AE or distilled water. Incubate at room temperature (15–25 °C) for 1 min, and then centrifuge at 6000 × *g* (8000 rpm) for 1 min (*see* **Note 2**).

3.1.2 Assessment of DNA Quality and Quantity

1. Analyze the quality of DNA on 1% w/v agarose gel using 3 μl of extracted DNA.
2. Quantify the DNA using 1 μl of DNA and analyzing by DNA BR assay kit as per manufacturer's protocol on Qubit 3.0.

3.1.3 PCR Amplification of Full-Length HLA Genes

Full-length targeted class I HLA-A, B, C, and Exon 2–4 of HLA-DRB1 and DQB1 genes was amplified using GenDx HLA primers as per the manufacturer's protocol [14] and high-fidelity long-range polymerase (Fig. 2).

1. All the reaction must be set on ice.

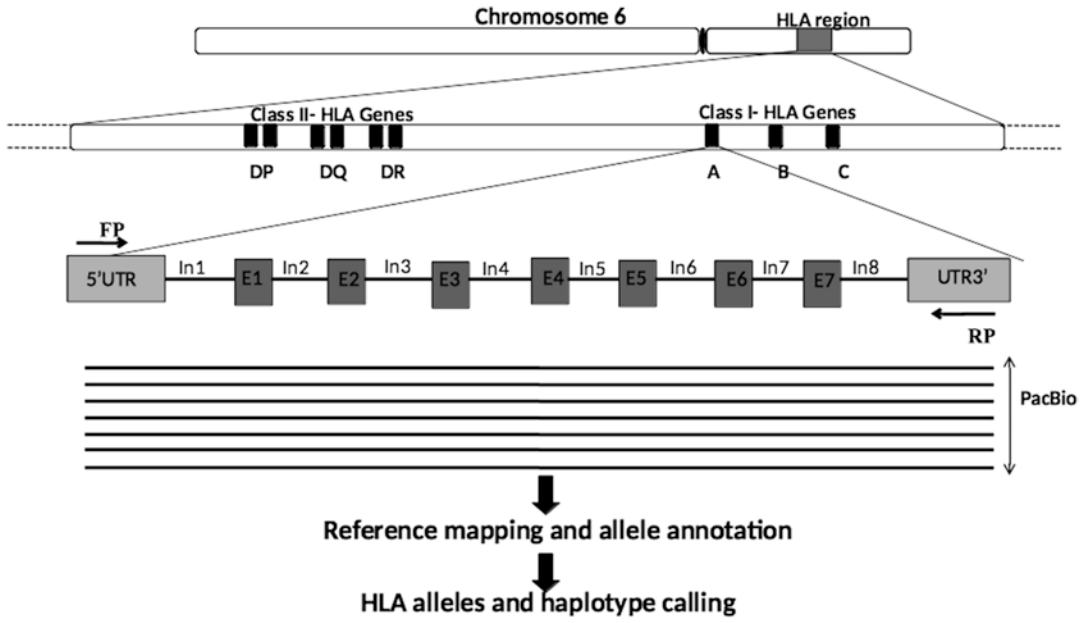


Fig. 2 HLA typing strategy for class I and II genes using PacBio methods. FP and RP represent the forward and reverse primers, respectively, for PCR amplification of HLA genes and E1–E7 represents Exon 1 to Exon 7 and In1–In8 represents Intron 1–8, UTR untranslated region

Table 1
Composition of reaction mix for locus-specific amplification

Component	All loci (except DQB1), µl	DQB1
Nuclease-free H_2O	15.85–18.85	10.45-13.45 µl
Q-Solution (5×)	–	5 µl
LongRange enzyme (5 U/µl)	0.4	0.8 µl
Long Range buffer (10×)	2.5	
dNTP mix (10 mM each)	1.25	
AmpX primer (red cap)	1	
Sample DNA (50 ng)	1–4	
Total Volume	25	

2. Thaw 10× Long Range PCR Buffer, dNTP mix, nuclease-free H_2O, and primer solutions. Mix the solutions thoroughly and centrifuge briefly before use.
3. Prepare a reaction mix as shown in Table 1. Prepare a separate reaction mix for each amplification primer.
4. HLA-DQB1 requires the addition of Q-Solution and double the amount of Long-Range enzyme per reaction (*see* **Note 3**).
5. Vortex the reaction mix thoroughly, and centrifuge briefly.
6. Add the reaction mix into each PCR tube. The appropriate volume is 25 µl minus the amount of DNA added in the next step.
7. Add 1–4 µl template DNA (50-200 ng) to each tube containing reaction mix.
8. Program the thermal cycler according to the manufacturer's instructions, using the conditions outlined in Table 2.
9. After amplification, store the samples overnight at 2–8 °C (*see* **Note 4**).

3.1.4 Quality Assessment of PCR Amplicon

1. Quantify the amplicons of class I and II genes using 1 µl of PCR assay and analyzing by DNA BR assay kit as per manufacturer's protocol on Qubit 3.0.
2. Analyze the PCR products on 1% w/v agarose gel using 3 µl of each PCR assay.
3. Confirm the size of amplicon as per Table 3.

3.1.5 PCR Product Purification

1. All amplicons of class I and II genes were purified using Agencourt AMPure beads.
2. Keep the AMPure beads at room temperature for half an hour.

Table 2
Cycling protocol for amplification

Step	Temperature, °C	Time	Number of cycles
Initial denaturation	95	3 min	1
Denaturation	95	15 s	35
Annealing	65	30 s	
Elongation	68	6 min	
Final elongation	68	10 min	1
Cooling	4	∞	

Table 3
Approximate size of PCR products

HLA locus	Expected size, Kb
A, B and C	3.1–3.4
DRB1	3.7–4.8
DRB3	3.8
DRB4	0.4 (exon 2) and 1.3 (exon 3)
DRB5	4.0
DQA1	5.4–5.8
DQB1	3.7–4.1

3. Add 0.6× of AMPure beads (e.g., for 100 µl add 60 µl of AMPure beads) to the PCR amplicon and mix by pipetting up and down (*see* **Note 5**).
4. Spin down the tube to collect the beads and keep it at room temperature for 15 min.
5. Keep the tube in a magnetic bead rack until the beads collect to the side of the tube and the solution appears clear.
6. Keep the tubes on the magnetic bead rack, remove the cleared supernatant to another tube by pipetting off (*see* **Note 6**).
7. Do not remove the tube from the magnetic bead rack and add freshly prepared 70% ethanol to the Eppendorf tube to the opposite side of beads.
8. Use a sufficient volume of 70% ethanol to fill the tube without disturbing the beads. Let the tube sit for 30 s.

9. After 30 s, pipette and discard the 70% ethanol.
10. Repeat **steps 6–8** above.
11. Remove residual 70% ethanol by short spin and place back on magnetic stand.
12. Pipette off any remaining 70% ethanol (*see* **Note 7**).
13. Allow beads to air-dry (with the tube caps open) for 30–60 s (*see* **Note 8**).
14. Remove the tube from the magnetic rack and add 50 μl of elution Buffer to the sample. Elute the DNA off the beads.
15. Vortex for 30 s and incubate at room temperature for 10 min.
16. Put the tube back on the magnetic rack to elute the sample from the beads and transfer the sample to a fresh tube. The bead tube can be discarded.

3.1.6 Quality Assessment of Purified PCR Amplicon

Quality and quantity of purified PCR amplicon using Caliper LabChip GX/Agilent 2100 Bioanalyzer and Qubit, respectively.

1. Quantify the purified amplicons of class I and II genes using 1 μl of DNA and analyzing by DNA BR assay kit as per manufacturer's protocol on Qubit 3.0.
2. Check the size of amplicon on Caliper LabChip GX (Perkin Elmer, USA) or Agilent 2100 Bioanalyzer using 12,000 DNA chip and reagents as per the manufacturer's protocol.
3. Confirm the size of amplicons as per the expected size given in Table 3.

3.2 HLA SMRTbell Library Preparation

1. Prepare the SMRTbell library using SMRT template preparation kit as per the manufacturer's protocol [15]. SMRT template preparation protocol consists of the following steps.
2. Equimolar pooling of purified PCR amplicon (*see* Subheading 3.2.1).
3. Repair DNA damage (*see* Subheading 3.2.2).
4. Purify the repaired DNA (*see* Subheading 3.2.3).
5. Blunt-end ligation of SMRTbell™ adapters to the end-repaired amplicons (*see* Subheading 3.2.4).
6. Add exonuclease and incubate (*see* Subheading 3.2.5).
7. Purify SMRTbell™ templates (*see* Subheading 3.2.6).
8. SMRTbell™ Library Quality Assessment (*see* Subheading 3.2.7).
9. Anneal primers to SMRTbell templates (*see* Subheading 3.2.8).
10. Bind polymerase to SMRTbell templates (*see* Subheading 3.2.9).

3.2.1 Equimolar Pooling of Purified PCR Amplicon

The PacBio SMRT library was prepared by pooling PCR amplicons of all HLA genes with equimolar concentration for each sample.

1. Convert the concentration of each amplicon into molar concentration using the below formula:

$$\text{Molar concentration} = \frac{\left(\text{Concentration of amplicon} \times 1000000\right) / 650\text{Da}}{\text{size in bases}}.$$

2. Pool equal molar of all the amplicons of each sample to make one sample. Pool equal molar concentration of HLA-A, HLA-B, HLA-C, HLA-DQB1 and HLA-DRB1 as one sample.
3. Keep target volume of final pool (μl) as 20 μl and target concentration per amplicon (nM) for equimolar pool as 4 nM.

3.2.2 Repair DNA Damage in Amplicons

1. Thaw the kit component on ice and add the reagents as per the Table 4 in a LoBind microfuge tube (recommended).
2. Mix the content by pipetting and spin down tube contents.
3. Incubate at 37 °C for 20 min or longer, then return reaction to 4 °C for 1 min.

3.2.3 Purify the Repaired DNA

1. Bring the bead reagent to room temperature and mix it well until the solution appears homogenous.
2. Add 0.6× volume of AMPure PB magnetic beads to the End-Repair reaction and mix by pipetting up and down.

Table 4
DNA repair reagent mix

Reagents	Stock concentration	Volume	Final
DNA amplicon pool	–	*X* μl for 1.0—5 μg	–
DNA damage repair buffer	10×	5.0 μl	1×
NAD^+	100×	0.5 μl	1×
ATP Hi	10 mM	5.0 μl	1 mM
dNTP	10 mM	0.5 μl	0.1 mM
DNA damage repair mix		2.0 μl	–
H_2O		*X* μl –to raise the final volume to 50 μl	–

3. Spin (or pulse) down the tube to collect the beads and transfer the tube to a VWR vortex mixer (recommended).
4. Allow the DNA to bind to beads by shaking at 2000 rpm for 10 min at room temperature until the bead/DNA mixture should appear homogenous.
5. Spin down the tube for a few seconds to collect beads.
6. Place the tube in a magnetic bead rack until the beads collect to the side of the tube and the solution appears clear (*see* **Note 9**).
7. Keeping the tubes on the magnetic bead rack, remove the cleared supernatant to another tube by pipetting off (*see* **Note 10**).
8. Do not remove the tube from the magnetic bead rack and add freshly prepared 70% ethanol to the Eppendorf tube to the opposite side of beads.
9. Use a sufficient volume of 70% ethanol to fill the tube without disturbing the beads. Let the tube sit for 30 s.
10. After 30 s, pipette and discard the 70% ethanol.
11. Repeat **steps 6–8** above.
12. Remove residual 70% ethanol by short spin and place back on magnetic stand.
13. Pipette off any remaining 70% ethanol (*see* **Note 7**).
14. Allow beads to air-dry (with the tube caps open) for 30–60 s (*see* **Note 8**).
15. Remove the tube from the magnetic rack and add 34 μl of Elution Buffer to the sample. Elute the DNA off the beads.
16. Vortex for 10 min at 2000 rpm.
17. Put the tube back on the magnetic rack to elute the sample from the beads and transfer the sample to a fresh tube. The bead tube can be discarded.
18. The End-Repaired DNA can be stored overnight at 4 °C or at −20 °C for longer duration (*see* **Note 11**).

3.2.4 Blunt-End Ligation of SMRTbell™ Adapters to the End-Repaired Amplicons

During this step, blunt end hairpin adapters (SMRTbell adapters) are ligated to repaired end of fragmented DNA.

1. Thaw the reagents on ice and mix them as per Table 5 in a LoBind microcentrifuge tube.
2. Add the adapter to the DNA whereas all other components should be added to the Master Mix (*see* **Note 12**). Make up the total volume to 40 μl with water.
3. Mix the reaction well by pipetting and spin down the tube contents.

4. Incubate at 25 °C for 15 min (*see* **Note 13**).
5. Incubate at 65 °C for 10 min to inactivate the ligase and store at 4 °C.
6. Caution: Don't stop and proceed with adding exonuclease after this step.

3.2.5 Add Exonuclease and Incubate

Add exonuclease to remove the un-ligated products.

1. Thaw the reagent on ice and mix the reaction well by pipetting.
2. Spin down the tube contents with a quick spin in a microfuge.
3. Add the reagents as per Table 6, mix by pipetting and spin down.
4. Incubate at 37 °C for 1 h, then return the reaction to 4 °C.
5. Caution: You must proceed with purification after this step.

3.2.6 Purify SMRTbell™ Templates

In this purification process, there are three distinct and consecutive AMPure PB bead purification steps. Refer *to Pacific Biosciences HLA Getting Started Guide*, page no. 25 for detailed steps. Perform all purification steps at room temperature to adequately remove enzymes (exonucleases, ligases, etc.) and ligation products smaller than 0.4 Kb, such as adapter dimers.

3.2.7 SMRTbell™ Library Quality Assessment

The libraries were validated by Agilent Bio-analyzer using 12,000 DNA Chip and quantified using Qubit 3.0.

1. Quantify the libraries by DNA BR assay kit using 1 μl of DNA as per the manufacturer's protocol on Qubit 3.0.
2. Check the size of library using Agilent 2100 Bioanalyzer using 12,000 DNA chip and reagents as per manufacturer's protocol (*see* **Note 14**).

3.2.8 Anneal Primers to SMRTbell Templates

Before starting the sequencing, the SMRTbell template must be annealed to sequencing primers at both ends of the SMRTbell template. This step is followed by binding of DNA polymerase to annealed template.

A Binding Calculator is provided to assist with setting up the annealing and binding reactions and setting up the sample plate for sequencing as per the manufacturer's protocol (Pacific Bioscience Template preparation and sequencing guide). Calculator can be downloaded from the web at http://calc.pacb.com or http://calc.PacificBiosciences.com.

1. In the calculator, click on number of SMRT Cells option which specifies how many SMRT Cells to prepare, and the Calculator determines the amount of sample necessary.

Table 5
Blunt end ligation reaction mix

Reagents	Stock concentration	Volume	Final
DNA (end repaired)	–	29–30 µl	–
Blunt adapter (20 µM)	20 µM	1.0 µl	0.5 µM
Mix before proceeding			
Template Prep Buffer	10×	4.0 µl	1×
ATP Lo	1 mM	5.0 µl	0.05 mM
Mix before proceeding			
Ligase[a]	30 U/µl	1.0 µl	0.75 U/µl
H_2O		Add to make 40 µl	–

[a]The Ligase Enzyme tube should remain closed and on ice when not frozen

Table 6
Exonuclease reaction mix

Reagents	Stock concentration	Volume (µl)
Ligated DNA		40
Exo III	100 U/µl	1.0
Exo VII	10 U/µl	1.0
Total		42

2. Add the library concentration and size of libraries. Eg 20 ng/µl and 4000 bases.
3. Protocol: Select the loading method as MagBead.
4. Binding Kit: Select the sequencing polymerase as P6v2.
5. Preparation Protocol: Small scale.
6. Long Term Storage: No.
7. DNA Control Complex: Yes.
8. Complex Reuse: No.
9. Standard Concentration: Yes
10. Click on Custom Parameters section.
11. Concentration on Plate: Use the default recommendation as 0.01 nM.
12. DNA Control Complex Ratio to Template: Use the default recommendation as 1.2%.

Table 7
Reagents for conditioning primer

Reagents	Volume, µl
Sequencing Primer v2	1.0
Elution buffer	32.3
Total volume	33.3

Table 8
Reaction mix for annealing primer

Reagents	Volume
Volume H_2O	6.1 µl
10× primer buffer	0.9 µl
Sample volume	0.97 µl
Diluted sequencing primer	1.0 µl
Total volume	9.0 µl
Final concentration	0.8333 nM

13. Polymerase: Template Ratio: Use the default recommendation as 10:1.
14. Primer: Template Ratio: Use the default recommendation as 20:1.

    For the Conditioning primer:

    (a) Dilute and preheat the Sequencing Primer from 5000 to 150 nM in Elution Buffer as given in Table 7.

    (b) Incubate the diluted primers at 80 °C for 2 min then hold at 4 °C (*see* **Note 15**).

    For the annealing primer:

    (a) In a fresh tube, add the appropriate amount of reagents in the order given in Table 8.

    (b) Mix the content with pipetting, spin down briefly, and incubate at 20 °C for 30 min.

    (c) Transfer to 4 °C location for immediate use or store at −20 °C.

Table 9
Polymerase dilution reaction mix

Reagents	Volume
SA-DNA polymerase P6 (500 nM)	1.5 µl
Binding buffer v2	13.5 µl
Total volume	15.0 µl
Final concentration	50 nM

Table 10
Polymerase binding reaction mix

Reagents	Volume
No of SMRT cells	1
dNTP	1.5 µl
DTT	1.5 µl
Binding buffer v2	1.5 µl
Annealed template	9 µl
Diluted polymerase	1.5 µl
Total volume	15 µl
Final concentration	0.5 nM

3.2.9 Bind Polymerase to SMRTbell Templates

In the binding reaction step, DNA sequencing polymerases are bound to the primer-annealed SMRTbell templates.

1. Thaw the reagents for polymerase dilution and prepare the reaction on ice.
2. Dilute the polymerase in 0.5 ml tube (Table 9), mix by pipetting, and spin down briefly.
3. For polymerase binding, add the components as mentioned in Table 10, mix by pipetting and spin down briefly.
4. Incubate at 30 °C for 30 min and hold at 4 °C (*see* **Note 16**).

3.2.10 Bind the DNA Complex to MagBead

Prior to sequencing, the template-polymerase complex must be transferred to a 96-well sample plate with concentrations and volumes specified by the Binding Calculator. The bound complex for sequencing is prepared as mentioned below.

The DNA Internal Control Complex (Pacific Biosciences) are SMRTbell templates already bound with the polymerase and these are added to the sample before loading on the instrument. DNA

Table 11
First dilution of internal control complex

Reagents	Volume, μl
MagBead binding buffer	99
Stock DNA control	1
Total volume	100

Table 12
Second dilution of internal control complex

Reagents	Volume, μl
MagBead binding buffer	49
First dilution	1
Total volume	50

Table 13
Dilution of the sample complex

Sample name	HLA
# of SMRT cells	1
MagBead wash buffer	0 μl
MagBead binding buffer	18.6 μl
Sample complex	0.38 μl
Total volume	19.0 μl (at 0.01 nM)

control helps in determining any problems that may occur during binding and the sequencing run. The amount of DNA Internal Control Complex to add to experimental templates is determined by the sample insert size and chosen chemistry. The Binding Calculator automatically recommends the amount of DNA Internal Control Complex to add to achieve the total number of reads (between 500 and 1000 reads per SMRT Cell, *see* **Note 17**).

1. Dilute the DNA Internal Control Complex as shown in Table 11.
2. Further dilute the first dilution of DNA control as shown in Table 12.
3. Dilute the sample complex as given in Table 13 (*see* **Note 18**).
4. Keep the magnetic beads cold.

5. Add 73.9 µl MagBeads to empty tube and place it on magnetic stand.
6. Collect beads. Remove the supernatant and discard.
7. Add 73.9 µl of MagBead wash buffer and wash by slowly aspirating and dispensing ten times.
8. Collect beads. Remove the supernatant and discard.
9. Add 73.9 µl of MagBead Binding Buffer and mix by slowly aspirating and dispensing ten times.
10. Add 73.9 µl of washed beads to a new tube and place it on a magnetic stand.
11. Collect beads. Remove the supernatant and discard.
12. Add 19 µl of diluted sample complex and mix by slowly aspirating and dispensing ten times.
13. Incubate in a rotator at 4 °C for 20 min (up to 2 h).
14. Keep the tube containing MagBead complex on a magnetic plate.
15. Collect beads. Remove clear supernatant and discard (*see* **Note 19**).
16. Add 19 µl of MagBead binding buffer and mix by slowly aspirating and dispensing ten times.
17. Collect beads by placing the tube on magnetic stand. Remove the supernatant and discard.
18. Add 19 µl of MagBead wash buffer and wash by slowly aspirating and dispensing ten times.
19. Collect beads by placing the tube on magnetic stand. Remove the supernatant and discard.
20. Add 1.2 µl of DNA Control Dilution.
21. Add 17.8 µl of MagBead Binding Buffer and mix by slowly aspirating and dispensing ten times.
22. Keep at 4 °C until use.

3.3 Sequencing

1. *Loading the sample plate*: Transfer the specified volumes (19 µl) into separate well of a new 96-well PCR plate for processing on PacBio RS (*see* **Note 18**).
2. *Load the reagents, tips and SMRT cells on the sequencer and start sequencing*: The reagents and SMRT cell vary with type of sequencer used, therefore refer to [16] for details.

3.4 Data Analysis Using Bioinformatics Tools

PacBio raw sequencing reads is converted into consensus sequences using Long amplicon analysis software by PacBio (http://pacbiodevnet.com/). Further data can de-multiplexed based on adapter sequence followed by trimming of primer sequences. Allele calling can be done by comparing with nearest IMGT HLA alleles using commercially available HLA typing softwares from connexion and GenDx.

3.4.1 Long Amplicon Analysis Software

Long amplicon analysis protocol software by PacBio generates de-novo consensus sequences from pooled amplicon samples. Each amplicon can come from a diploid or polyploid organism; the software uses any differences between the alleles to split the consensus sequence into multiple haplotypes. It allows for accurate allelic phasing and variant calling in large genomic intervals. The software first splits the reads by barcode, then the reads for each barcode are processed independently.

The Long Amplicon Analysis software includes four main steps:

1. *Coarse clustering*: This step group reads from different amplicons into different clusters. The coarse clustering step is generally successful in separating HLA-A, B, and C genes of class I and DR, DP and DQ gene of class II into separate clusters.
2. *Phasing*: This step load the reads for each cluster into the Quiver consensus software (http://pacbiodevnet.com/) and find an initial consensus. Recursively split reads from different haplotypes or other PCR products based on high scoring mutations proposed by Quiver.
3. *Consensus*: Consensus generates a final consensus for each haplotype or PCR product using Quiver.
4. *Post-processing filters*: This step removes PCR artifacts. Chimeric sequences are identified, and other PCR artifacts are identified by overall consensus quality.

3.4.2 HLA Typing Softwares

HLA alleles can be typed using commercially available softwares like, NGSengine-GenDx (http://www.gendx.com/products/ngsgo-pb), and HLA typing software-connexion (http://www.conexio-genomics.com).

4 Notes

1. Do not add QIAGEN protease or proteinase K directly to Buffer AL.
2. Incubating the QIAamp Mini spin column loaded with Buffer AE or water for 5 min at room temperature before centrifugation generally increases DNA yield. A second elution step with a further 200 µl Buffer AE will increase yields by up to 15%.
3. It is extremely important to include at least one negative control in every PCR setup that lacks template nucleic acid to detect possible contamination.
4. Clean-up of the PCR products should be carried out within 24 h.

5. 0.6× AMPure beads is required to purify the DNA fragment of 1–5 Kb.
6. Avoid disturbing the bead pellet.
7. Evaporation of ethanol can be confirmed by turning the tube. If the bead pellet is dry, it will stay on the wall in the same spot even though it is not on the magnet.
8. Do not exceed 2 min that can lead to over drying of the beads.
9. The actual time required to collect the beads to the side depends on the volume of beads added.
10. Avoid disturbing the bead pellet.
11. The protocol can be stopped at this step and reaction mix can be stored for longer duration.
12. The adapter should be added to DNA and NOT to the ligase enzyme master mix.
13. For blunt end ligation, overnight incubation at 25 °C is recommended for input material less than 1 μg. At this point, the ligation can be extended up to 24 h to maximize ligation efficiency, or cooled to 4 °C for storage of up to 24 h.
14. Typical library yields will require at least a 1:10 dilution prior to analysis on the Bioanalyzer instrument to ensure reliable quantitation.
15. Conditioned primer may be stored at −20 °C and used for up to 30 days.
16. Once the polymerase-SMRTbell template complex is formed, it should either be immediately used or stored at 4 °C for up to 3 days. Yield may be impacted if stored longer than 7 days.
17. The calculations are considering following the Insert size—4000 bases, Library Concentration—20 ng and Number of SMRT cell used—1.
18. The calculations are for the preparation of one SMRT cell, the amount will vary depending upon the number of SMRT cells used.
19. Optional: you can save 5 μl of supernatant for QC purposes.

Acknowledgments

We are thankful to Dr. Sudhir Krishna, NCBS for his scientific guidance and funding this research. We thank Dr. Latha Jagannathan and Dr. Nutan Dighe, Bangalore Medical Services Trust, Bangalore for their constant support and providing the samples for research work. We thank Mohammad Zahid from Shiva Scientific/GenDx for providing PacBio HLA primers. Thanks to

Dr. Anil Singh and Mr. Mohit from Institute of Himalayan Bioresource Technology, Palampur and also thanks to Paras Yadav, Imperial Life Science (P) LTD., Delhi for their help to access PacBio sequencer. We appreciate Centre for Cellular and Molecular Platforms and Department of Biotechnology, Government of India for their support for this project.

References

1. Carrington M, O'Brien SJ (2003) The influence of HLA genotype on AIDS. Annu Rev Med 54:535–551
2. Shina T, Hosomichi K, Inoko H, Kulski JK (2009) The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet 5:15–39
3. De Santis D, Dinauer D, Duke J, Erlich HA, Holcomb CL et al (2013) 16th IHIW: review of HLA typing by NGS. Int J Immunogenet 40:72–76
4. Morishima Y, Sasazuki T, Inoko H, Juji T, Akaza T, Yamamoto K et al (2002) Matched unrelated donors. Blood 99:4200–4206
5. Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J et al (2008) Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. Immunogenetics 60:1–18
6. Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, Slavich L et al (2010) Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum Immunol 71:1033–1042
7. Bentley G, Higuchi R, Hoglund B, Goodridge D, Sayer D, Trachtenberg EA, Erlich HA (2009) High-resolution, high-throughput HLA genotyping by next generation sequencing. Tissue Antigens 74:393–403
8. Enrich RL, Jia X, Anderson S, Banks E, Gao X et al (2011) Next generation sequencing for HLA typing of class I loci. BMC Genomics 12:42
9. Moonsamy PV, Williams T, Bonella P, Holcomb CL, Höglund BN et al (2013) High throughput HLA genotyping using 454 sequencing and the Fluidigm Access Array™ system for simplified amplicon library preparation. Tissue Antigens 81:141–114
10. Gowda M, Ambardar S, Dighe N, Manjunath A, Shankaralingu C, Hallappa P, Harting J, Ranade S, Jagannathan L, Krishna S (2016) Comparative analyses of low, medium and high-resolution HLA typing technologies. J Clin Cell Immunol 7(2). https://doi.org/10.4172/2155-9899.1000399
11. Nelson WC, Pyo CW, Vogan D, Wang R, Pyon YS et al (2015) An integrated genotyping approach for HLA and other complex genetic systems. Hum Immunol 76:928–938
12. Mayor NP, Robinson J, McWhinnie AJM, Ranade S, Eng K et al (2015) HLA typing for the next generation. PLoS One 10:e0127153
13. QIAGEN (2016) QIAamp® DNA mini and blood mini handbook (cat. no. 51304). https://www.qiagen.com/de/resources/resourcedetail?id=62a200d6-faf4-469b-b50f-2b59cf738962&lang=en. Accessed 10 Dec 2017
14. GenDX (2017) NGSgo-AmpX instruction for use (cat. no. 2341102 & 2341502). http://www.gendx.com/ngsgo-ampx. Accessed 10 Dec 2017
15. Pacific Bioscience (2016) HLA getting started guide (cat. no. 100-259-100). http://www.pacb.com/documentation/hla-getting-started-guide/. Accessed 10 Dec 2017
16. Pacific Bioscience (2015) Template preparation and sequencing guide (cat. no. 100-259-100). http://www.pacb.com/documentation/guide-pacific-biosciences-template-preparation-and-sequencing/. Accessed 10 Dec 2017

Chapter 10

Full-Length HLA Class I Genotyping with the MinION Nanopore Sequencer

Kathrin Lang, Vineeth Surendranath, Philipp Quenzel, Gerhard Schöfl, Alexander H. Schmidt, and Vinzenz Lange

Abstract

Nanopore sequencing, a paradigm change in sequencing technologies, offers a new cost-effective and scalable platform for HLA genotyping. Among the new generation of high-throughput sequencing technologies, the MinION nanopore sequencer is the first to offer a non-template-based direct DNA sensing sequencing technology. Oxford Nanopore Technologies (ONT) introduced the first version of the MinION in 2014; since then, the platform has gone through multiple iterations resulting in higher throughput and sequencing accuracy. The "what you put in is what you get" nature of the platform enables molecules to be sequenced without fragmentation. This results in ultra-long read lengths in the order of tens of kilobases enabling entire genes to be characterized with fully phased sequence information. With release R9.5, the MinION platform has reached a quality that enables HLA genotyping with minor shortcomings in long homopolymer regions. Within this chapter, we describe a protocol for sequencing and genotyping HLA Class I alleles using the MinION.

Key words HLA, Full-length, Genotyping, Nanopore sequencing, MinION

1 Introduction

With dimensions of 10 × 3 × 2 cm and 90 g weight, the MinION [1] is a portable sequencer that plugs directly into a standard USB3 port on a desktop computer that runs the MinKNOW software which is required to run and control the MinION sequencer (available for download from the ONT community portal). The MinION sequences unfragmented molecules in the order of tens of kilobases by the translocation of a DNA molecule through an orifice (referred to as pore) under the influence of a uniform electric field across a lipid membrane. This translocation causes characteristic changes in electric current density across the membrane, which

Electronic supplementary material: The online version of this article (https://doi.org/10.1007/978-1-4939-8546-3_10) contains supplementary material, which is available to authorized users.

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_10,

are then translated into nucleotide sequences. Since its introduction, the MinION platform has undergone rapid development with three changes of the pore, six changes of the chemistry involved in sequencing, and multiple software iterations. These changes have led to significant improvements in sequencing throughput, mainly by increasing translocation speed from 30 to 450 nucleotides per second, and basecalling accuracy, reducing single read error rate from 30% to 10% [2, 3].

Nanopore sequencing technology promises a cost-effective, practical, and efficient solution to characterize and genotype HLA alleles in their entirety, thus voiding phasing issues. Despite the relatively high per-read error rate, accurate mapping, consensus definition, and subsequent accurate genotyping are possible with a few hundreds of reads [1, 4, 5]. This also allows for multiplexing tens of samples within a single run with only approximately a hundred reads (in the order of 5 kb length) necessary to accurately genotype HLA class I loci [4, 5]. HLA genotyping using the MinION currently suffers from the inability to distinguish a limited number of specific alleles owing to the platform's failure in accurately characterizing long homopolymer stretches [6]. Excluding these regions from analysis while genotyping provides a temporary workaround; given the pace at which the MinION platform has evolved, a timely resolution of this issue is to be expected.

Within this chapter, the steps, both in the laboratory and in silico, required to characterize HLA Class I genotypes from genomic DNA are elucidated. Starting from genomic DNA, this protocol involves the use of primers targeting the three HLA Class I loci [7]. Molecular barcodes are attached in a second PCR to enable pooling of up to 96 PCR products. After a couple of sequencing library preparation steps, the PCR products are loaded on the MinION device [8]. The output files in the FAST5 format are first converted to FASTQ files, then demultiplexed and finally genotyped with commercially available genotyping software [9].

2 Materials

2.1 PCR

1. Genomic DNA.
2. HLA-specific primers extended by the following overlap sequences for MID-PCR:
   Forward primer: ACTTCGTACGTACGGCGTCTTATAC<Target specific Forward Primer>.
   Reverse primer: GAGACACGTCCGATTACGACTTGAC<Target specific Reverse Primer>.
   For HLA Class I target-specific primer sequences, *see* ref. 7.
3. MID primers.
   Forward primer: GGTGCTG**[MID]**TTAACCTACTTCGTACGTACGGCGTCTTATAC.

Reverse primer: AGGTTAA[**MID**]CAGCACCGAGACAC GTCCGATTACGACTTGAC.

For **[MID]** sequences *see* Supplementary Document 1.

4. PCR Kit of choice.
5. Thermocycler.
6. Gel-electrophoresis system.

2.2 Purification

1. SPRIselect beads.
2. Magnetic rack.
3. 70% ethanol.

2.3 Library Preparation

1. DNA low-bind tubes.
2. NEBNext End Repair/dA-tailing Module Kit [NEB].
3. NEB Blunt/TA Ligase Mastermix [NEB].
4. $1D^2$ sequencing Kit for genomic DNA [Oxford Nanopore; SQK-LSK 308].
5. Library Loading Bead Kit [Oxford Nanopore; EXP-LLB001].

2.4 Sequencing, Basecalling, and Demultiplexing

1. MinION MK1B.
2. R9.5 flow cell (FLO-MIN107).
3. Linux PC with at least 12 CPU cores and a minimum of 64 GB RAM with the software listed below, installed on it.
4. MinKNOW software (downloadable from https://community.nanoporetech.com).
5. Albacore software (downloadable from https://community.nanoporetech.com).
6. LAST aligner (downloadable from http://last.cbrc.jp/).
7. Biopieces framework (downloadable from http://maasha.github.io/biopieces/).
8. Demultiplexing scripts (downloadable from https://github.com/DKMS-LSL/ont_lastlopper).

2.5 Genotyping

1. Windows Desktop PC with a minimum of 8 GB RAM.
2. NGSengine (GenDx)/HLA specific software installed on the Windows Desktop PC.

3 Methods

Carry out all steps at room temperature. Pipette enzymes and ONT reagents on ice or cooling block. From DNA end repair onward, for all library preparation steps, DNA low-bind tubes are recommended. Avoid fast pipetting through thin pipette tips to

prevent shearing of the long amplicons. In particular, do not mix by pipetting. Instead, mix by inversion or flicking the tube with a finger.

3.1 PCR and Pre-Library Steps

1. Perform individual long-range PCRs with an appropriate long-range PCR kit with 50–200 ng genomic DNA and primers of choice to amplify HLA class I genes [7]. Verify amplification success and specificity by gel-electrophoresis (*see* **Note 1**).
2. Perform MID-PCRs with an appropriate long-range PCR kit using distinct barcodes/MIDs (*see* Subheading 2.1 and Supplementary Document 1) for each reaction.
3. Pool up to 96 MID-PCR products. Optionally, PCR product quantity may be equalized by adjusting the volume of pooled PCR product based on the gel-electrophoresis band intensities.
4. Purify at least 400 µl of the amplicon pool with SPRIselect beads in a bead/DNA volume ratio 0.6×. Make sure the beads are suspended completely by vortexing. Mix by inversion and spin down only for a short time such that the beads remain dispersed. Incubate for 5–10 min. Separate beads on a magnetic rack. Discard the supernatant. Perform a wash step with 500 µl 70% ethanol leaving the tube at the magnetic rack. Discard the supernatant. Repeat the wash step. Spin down shortly to gather the ethanol at the bottom of the tube, separate the ethanol from beads using a magnetic rack, and remove completely (*see* **Note 2**). Air-dry the pellet for 2 min. Elute with 150 µl of water, mix by inversion, and spin down shortly. Incubate for 5 min. Separate beads using the magnetic rack. Transfer the supernatant to a fresh tube.
5. Measure the DNA concentration preferably using a fluorescent-based method.

3.2 Library Preparation

1. Prepare the DNA end-repair/dA-tailing reaction according to the manufacturer's guidelines in a total volume of 60 µl using 500 fmol (1.5 µg at 4.5 kb amplicon length) of the initial purified pool.
2. After heat-inactivation of the DNA end-repair/dA-tailing reaction, spin down shortly and transfer 100 µl to a 1.5 ml tube. Add 60 µl (0.6×) of SPRIselect beads for purification to the DNA end-repair/dA-tailing reaction. Purify as described in **step 4** of Subheading 3.1.
3. Quantify recovery by fluorescence. Ensure a recovery of 700 ng to 1 µg of DNA (*see* **Note 3**).
4. Mix 22.5 µl of the eluted amplicon pool with 2.5 µl 1D adapters (ONT kit) and 25 µl Blunt/TA Ligase Master Mix (*see* **Note 4**) and incubate for 10 min. Use water to reach a total

volume of 100 µl and then add 60 µl (0.6×) SPRIselect beads. Purify as described in **step 4** of Subheading 3.1.

5. Elute with 46 µl water and incubate for 5 min. Use the magnetic rack for separation and transfer the supernatant to a fresh tube.
6. Mix 45 µl from **step 5** with 5 µl BAM (ONT kit) and 50 µl Blunt/TA Ligase Master Mix, and incubate for 10 min.
7. Purify by adding 60 µl (0.6×) SPRIselect beads. Incubate for 5–10 min before using a magnetic rack for separation. Discard the supernatant. Perform a wash step with 140 µl of "Adapter Bead Binding" buffer (ONT kit). Close the tube and resuspend the beads by flicking the tube. Place the tube back on the magnetic rack. Discard the supernatant. Repeat the wash step with 140 µl of "Adapter Bead Binding" buffer.
8. Air-dry the pellet for 2 min and then elute with 15 µl elution buffer (ONT kit). Incubate for 5 min and transfer the supernatant to a fresh tube after using a magnetic rack for separation.
9. Quantify recovery by fluorescence. Ensure a recovery of more than 250 ng of DNA (*see* **Note 3**). The library is now ready for sequencing and should be kept on ice.

3.3 Flow Cell Quality Control, Priming, and Sequencing

1. Place the flow cell in a MinION MK1B and connect the MinION to the computer on which the MinKNOW GUI is installed.
2. Label your experiment in the MinKNOW GUI, e.g., Sample ID and flow cell ID, and press the "Submit" button. Select Platform Quality Control (QC) script under "Choose operation" and press the "Execute" button. When the check is complete, the software will return to the start page. To see the active pore report, click on the "notification panel." The number of active pores should be more than 800 (*see* **Note 5**).
3. Inspect the flow cell for air bubbles in the channel between priming port and array. To remove air bubbles, open the "Priming Port" and draw out buffer and bubbles; while removing buffer ensures that the sensor array is always covered by buffer (*see* **Note 6**).
4. Mix 480 µl Running Buffer (RBF) (ONT kit) with 520 µl water as priming solution. With a 1000 µl pipette, add 800 µl of the freshly prepared priming solution via the "Priming Port." Close the "Priming Port" and wait for 5 min. Open the "SpotON Sample Port" and add the remaining 200 µl via the "Priming Port." Take care to not introduce air bubbles while pipetting, preferably by not releasing the plunger (*see* **Notes 6** and **7**).

5. In a fresh tube mix 35 µl Running Buffer (RBF), 25.5 µl Loading Beads, 2.5 µl water, and 12 µl of the sequencing library (from Subheading 3.2). Dropwise, add 75 µl of this mix in the "SpotOn Sample Port." Replace the SpotON Sample Port cover and close the Priming Port. Close the MinION lid.
6. Select the appropriate protocol script under "Choose Operation" (*see* **Note 8**) and launch sequencing by pressing the "Execute" button. The GUI will keep you updated on sequencing progression, the time required for sequencing being dependent on the chosen sequencing protocol.

3.4 FASTQ Extraction and Demultiplexing

1. Determine the location of the FAST5 files output from the sequencer within /var./lib/MinKNOW/data/reads (*see* **Note 9**).
2. Create a directory to store the FASTQ files.
3. Run albacore with the following command:

```
read_fast5_basecaller.py --flowcell [flowcell used, in the form
FLO-MINxxx] --kit [sequencing kit used, in the form SQK-LSKxxx]
-t [number of cores] -o fastq -r -i [location of FAST5 files from
step 1] -s [directory from step 2]
```

(*see* **Note 10**).

4. Concatenate the resulting FASTQ files into 1 FASTQ file using the command:

```
cd [directory from step 2]; for fastq_file in `ls *. fastq` do
cat $fastq_file > [name of FASTQ file]; done
```

5. Download the demultiplexing scripts into a new directory and copy the newly created FASTQ file into the same directory.
6. Make the wrapper script executable with the command:

```
cd [location of the new directory from step 5.]; chmod +x de-
multiplex_vlast.sh
```

7. Copy the "Supplementary_2.fa" (Supplementary Document 2 in FASTA format) into the directory from **step 5**.
8. Run the demultiplexing routine with the command:

```
./demultiplex_vlast.sh [name of the FASTQ file with all the reads]
[name of the MID FASTA file]
```

3.5 Genotyping

1. Open NGSengine (GenDx).
2. Open a "New project" via the file-button in the left-hand corner at the top.
3. Define a name for your project and a folder where it should be saved, press "Next."

4. Add the folder where the demultiplexed fastq files are stored, press "Next."
5. A summary is shown, press "Next" to see a data overview for further analysis.
6. Open "Preferences" via the File tab in the left hand corner at the top:
   (a) Within the "General" section, set the maximum number of reads to 10,000 (*see* **Note 11**), set platform to "Oxford Nanopore" as the sequencing platform.
   (b) Within the "Locus default settings" section, set "Amplicon" to "Auto," and "Analysis region" to "Amplicon." Exclude problematic stretches, such as homopolymers, by excluding them in the "Ignore regions" field after setting it to "Custom." Use the IPD-IMGT/HLA genomic coordinate system to set the regions (*see* **Note 12**).
   (c) Within the "Locus Selection" section, define the loci that were in the sequencing run. Put them in the "Analysis" column.
7. Define the Loci by holding the mouse over the sample to be analyzed; via right mouse click, set the HLA locus/loci that should be typed.
8. Press "Analyze" for all samples at the left-hand site of the screen or for a specific sample at the right-hand side at the end of the sample line.
9. When the data analysis is complete, an overview of the genotyped alleles for the samples is shown.

4 Notes

1. High yield of the targeted product is important for successful data analysis. Smaller unspecific amplification products detectable in the gel-electrophoresis picture may indicate side products present at higher molarity than the targeted product and may therefore dominate the sequencing output.
2. Ethanol is known to interfere with downstream library preparation steps.
3. Lower recoveries indicate suboptimal handling and will result in lower sequencing output.
4. If the Blunt/TA Ligase Master Mix has partially precipitated, dissolve completely by vortexing or pipetting.
5. Flow cells with less than 800 active pores will result in lower sequencing yield and should therefore be returned to ONT for replacement.

6. Air contact will damage the pores.
7. At times the priming solution can bubble out of the "SpotON Port" but will be retracted by capillary action, thus spreading over the array; this should not raise any concerns.
8. MinKNOW will automatically suggest the appropriate sequencing protocol based on the already chosen flow cell type and ONT kit used for sequencing preparation.
9. Every run has two date and time stamped subfolders within the "reads" folder. The folder with the earlier time stamp is from an instrument control run and the folder with the later time stamp contains the FAST5 files pertaining to the samples sequenced.
10. Though the ONT kit used is a $1D^2$ (where both the sense and antisense strands of a molecule are sequenced, in contrast to 1D reads where only one of the strands is sequenced) kit, basecalling $1D^2$ reads for amplicon sequencing is prohibitively time consuming, at the time of writing, and should be avoided.
11. Choosing to use all the reads can be quite time intensive, setting this number to 10,000 enables faster genotyping without loss of resolution.
12. Homopolymers longer than 6 nucleotides are underestimated by the basecaller, at the time of writing, and can lead to erroneous genotyping; such regions can be excluded from further analysis by NGSengine reducing the resolution but preventing erroneous results.

References

1. Loman NJ, Quick J, Simpson JT (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12:733–735. https://doi.org/10.1038/nmeth.3444
2. Ip CLC, Loose M, Tyson JR et al (2015) MinION analysis and reference consortium: phase 1 data release and analysis. F1000Research 4:1075. https://doi.org/10.12688/f1000research.7201.1
3. Jain M, Tyson JR, Loose M et al (2017) MinION analysis and reference consortium: phase 2 data release and analysis of R9.0 chemistry. F1000Research 6:760. https://doi.org/10.12688/f1000research.11354.1
4. Ton KNT, Cree SL, Gronert-Sum SJ, et al (2017) Multiplexed nanopore sequencing of HLA-B locus in Māori and Polynesian samples. bioRxiv. doi:https://doi.org/10.1101/169078
5. Liu C, Xiao F, Hoisington-Lopez J, et al (2017) Accurate typing of class I human leukocyte antigen by Oxford nanopore sequencing. bioRxiv. doi:https://doi.org/10.1101/178590
6. Jain M, Koren S, Quick J, et al (2017) Nanopore sequencing and assembly of a human genome with ultra-long reads. bioRxiv. doi:https://doi.org/10.1101/128835
7. Ehrenberg PK, Geretz A, Sindhu RK et al (2017) High-throughput next-generation sequencing to genotype six classical HLA loci from 96 donors in a single MiSeq run. HLA 90:284. https://doi.org/10.1111/tan.13133
8. Jain M, Fiddes IT, Miga KH et al (2015) Improved data analysis for the MinION nanopore sequencer. Nat Methods 12:351–356. https://doi.org/10.1038/nmeth.3290
9. Duke JL, Lind C, Mackiewicz K et al (2016) Determining performance characteristics of an NGS-based HLA typing method for clinical applications. HLA 87:141–152. https://doi.org/10.1111/tan.12736

Chapter 11

Imputation-Based HLA Typing with SNPs in GWAS Studies

Xiuwen Zheng

Abstract

SNP-based imputation approaches for human leukocyte antigen (HLA) typing take advantage of the extended haplotype structure within the major histocompatibility complex (MHC) to predict classical HLA alleles using dense SNP genotypes, such as those available on chip panels of genome-wide association study (GWAS). These methods enable HLA analyses of classical alleles on existing SNP datasets genotyped in GWAS studies at no extra cost. Here, I describe the workflow of HIBAG, an imputation method with attribute bagging, for obtaining a sample's HLA class I and II genotypes of two-field resolution using SNP data. Two examples are provided to illustrate with a publicly available HLA and SNP dataset: genotype imputation with pre-fit classifiers in GWAS, and model training to build a new classifier.

Key words HLA, MHC, Imputation, SNP, GWAS, HIBAG

1 Introduction

The human leukocyte antigen (HLA) system, located in the major histocompatibility complex (MHC) on chromosome 6p21, is highly polymorphic and contains 226 genes with essential roles in the immune system. It has been shown to be important in human disease, adverse drug reactions, and organ transplantation [1], and more than 100 diseases and cancers are found to be associated with HLA variation [2]. HLA genes present a wide range of peptides to provide defense against a great diversity of environmental microbes. Evolutionary pressure has given rise to a great deal of functional diversity. Among these HLA loci, the class I genes (HLA-A, -B, and -C) and the class II genes (HLA-DRB1, -DQA1, -DQB1, and -DPB1) are the most frequently studied.

Due to the inherent highly polymorphic nature of the HLA region, high-resolution HLA typing remains challenging [3, 4]. Sequence-based typing (SBT) approaches have been considered the gold standard, however SBT approaches are relatively time-consuming and cost-prohibitive for large-scale studies compared to recent methods using next-generation sequencing (NGS). An

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_11,

alternative to SBT and NGS approaches is to infer or "impute" HLA types using SNP data in a genome-wide association study (GWAS). The typical SNPs used in HLA imputation are in the flanking region of a specific HLA gene, and the imputation methods rely on linkage disequilibrium between the presence of a SNP allele and specific HLA alleles. These methods require a training sample with reference information about how SNP variants are associated with HLA alleles. The HLA types of a new sample are determined by its SNP profile and a prediction model.

Although imputation-based methods are generally less accurate than direct sequencing approaches, one of the motivations using imputation is to leverage SNP datasets of millions of individuals genotyped in GWAS studies in the last decade, without additional cost involved [5]. A new method for HLA imputation using attribute bagging, HIBAG, was proposed recently, which is highly accurate and computationally tractable [6]. It can be used with published parameter estimates and eliminate the need to access large training samples. HIBAG combines the concepts of attribute bagging with haplotype inference from unphased SNPs and HLA types. Attribute bagging is a technique for improving the accuracy and stability of classifier ensembles using bootstrap aggregating and random subsets of variables [7, 8]. This method was independently validated using population-specific references and 1000 genomes HLA data in Japanese, Brazilian, and African American study samples [9–11]. The ability to accurately impute 2-field HLA genotypes in the Human Genome Diversity Project cell panel was further assessed in a recent collaborative study, suggesting the importance of large and diverse reference datasets [12].

The accuracy of HIBAG imputation depends on various factors, including whether the imputed population is evolutionary closely related to reference [5], the sample size of training data, the degree of polymorphism of HLA genes, the frequency of the HLA allele, the SNP density in the HLA region available on genotyping platforms, and genomic structure deletion resulting in systematic exclusion of SNPs from chip panels [12]. Our previous study found that the 2-field accuracies were between 92% and 99% for Europeans, but with a drop when inferring alleles for samples of Asian, Hispanic, and African ancestries. In general, a well-matched training set is necessary for high imputation accuracy, e.g., 90% or greater.

Our HIBAG method is freely available in the R/Bioconductor package (http://www.bioconductor.org/packages/HIBAG). The pre-fit classifiers have been published as an initial set of parameter estimates for the Illumina 1M Duo, OmniQuad, OmniExpress, 660K and 550K platforms, using the common SNP markers among these platforms [6]. A typical parameter file for imputing HLA types contains only haplotype frequencies at randomly selected SNP sites rather than individual training genotypes. Therefore,

HIBAG does not require the uploading of genotype information to a website for web-implemented methods, which could raise concerns over data privacy, or having access to large training HLA datasets. As shown in Fig. 1, the HIBAG workflow consists of training and prediction parts. The output of model fitting is a set of parameter estimates that are used in the prediction as an input. The pre-fit classifiers can be shared in research communities with no need of individual genotypes.

The model training with the HIBAG algorithm is a time-consuming process, while the computation time using pre-fit classifiers for prediction is much less. Here, I introduce a high-performance implementation using graphics processing unit (GPU), which is available in the HIBAG.gpu R package (https://github.com/zhengxwen/HIBAG.gpu). The GPU-based implementation enables up to 125-fold speedup with one GPU card for the process of model building, compared to the CPU calculation with one core. This makes training a very large number of samples computationally feasible. The HIBAG.gpu package is built based on the open computing language (OpenCL) framework, which is an open and royalty-free standard for parallel programming of heterogeneous systems (https://www.khronos.org/opencl).

As a complement to the initial published parameter estimates for Illumina 1M Duo, OmniQuad, OmniExpress, 660K and 550K chips, the HIBAG models have been built on more than 20 chip-based platforms, for example, Illumina ImmunoChip, Infinium OmniExpress, and Affymetrix Genome-Wide Human SNP Array 6.0. Platform-specific pre-built models tailor existing SNP markers to targeted chips, and maximize the prediction capability without using other SNP imputation tools.

Here, I describe how to use HIBAG to quickly and easily determine HLA class I and II genotypes using SNP data in GWAS studies. I also demonstrate how to build a HIBAG model with HLA and SNP genotypes, using multi-core acceleration and GPU computing.

2 Materials

To determine the intermediate-resolution (two-field) HLA types of a sample in GWAS, the HIBAG package and additional software tools must be installed on a computer. In addition, the pre-fit classifiers according to the GWAS genotyping platform need to be downloaded from the HIBAG website. The examples described in Chapter 3 can be executed on a publically available HLA dataset in the 1000 genomes project [12, 13] or the samples from your own experiments.

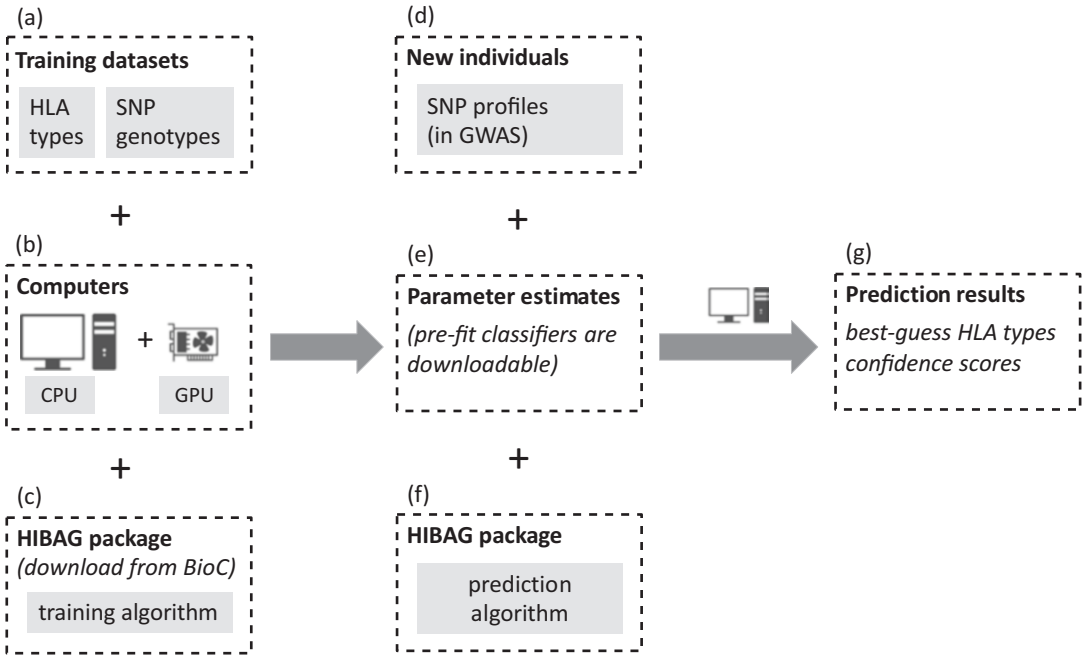


Fig. 1 Computational workflow for HLA imputation using SNP data. (**a**) the training datasets consist of HLA and SNP genotypes; (**b**) computing equipment can be a desktop, a cluster of compute node, or a graphics processing unit (GPU); (**c**) the HIBAG R package can be downloaded from Bioconductor (BioC) and the training algorithm is freely available in the package; (**d**) SNP profiles are required for imputing HLA types of new individuals; (**e**) the parameter estimates are the output of the training algorithm and the pre-fit classifiers are published online; (**f**) the prediction algorithm is freely available in the HIBAG package; (**g**) the prediction results are best-guess HLA types and their confidence scores

2.1 Hardware

The HIBAG method is publicly available as an R/Bioconductor package, and it can be installed and executed on multiple operating systems, e.g., Linux, MacOS and Windows. To utilize the GPU functionalities, an OpenCL-compatible hardware accelerator should be installed on your desktop or workstation, e.g., NVIDIA Tesla K80 for general-purpose computing on graphics processing units (GPGPU). Internet access is not necessary if the required datasets and pre-fit classifiers have been downloaded.

2.2 Software Dependencies

The R programming environment (http://www.r-project.org) should be installed and R v3.0 or a newer version is recommended. HIBAG v1.14.0 from Bioconductor or github is required to complete the exercises in this chapter (http://www.bioconductor.org/packages/HIBAG). The kernel of the HIBAG package is written in C++. To install the package from the source codes, a C++ compiler should be available on your operating system, e.g., GNU g++ compiler. To enable a GPU application, an appropriate driver should be installed including OpenCL headers and libraries, which

are included in the OpenCL SDK from your favorite vendor. The HIBAG.gpu package (v0.9.0) can be downloaded and installed from github (https://github.com/zhengxwen/HIBAG.gpu).

As an input of the HIBAG algorithm, SNP genotypes in GWAS are stored in either PLINK BED files [14] or SNP GDS files [15]. Optional software is PLINK (https://www.cog-genomics.org/plink2), a genome-wide association and population-based linkage analysis toolset. Here, PLINK is used for file format conversion in the examples.

2.3 HLA and SNP Datasets

The HLA genotypes of study individuals in the 1000 genomes project phase 1 are publicly available on the website of the 1000 Genomes [13]. To simply the demonstration, we use the standard datasets of HLA and SNPs from the 1000 genomes project, which have been prepared in the ImmPute project (http://immpute-project.immunogenomics.org) for benchmarking [12]. The ZIP file distributed to the ImmPute participants can be downloaded at http://igdawg.org/pubs/ImmPuteDataPackage.zip, which contains 10,268 SNPs in the extended human MHC and 930 subjects.

The pre-fit classifiers built on GSK HLARES data [6] can be downloaded from either http://www.biostat.washington.edu/~bsweir/HIBAG or http://zhengxwen.github.io/HIBAG/platforms.html. Platform-specific parameter estimates are available online for more than 20 common SNP chips.

3 Methods

This section comprises two exercises. In the first exercise, we will download a pre-fit classifier from the HIBAG website, apply it to a SNP dataset, and assess its predictive performance. The second exercise focuses on imputation model training and the approaches for software speedup will be discussed. The complete R scripts are also available online http://zhengxwen.github.io/HIBAG/tutorial_for_mimb.

3.1 Imputation with Pre-fit Classifiers

1. Go to the website of the ImmPute project (http://immpute-project.immunogenomics.org) and download the ZIP file "ImmPuteDataPackage.zip," which contains 1000 genomes SNP and HLA genotypic data. The SNP data in the ImmPute project were tailored to the Illumina ImmunoChip platform.
2. After downloading and uncompressing the ZIP file, use the PLINK software to convert the genotype PED file to the binary PED (BED) file:

```
plink --file KG --out KG --make-bed
```

3. Go to the webpage of HIBAG platform-specific parameter estimates (http://zhengxwen.github.io/HIBAG/platforms.html) and download the two-field pre-fit classifiers of Illumina ImmunoChip on hg19 human genomes "ImmunoChip-Broad-HLARES-HLA4-hg19.RData." These models were built with multi-ethnic GSK HLARES samples, and enable determining 2-field HLA types at the HLA-A, -B, -C, -DRB1, -DQA1, -DQB1, and -DPB1 loci.

4. Start an R session and import the models and the SNP dataset of 930 individuals:

```
library(HIBAG)
mlst <- get(load("ImmunoChip-Broad-HLARES-HLA4-hg19.RData"))
geno <- hlaBED2Geno("KG.bed", "KG.fam", "KG.bim")
```

5. Run the HIBAG prediction algorithm for HLA–A:

```
# load HLA-A pre-fit classifier to memory
model <- hlaModelFromObj(mlst$A)
# run prediction
hla_a <- hlaPredict(model, geno)
```

This process takes about 35 min to run, and we can utilize multiple cores to speed up the calculation:

```
# run with 8 cores
hla_a <- hlaPredict(model, geno, cl=8)
summary(hla_a)
## Gene: A
## Range: [29910247bp, 29913661bp] on hg19
## # of samples: 930
## # of unique HLA alleles: 38
## # of unique HLA genotypes: 218
## Posterior probability:
## [0,0.25) [0.25,0.5) [0.5,0.75) [0.75,1]
## 1 (0.1%) 36 (3.9%) 114 (12.3%) 779 (83.8%)
head(hla_a$value)
## sample.id allele1 allele2 prob matching
## 1 HG00096 01:01 29:02 0.9999180 0.0012829705
## 2 HG00097 03:01 24:02 0.9491243 0.0004386126
## 3 HG00099 01:01 68:01 0.9983373 0.0008045432
## ...
# write the imputation results to a text file
write.table(hla_a$value, file="result.txt", sep="\t",
    quote=FALSE, row.names=FALSE)
```

6. The posterior probabilities (the column "prob") calculated in the imputation process are used as confidence scores, and a value of 0.5 is suggested as the call threshold [6]. The call rate (the fraction of individuals successfully imputed) is 96% in this example according to the posterior probabilities greater than 0.5. The column "matching" is a measure or proportion describing how the SNP profile matches the SNP haplotypes observed in the training set, i.e., the likelihood of SNP profile in a random-mating population consisting of training haplotypes. Matching proportion is not directly related to confi-

dence score, but a very low value of "matching" indicates that it is underrepresented in the training set.

7. The ZIP file "ImmPuteDataPackage.zip" contains the actual HLA-A, -B, -C, -DRB1 and -DQB1 types of 930 individuals in the 1000 genomes project. We use these HLA genotypes to evaluate the HIBAG imputation at the HLA-A locus:

```
# read HLA data in the excel file
library(readxl)
hla <- read_excel("IDAWG_KG_HLAformattedIMMPUTE.xlsx")
# create an object of HLA genotypes with true alleles
true_a <- hlaAllele(hla$id, hla$A__1,
    max.resolution="4-digit", locus="A")
# evaludate the prediction w/o and with call threshold
rv_ct0 <- hlaCompareAllele(true_a, hla_a, call.threshold=0)
rv_ct5 <- hlaCompareAllele(true_a, hla_a, call.threshold=0.5)
```

The function hlaCompareAllele() returns the details of comparison between the imputed alleles and the true ones, including overall prediction accuracies at the individual and allelic levels, a confusion matrix, sensitivity, specificity, positive predictive value and negative predictive value for each allele. For example, the overall allelic accuracies are 94.8% and 93.4% with and without call threshold respectively, i.e., the number of correctly imputed alleles over the total number of alleles:

```
rv_ct0$overall$acc.haplo
## 0.9344086
rv_ct5$overall$acc.haplo
## 0.9479283
```

8. The HIBAG package provides functionalities for formatting the prediction evaluation in a table with text, html, tex, or markdown format. Running the function hlaReport() as follows and the output file is "allele.md" as shown in Fig. 2. The columns in the table are allele name, accuracy, sensitivity, specificity, positive predictive value and negative predictive value, etc.

```
hlaReport(rv_ct0, export.fn="allele.md", type="markdown")
```

9. The relationships among imputation accuracy, call rate, and call threshold are shown in Fig. 3a, b. The prediction accuracy increases with more restrictive call threshold. The function hlaReportPlot() creates a figure for call rate and call threshold respectively:

```
hlaReportPlot(hla_a, true_a, fig="call.rate")
hlaReportPlot(hla_a, true_a, fig="call. threshold")
```

10. Repeat **step 5–9** for HLA-B, -C, -DRB1, and -DQB1 respectively. The overall accuracies of five HLA genes are shown in Table 1, ranging from 90% to 97%, with and without call threshold.

Overall accuracy: 93.4%, Call rate: 100.0%

Allele	# Valid.	Freq. Valid.	CR (%)	ACC (%)	SEN (%)	SPE (%)	PPV (%)	NPV (%)	Miscall (%)
01:01	111	0.0597	100.0	99.6	97.3	99.8	96.4	99.8	03:01 (67)
01:02	5	0.0027	100.0	99.8	20.0	100.0	100.0	99.8	01:01 (100)
02:01	362	0.1946	100.0	97.4	96.4	97.7	90.9	99.1	02:07 (92)
02:02	22	0.0118	100.0	99.6	68.2	100.0	100.0	99.6	02:05 (100)
02:03	17	0.0091	100.0	99.1	0.0	100.0	–	99.1	02:01 (100)
02:04	1	0.0005	100.0	99.9	0.0	100.0	–	99.9	02:01 (100)
02:05	15	0.0081	100.0	99.5	100.0	99.5	62.5	100.0	–
02:06	41	0.0220	100.0	99.9	100.0	99.9	95.3	100.0	–
02:07	35	0.0188	100.0	99.1	85.7	99.3	71.4	99.7	02:01 (100)
02:10	2	0.0011	100.0	99.9	0.0	100.0	–	99.9	02:06 (100)
02:11	5	0.0027	100.0	99.6	0.0	99.9	0.0	99.7	02:01 (100)
02:14	1	0.0005	100.0	99.9	0.0	100.0	–	99.9	02:05 (100)

Fig. 2 Descriptive summaries with markdown format for per-allele sensitivity, specificity, positive predictive value, and negative predictive value of HLA-A imputation

3.2 Model Training

In this section, the 1000 genomes SNP and HLA data are used as a training set to build a multi-ethnic model for HLA-A prediction. The model training is a time-consuming process, and the approaches for accelerating the calculation will be discussed.

3.2.1 Model Training Under Regular Settings

1. After downloading and uncompressing "ImmPuteDataPackage.zip," the PLINK software is used to create a PLINK BED file for SNP genotypes (*see* **step 2** in Subheading 3.1).
2. Start an R session and import the SNP and HLA genotypes, consisting of 930 individuals in the 1000 genomes project:

```
library(HIBAG)
# import SNP data
geno <- hlaBED2Geno("KG.bed", "KG.fam", "KG.bim")
# read HLA data in the excel file
library(readxl)
hla <- read_excel("IDAWG_KG_HLAformattedIMMPUTE.xlsx")
```

3. Use the HLA data stored in the excel file to create an R object of true HLA–A genotypes, and determine the flanking SNPs to be trained. It is important to specify the human genome reference "hg19" according to the SNP data, since the location of the specified HLA gene is required to determine the

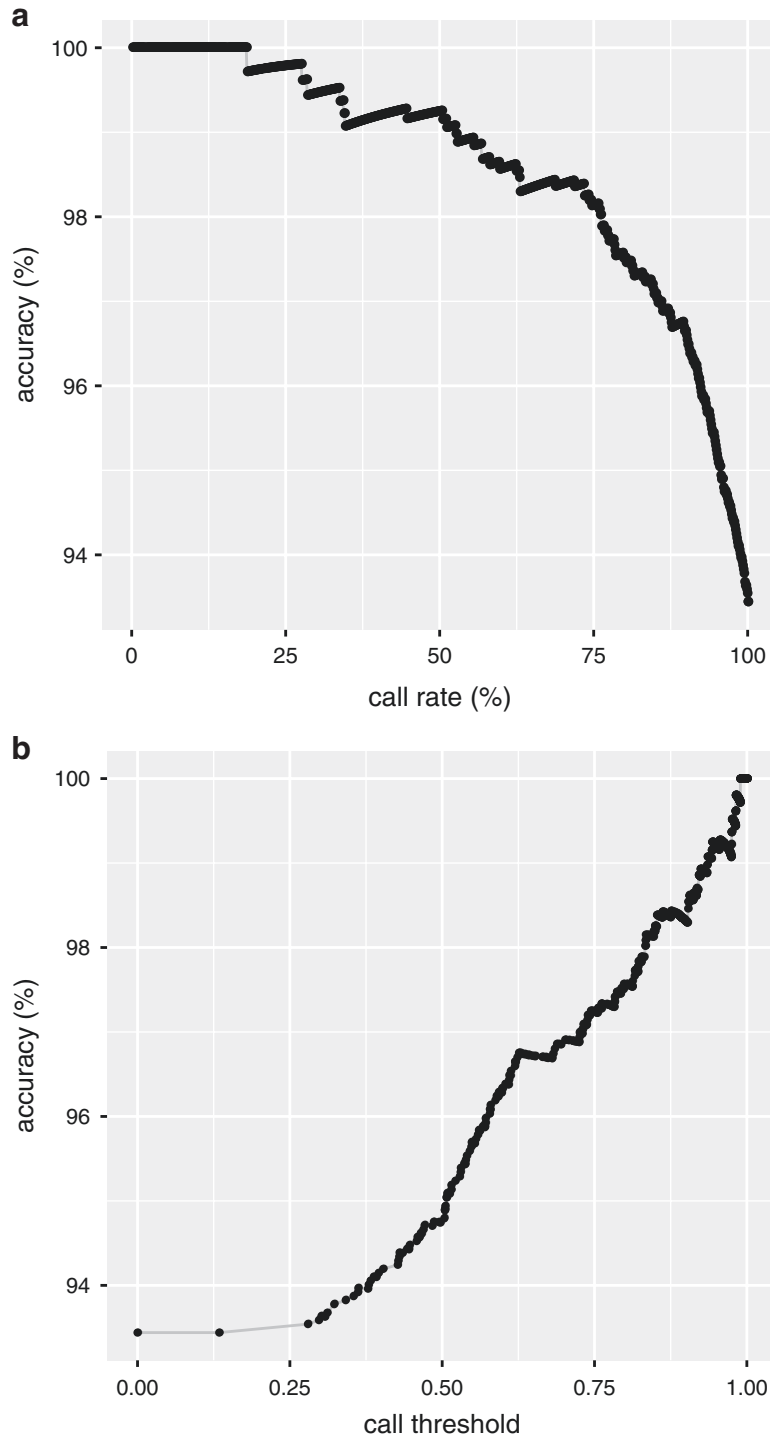


Fig. 3 The relationships among HLA-A accuracy, call rate, and call threshold using the HLARES pre-fit classifiers and the validation samples of 1000 Genomes

Table 1
Summary of the 2-field prediction accuracies and call rates for the HLARES pre-fit classifiers, using 1000 genomes HLA data as validation samples

	HLA-A	HLA-B	HLA-C	HLA-DRB1	HLA-DQB1
# of training samples	2901	3886	2916	3713	2985
# of validation samples	930	930	930	930	930
Accuracy, no call threshold					
	93.4%	91.1%	97.2%	90.8%	91.2%
Accuracy, call threshold=0.5					
	94.8%	96.1%	97.7%	96.4%	91.6%
Call rate (%)	96.0	84.8	97.7	79.5	98.4

SNPs in the flanking region of that HLA gene. A flanking region of 500 kb on each side is recommended.

```
# HLA genotypes
hla_a <- hlaAllele(hla$id, hla$A, hla$A__1,
    max.resolution="4-digit", locus="A", assembly="hg19")
# the SNP set used in model training
snpsel <- hlaFlankingSNP(geno$snp.id, geno$snp.position,
    "A", flank.bp=500*1000, assembly="hg19")
# SNP genotypes used in model traning
train.geno <- hlaGenoSubset(geno,
    snp.sel=geno$snp.id %in% snpsel)
summary(train.geno)
```

4. The training algorithm is available in the function hlaAttrBagging(). By default, 100 individual classifiers will be generated to build an ensemble model, however this process is very time-consuming and will take ~4 days to run on a single core (Intel Xeon CPU E5-2630L @2.40GHz):

```
model <- hlaAttrBagging(hla_a, train.geno, nclassifier=100)
## Accuracy with training data: 98.7%
## Out-of-bag accuracy: 95.8%
```

To finish this example and avoid the long waiting, you could try building one classifier (nclassifier = 1) instead of 100. The approaches for accelerating the calculation will be shown in Subheadings 3.2.2, 3.2.3, and 3.2.4.

5. After model training completes, we save this model into an R object file for future uses:

```
mobj <- hlaModelToObj(model)
save(mobj, file="hla_a_model.rdata")
```

6. The HIBAG model can be visualized in a scatterplot showing the relationship between the frequency of SNP uses in an individual classifier and genome coordinate (shown in Fig. 4):

```
plot(model)
```

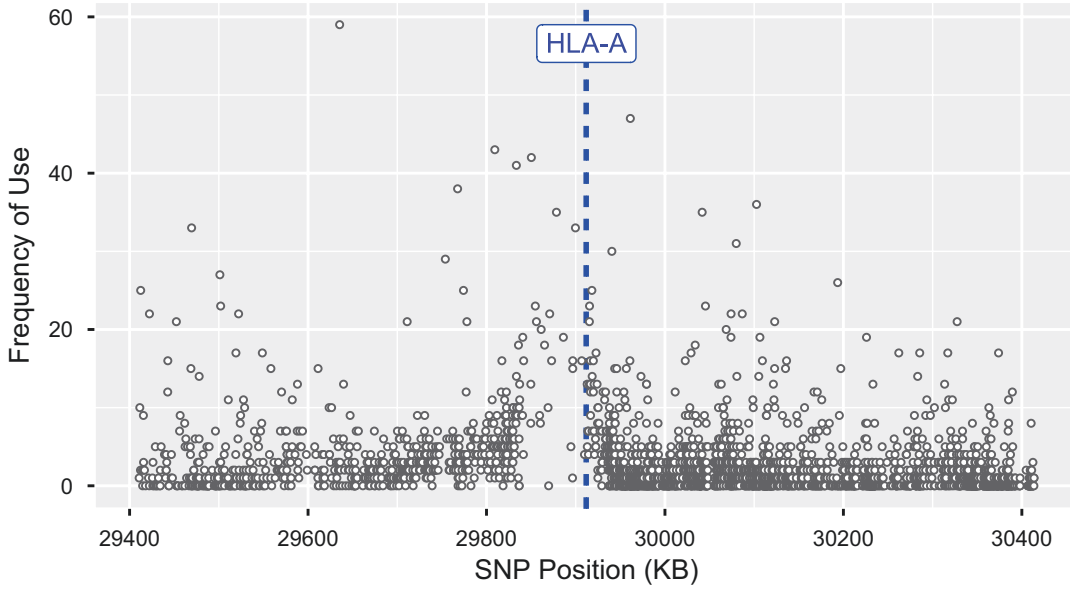


Fig. 4 The number of classifiers used in the HIBAG model for each SNP marker. The ensemble model consists of 100 individual classifiers, and more important SNP markers tend to be used more frequently

3.2.2 Software Optimization with Intel POPCNT Intrinsic

If users install the HIBAG package from Bioconductor via "biocLite("HIBAG")", by default the installation does not enable the compiler optimization according to the native CPU architecture. In order to utilize the latest Intel/AMD intrinsics, e.g., POPCNT for obtaining the count of number of bits set to one, we need to customize the package compilation.

Following the R CRAN instruction http://cran.r-project.org/doc/manuals/r-release/R-admin.html#Customizing-package-compilation, we modify *HOME*/.R/Makevars to enable native intrinsics when using GNU compiler:

```
HIBAG (HLA Genotype Imputation with Attribute Bagging)
Kernel Version: v1.4
Supported by Streaming SIMD Extensions (SSE2 + POPCNT)
```

```
## for C code
CFLAGS=-g -O3 -march=native -mtune=native
## for C++ code
CXXFLAGS=-g -O3 -march=native -mtune=native
```

If the package compilation succeeds, you should see a welcome message after loading the package:

The HIBAG algorithm internally measures how a pair of SNP haplotypes matches with observed SNP genotypes. The SNP alleles are stored in a bit vector, and the Hamming distance is calculated

via built-in POPCNT intrinsic when comparing genotypes. We rerun the model fitting:

```
model <- hlaAttrBagging(hla_a, train.geno, nclassifier=100)
## Accuracy with training data: 98.7%
## Out-of-bag accuracy: 95.8%
```

The training process takes ~2.4 days, saving 40% of running time.

3.2.3 Multi-Core Settings

The HIBAG ensemble model consists of independent individual classifiers, therefore the parallel implementation is straightforward, i.e., building individual classifiers separately on different cores. Below, eight cores are used to run model training:

```
model <- hlaParallelAttrBagging(8, hla_a, train.geno,
    nclassifier=100)
## Accuracy with training data: 98.5%
## Out-of-bag accuracy: 95.7%
```

The job-level parallelism on loosely coupled compute clusters is also supported via communication over sockets. Using the framework implemented in the R package "parallel", a compute cluster object can be passed to the first argument of hlaParallelAttrBagging().

The accuracy with training data and out-of-bag accuracy may not be the same in each running, since bootstrap sampling and random variable selection in the HIBAG algorithm introduce some degree of randomness.

3.2.4 GPU Computing

To complete the example in this section, a GPU card should be available on your desktop or workstation as well as the OpenCL headers and library. After installing the HIBAG.gpu package, we start a new R session and run the following scripts:

```
Loading required package: OpenCL
Available OpenCL platform(s):
NVIDIA CUDA, OpenCL 1.2 CUDA 8.0.0
Device #1: NVIDIA Corporation Tesla M80
Device #2: NVIDIA Corporation Tesla M80
Using Dev#1: NVIDIA Corporation Tesla M80
GPU device supports 64-bit floating-point numbers.
```

```
library(HIBAG.gpu)
```

Here is a typical welcome message after loading the package,

When there are more than one GPU device available in the same machine, we can switch to other devices instead of the default one:

```
hlaGPU_Init(2)
```

```
Using Dev#2: NVIDIA Corporation Tesla M80
GPU device supports 64-bit floating-point numbers
```

When the computing equipment is ready, model training is as simple as calling the function hlaAttrBagging_gpu() in the HIBAG.gpu package:

```
model <- hlaAttrBagging_gpu(true_a, train.geno,
    nclassifier=100)
## Accuracy with training data: 98.7%
## Out-of-bag accuracy: 95.6%
```

The GPU computing process takes ~3.2 h and it is almost 18 times faster than one-core implementation in Subheading 3.2.2.

4 Discussion

It is important to realize the potential limitations, while using HIBAG for HLA imputation:

1. The numbers of HLA alleles documented in the IMGT-HLA database [16] are much larger than the numbers investigated in the studies of HLA imputation. For example, the numbers of protein-level HLA alleles from IMGT are 2781, 3501, and 2490 at the HLA-A, -B, and -C loci listed in June 2017, and new alleles are routinely being discovered. However, we have only 85, 144, and 49 alleles in GSK HLARES pre-fit classifiers. Quite large training sets are required to successfully predict more HLA alleles in the IMGT-HLA dataset.
2. As shown in our previous study [6], the overall accuracies of European ancestry increase with the training sample size, but are only slightly improved after 500 training samples. Rare alleles with frequency <1% have significantly lower prediction accuracies than the common alleles. The size of sample sets required to accurately type rare alleles is impractical using an imputation methodology.
3. The accuracy of HIBAG imputation depends on various factors. One of the most important factors is how the imputed population is evolutionary closely related to reference. The subpopulations in Europe include British, French, German, Finnish, and Italian according to geography, culture, genetics, and other factors. The HLARES pre-fit classifiers of European ancestry were built on individuals from Europe, North America, and others. These models may not include a sufficiently large number of samples in ethnic minorities, hence the prediction accuracy is not always as high as what reported in the publication.
4. In general, the HIBAG method is robust to missing SNP markers of targeted genotyping platform. So, it is feasible to

impute samples on a new GWAS genotyping chip using existing platform-specific classifiers if there is no too much difference in SNP markers between the two platforms.

5. Multi-ethnic pre-fit classifiers are downloadable as well as ethnic-specific models. Multi-ethnic models may be more effective than ethnic-specific models, especially when racial information of imputed samples is unknown or admixed samples are underrepresented in the training set. For the samples with known ancestries, multi-ethnic classifiers could be used in the sensitivity analysis and we could compare the multi-ethnic prediction with ethnic-specific results, since the similar imputation results are expected.

References

1. Shiina T, Hosomichi K, Inoko H, Kulski JK (2009) The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet 54(1):15–39
2. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L et al (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42(Database issue):D1001–D1006
3. Bauer DC, Zadoorian A, Wilson LO, Thorne NP, Alliance MGH (2016) Evaluation of computational programs to predict HLA genotypes from genomic sequencing data. Brief Bioinform pii:bbw097
4. Erlich H (2012) HLA DNA typing: past, present, and future. Tissue Antigens 80(1):1–11
5. Meyer D, Nunes K (2017) HLA imputation, what is it good for? Hum Immunol 78(3):239–241
6. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, Weir BS (2014) HIBAG–HLA genotype imputation with attribute bagging. Pharmacogenomics J 14(2):192–200
7. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140
8. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
9. Khor SS, Yang W, Kawashima M, Kamitsuji S, Zheng X, Nishida N, Sawai H, Toyoda H, Miyagawa T, Honda M et al (2015) High-accuracy imputation for HLA class I and II genes based on high-resolution SNP data of population-specific references. Pharmacogenomics J 15(6):530–537
10. Levin AM, Adrianto I, Datta I, Iannuzzi MC, Trudeau S, McKeigue P, Montgomery CG, Rybicki BA (2014) Performance of HLA allele prediction methods in African Americans for class II genes HLA-DRB1, -DQB1, and -DPB1. BMC Genet 15:72
11. Nunes K, Zheng X, Torres M, Moraes ME, Piovezan BZ, Pontes GN, Kimura L, Carnavalli JE, Mingroni Netto RC, Meyer D (2016) HLA imputation in an admixed population: an assessment of the 1000 genomes data as a training set. Hum Immunol 77(3):307–312
12. Pappas DJ, Lizee A, Paunic V, Beutner KR, Motyer A, Vukcevic D, Leslie S, Biesiada J, Meller J, Taylor KD et al (2017) Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. Pharmacogenomics J. https://doi.org/10.1038/tpj.2017.7
13. Gourraud PA, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, Rioux JD, Hauser S, Oksenberg J (2014) HLA diversity in the 1000 genomes dataset. PLoS One 9(7):e97282
14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575
15. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28(24):3326–3328
16. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG (2015) The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 43(Database issue): D423–D431

Chapter 12

In Silico Typing of Classical and Non-classical HLA Alleles from Standard RNA-Seq Reads

Sebastian Boegel, Thomas Bukur, John C. Castle, and Ugur Sahin

Abstract

Next-Generation Sequencing (NGS) enables the rapid generation of billions of short nucleic acid sequence fragments (i.e., "sequencing reads"). Especially, the adoption of gene expression profiling using whole transcriptome sequencing (i.e., "RNA-Seq") has been rapid. Here, we describe an *in silico* method, seq2HLA, that takes standard RNA-Seq reads as input and determines a sample's (classical and non-classical) HLA class I and class II types as well as HLA expression. We demonstrate the application of seq2HLA using publicly available RNA-Seq data from the Burkitt's lymphoma cell line DAUDI and the choriocarcinoma cell line JEG-3.

Key words HLA type, HLA expression, NGS, RNA-Seq, Immunoinformatics, In silico, Non-classical HLA class I

1 Introduction

Human Leukocyte Antigen (HLA) proteins are one of the key players in the adaptive immune system as they present peptides to T lymphocytes. They are polygenic and encoded by highly polymorphic genes located on chromosome 6. The classical (HLA-A, -B, -C) HLA class I heavy chains assemble with β2-microglobulin (B2M) to form a stable heterodimer. In contrast, the HLA class II molecules HLA-DP, -DQ, and -DR consist of two membrane-spanning chains, α and β, which are each produced by separate HLA genes. Both HLA classes are highly polymorphic, i.e., there are many different variants of each HLA gene within the human population (Fig. 1). In addition, the HLA genes are expressed co-dominantly, such that every individual expresses (under physiological conditions) two alleles of each gene. Non-classical HLA class I molecules (HLA-E, -F, -G) have a similar protein structure to that of the classical HLA class I alleles and also require a bound peptide in the binding groove to form a stable complex. However,

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_12,

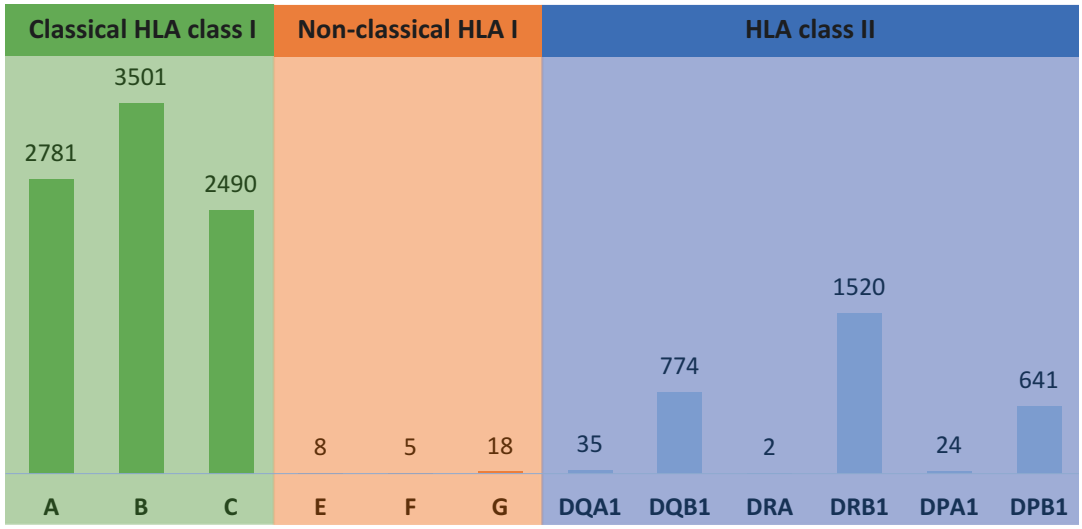


Fig. 1 HLA class I and class II (protein-coding) allele frequencies in humans. The HLA system is polygenic and highly polymorphic. Classical HLA class I (green) comprises three genes exhibiting a large variability. Non-classical HLA class I (red) is similar in structure and number of genes, which however show a much lower degree of variability. HLA class II (red) consists of three major loci (DR, DQ, DP) with each heavy chain (α and β) encoded in a separate gene. Numbers assigned as of October 2017 from http://hla.alleles.org/nomenclature/stats.html

they are characterized by few allelic polymorphisms and play a role in regulating innate immune responses [1].

Next-Generation Sequencing (NGS) enables the rapid generation of billions of short nucleic acid sequence fragments (i.e., "sequencing reads"). Especially, the adoption of gene expression profiling using whole transcriptome sequencing (i.e., "RNA-Seq") has been rapid with clinical and research laboratories worldwide depositing over 184,000 "human RNA-Seq" samples (as of October 2016) into the public repository Sequence Read Archive (SRA) (Table 1). In addition, due to efforts made by consortia such as the Genotype-Tissue Expression (GTEx) [9], The Cancer Genome Atlas (TCGA) [5] and the International Cancer Genome Consortium (ICGC) [10], many tumor and normal tissues, as well as cell line transcriptomes have been sequenced and raw datasets made publicly available.

Given the plethora of RNA-Seq datasets in the public domain and our efforts to develop an individualized T-cell-mediated cancer immunotherapy, in which the HLA type of a patient is a key parameter to prioritize neo-epitopes for vaccination [11–14], we sought to develop an algorithm to utilize the sequence content of RNA-Seq reads to determine HLA type and expression. However, one of the main challenges is the correct mapping of the sequence reads to reconstruct the HLA composition of the sample, which is aggravated by the polymorphic nature of the HLA loci and

Table 1
Sources for the retrieval of NGS datasets

Source	Description	URL	Refs
Sequence Read Archive (SRA)	Public repository of raw sequencing data from various NGS platforms	https://www.ncbi.nlm.nih.gov/sra	[2]
European Nucleotide Archive (ENA)	Repository of nucleotide sequencing information	https://www.ebi.ac.uk/ena	[3]
European Genome-Phenome Archive (EGA)	Archive of publicly available and protected access genetic and phenotypic datasets	https://ega-archive.org/	[4]
International Cancer Genome Consortium (ICGC)	Protected access NGS datasets from various cancer tissues	https://icgc.org/	[5]
The database of Genotypes and Phenotypes (dbGaP)	Archive and distribution platform for sequencing data from various studies. Protected access	https://www.ncbi.nlm.nih.gov/gap	[6]
Repositive	Online platform indexing genomic data that is stored in repositories	https://repositive.io/	[7]
MetaSRA	Normalized metadata for SRA; searchable web interface	http://metasra.biostat.wisc.edu/	[8]

resulting in a highly sequence similarity between the different alleles. Furthermore, the single human reference genome (e.g., GRCh37/hg19) does not adequately reflect the highly polymorphic nature of the HLA loci in the human population, confounding read alignment to a standard reference [8]. In addition, distinguishing between homozygous (e.g., HLA-A*02:01/HLA-A*02:01) and heterozygous loci (e.g., HLA-A*02:01/HLA-A*24:01) is aggravated, as the transcriptome lacks information about the maternal and paternal alleles. Using transcriptome-based data rather than genomic data provides challenges as well as benefits: the transcriptome reads reflect both the HLA types (the sequence) and the HLA expression levels (the count). If an HLA locus or allele is not expressed, it cannot be determined by RNA sequencing. On the other hand, knowledge of the HLA expression levels can be extremely informative: for example, HLA downregulation or loss is an established tumor escape mechanism [15, 16].

Acknowledging this advantage, we developed an efficient algorithm, called "seq2HLA," which takes standard RNA-Seq reads as input and outputs the most likely classical and non-classical HLA class I as well as classical HLA class II types, a certainty score for each call as well as the expression each of locus. Here, we will present the application of seq2HLA using publicly available RNA-Seq data from two human cancer cell lines.

2 Materials

To promote broad acceptance in the scientific community, seq2HLA exploits widely used, open source tools enabling easy installation and usage (Fig. 2).

2.1 Hardware

Seq2HLA was developed on SUSE Linux Enterprise Server 11 (×86_64). In order to run seq2HLA a computer with a POSIX environment, such as Linux or Mac OS, is required.

2.2 Software

To run seq2HLA, the following tools and packages must be installed:

1. Seq2HLA was written in *Python 2.6.8* (https://www.python.org). We recommend using Python 2.6.8 or a newer 2.x version. In addition, seq2HLA uses *biopython* (v 1.58 or newer) [17] and *numpy* (version 1.3.0 or newer, http://www.numpy.org).

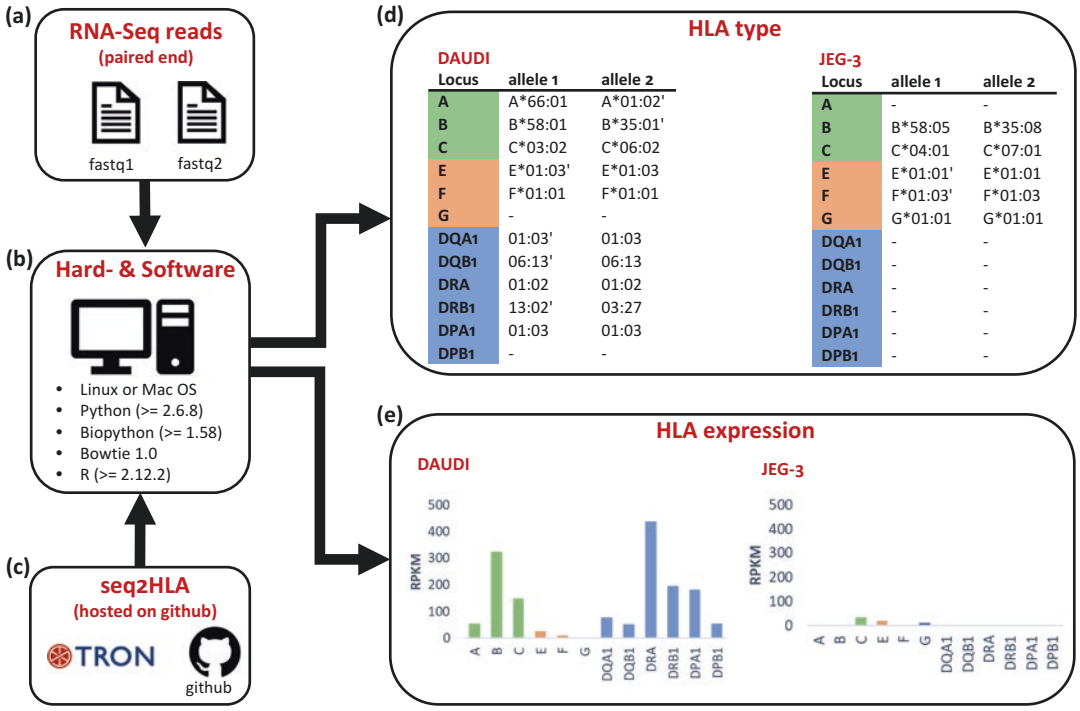


Fig. 2 Workflow of seq2HLA. (**a**) RNA-Seq samples in fastq format can be derived from public resources or own experiments. (**b**) In order to run seq2HLA a computer with a POSIX environment, such as Linux or Mac OS is required. Additionally, the following tools need to be available: Python, Biopython, Bowtie, and R. (**c**) The latest version of seq2HLA is available on github (https://github.com/TRON-Bioinformatics/seq2HLA). From the paired-end RNA-Seq data, seq2HLA determines the 4-digit (classical and non-classical) HLA class I and class II type (**d**) as well as expression levels (**e**). Shown here are the results for two cancer cell lines, DAUDI and JEG-3, that we use as example throughout this protocol

2. Seq2HLA uses a bowtie index comprising known HLA alleles and *Bowtie 1.0* [18] as a read aligner. It is one of the most widely used programs for aligning short sequencing reads to a reference genome. The environment variable must be set, such that Bowtie can be executed by the command "bowtie."
3. For the calculation of confidence scores, the statistical programming language *R* (version 2.12.2 or newer, https://www.r-project.org/) is used.

2.3 Download seq2HLA

Source code of seq2HLA is available via a Github repository (https://github.com/TRON-Bioinformatics/seq2HLA). A clone of the software can be retrieved using the Linux command line tool "git" with the following command:

```
git clone https://github.com/TRON-Bioinformatics/seq2HLA.git
```

This will copy the complete repository to the user's computer, including the Python executable seq2HLA.py. Additionally, seq2HLA can be downloaded interactively from the mentioned repository via the "Download ZIP" button, resulting in the download of a compressed folder containing the complete repository.

2.4 Get RNA-Seq Datasets

1. Seq2HLA was developed to run with standard paired-end RNA-Seq data produced by Illumina sequencing instruments. Here, "standard" refers to the library preparation: seq2HLA requires no special protocol isolating and enriching for HLA genes prior sequencing. Input of seq2HLA can be uncompressed or gzipped fastq files [ref], which contains the sequencing reads and the associated quality score for each base call. It has become the de facto standard for storing and distributing the output of NGS-based experiments.
2. As seq2HLA uses standard RNA-Seq reads not requiring a change to laboratory protocols, it is applicable to both existing and future datasets, either from own sequencing experiments or from one of the many public databases (*see* Table 1).
3. Throughout this protocol, we use paired-end RNA-Seq data from two widely used human cancer cell lines, downloaded from the Sequence Read Archive (SRA): Burkitt's lymphoma cell line DAUDI (SRA-ID: SRR387401) [19] and choriocarcinoma JEG-3 (SRA-ID: SRR3317028) [20], which is a trophoblastic model system known used to study the tissue-specific expression of HLA-G.
4. To download sequenced data from SRA, *sratoolkit* (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software) is needed. If **</path/to/sratoolkit/>** denotes the full path to your local copy of sratoolkit, we are able to download the RNA-Seq data for or example cell line DAUDI using its SRA-ID SRR387401 with the following command:

/path/to/sratoolkit/bin/fastq-dump --split-3 --gzip SRR387401.

This command downloads the SRA-file, splits it into two fastq files (as it is paired-end), and compresses them with gzip. Notwithstanding, seq2HLA accepts uncompressed and compressed fastq files as input.

3 Methods

3.1 Running seq2HLA

1. If **/path/to/seq2HLA/** is the full path to your local copy of seq2HLA, then the tool can be executed by using the console call:

```
/path/to/seq2HLA/seq2HLA.py-1<fastq1>-2<fastq1>-r "<runname>"[-p < int>]
[-3<int>]
```

2. **<fastq1>** and **<fastq2>** enotes the paths to the (uncompressed or gzipped) fastq files of the paired-end RNA-Seq reads.
3. **<runname>** denotes the preix every output file of this run. If you want to store the output in a different folder, you need to give the full path and the prefix, e.g., "/path/to/output/DAUDI_ SRR387401."
4. **-p** sets the number of parallel search threads for bowtie. This parameter is optional and per default threads are used.
5. **−3** is a bowtie parameter allowing us to trim a given number of nucleotides from the low-quality end of each read. This is again optional and the default value is 0.
6. Using the paired-end fastq files of DAUDI as an example, seq2HLA can be executed by the command:

```
python/path/to/seq2HLA/seq2HLA.py-1/path/to/seq/ SRR387401_1.fastq.gz-2/path/
to/seq/SRR387401_2.fastq.gz -r "DAUDI_ SRR387401".
```

3.2 Interpreting the Output

All (final and intermediate) results of seq2HLA are output to *stdout* and stored into text files in the working directory. At the end of a seq2HLA run, all intermediate files are deleted. Only the files containing final results are kept. Of these, the most important files are listed in Table 2.

In addition, for each HLA class, the output of bowtie containing mapping statistics is stored in textfiles with the suffix ".bowtielog."

3.2.1 HLA Typing

1. By aligning the RNA-Seq reads against a reference database of HLA alleles, seq2HLA first determines the HLA groups (i.e., 2-digit resolution), zygosity and calculates a confidence score for each HLA allele call [21].
2. The confidence score is a measure for the gap between the top HLA group (which has the reads mapped to it) and the

Table 2
Most important output files produced by seq2HLA using the run with runname "DAUDI_ SRR387401" as an example

Filename	Description
DAUDI_SRR387401-ClassI-class.HLAgenotype4digits	Classical HLA class I typings and associated certainty scores
DAUDI_SRR387401-ClassI-nonclass.HLAgenotype4digits	Non-classical HLA class I typings and associated certainty scores
DAUDI_SRR387401-ClassII.HLAgenotype4digits	HLA class II typings and associated certainty scores
DAUDI_SRR387401-ClassI-class.expression	Expression values of classical HLA class I loci in RPKM
DAUDI_SRR387401-ClassI-nonclass.expression	Expression values of non-classical HLA class I loci in RPKM
DAUDI_SRR387401-ClassII.expression	Expression values of HLA class II loci in RPKM
DAUDI_SRR387401.ambiguity	In case of typing ambiguities, all possible solutions are reported here

background mappings (i.e., reads mapping to HLA groups containing alleles with high sequence similarities). The confidence score is the result of a test determining whether the top HLA group is an outlier given the distribution of all HLA group read counts. A larger gap results in a clearer separation, a smaller confidence score, and thus in a more confident typing result (Fig. 3). In cases, in which either no reads map at all or the reads map to one HLA group, "NA" is returned as no outlier can be calculated from a distribution consisting of only one or no value.

3. In the next step, seq2HLA refines the typings and confidence scores to determine the 4-digit resolution (i.e., the actual protein presented on the cell surface) [23].
4. A consequence of the highly polymorphic nature of the HLA system is the sequence similarity of the alleles within a group. Using short reads increases the complexity of reconstructing the individual HLA composition of a sample as reads may map to multiple alleles. Thus, it is not always possible to find a unique typing. In these cases, the most likely 4-digit typing with an ambiguity flag (",") attached and all possible solutions are stored in the file "<prefix>.amiguity." In our DAUI example, there are typing ambiguities at HLA-A*01:02, B*35:01 (Table 3a) and at HLA-DQA1*01:03, HLA-DQB1*06:13 and HLA-DRB1*13:02 (Table 4a) as well as at HLA-E*01:03 (Table 5a).

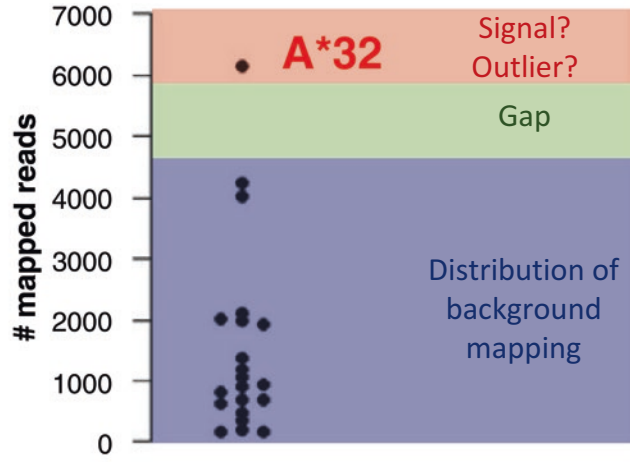


Fig. 3 Interpretation of the confidence score. It measures the gap between signal (i.e., reads mapping to the true HLA group) and noise (i.e., reads mapping to any other HLA group due to the polymorphic nature of HLA) by calculating the probability of the signal being an outlier given the distribution of the mapped reads per HLA group (black circles). The resulting score indicates the goodness of the separation between the true solution and the background. The larger the gap, the more confident is the typing (detection of the signal) and this is reflected by a smaller confidence score (Figure adapted from [22])

5. Using transcriptome data as input, seq2HLA is only able to determine an HLA allele, if it is expressed. In our DAUI example, no HLA-G transcripts can be detected (Table 6a) thereby preventing HLA-G typing (Table 5a). Similarly, in our JEG-3 example, HLA-DRB1 is the only locus for which seq2HLA was able to determine the HLA allele with a low (good) confidence score (Table 4b). However, expression of this allele is so low (0.08 RPKM, Table 6b), it is very much likely that this mapping is due to chance and thus not a real signal (*see* **Note 1**).

3.2.2 HLA Expression

1. Using RNA reads, seq2HLA is not only able to utilize the sequence content to determine HLA type, but also the abundance of the typed alleles. Seq2HLA provides locus-specific expression values for HLA class I (classical and non-classical) and HLA class II (Fig. 2, Table 6), normalized according to "reads per kilobase of exon model per million mapped reads" (RPKM) [24], which is a still widely used normalization measure. RPKM normalizes the number of reads mapping to transcripts in the reference file by the length of this transcript (*see* **Note 2**) and the total number of reads.
2. In the downloaded DAUDI example, classical HLA class I and HLA class II molecules are highly expressed (53–326 and

Table 3
HLA class I types in 4-digit resolution for a) DAUDI (SRR387401) stored in DAUDI_SRR387401-ClassI-class.HLAgenotype4digits and b) JEG-3 (SRR3317028) stored in JEG3_SRR3317028-ClassI-class.HLAgenotype4digits

	(a) DAUDI				(b) JEG-3			
Locus	Allele 1	Confidence	Allele 2	Confidence	Allele 1	Confidence	Allele 2	Confidence
A	A*66:01	0.049	A*01:02’	0.009	No	NA	No	NA
B	B*58:01	0.0	B*35:01’	0.0002	B*58:05	0.4	B*35:08	0.3
C	C*03:02	0.009	C*06:02	0.02	C*04:01	0.005	C*07:01′	0.01

Table 4
HLA class II types in 4-digit resolution for (a) DAUDI (SRR387401) stored in SRR387401-ClassII.HLAgenotype4digits and (b) JEG-3 (SRR3317028) stored in JEG3_SRR3317028-ClassII.HLAgenotype4digits

	(a) DAUDI				(b) JEG-3			
Locus	Allele 1	Confidence	Allele 2	Confidence	Allele 1	Confidence	Allele 2	Confidence
DQA1	01:03’	0.0	01:03	NA	No	NA	No	NA
DQB1	06:13’	0.00009	06:13	NA	No	NA	No	NA
DRB1	13:02′	0.0002	03:27	0.035	No	NA	No	NA
DRA	01:02	NA	01:02	NA	No	NA	No	NA
DPA1	01:03	0.69	01:03	0.06	No	NA	No	NA
DPB1	106:01	0.05	141:01	0.002	96:01	0	96:01	NA

52–440 RPKM, respectively). HLA-E and HLA-F are expressed at medium (25 RPKM) and low levels (8 RPKM), respectively, whereas HLA-G is not present at all (Table 6).

3. JEG-3 is a widely used trophoblastic model cell line, which is known to express HLA-G and HLA-C [25]. In our downloaded JEG-3 sample, we find indeed expression of HLA-C and -G (35 and 11 RPKM, respectively), very low abundance of HLA-B (0.6 RPKM), and no evidence of HLA-A and HLA class II (Table 6).

3.3 Example Applications

1. Using the host immune system as means to control cancer is the goal of cancer immunotherapy, which was announced as breakthrough of the year 2013 [26]. Mutanome directed cancer immunotherapy [27] is a novel concept exploiting the landscape of somatic nonsynonymous mutations displayed by a tumor (i.e., the mutanome) as targets for a personalized immunotherapy. A fraction of the mutations is presented on HLA class I or HLA class II molecules on the surface of the tumor or on antigen presenting cells (APCs). Knowing the HLA type of the patient allows us to predict HLA class I and class II neo-epitopes (i.e., mutated peptides, which are likely to induce a T-cell response and exert a potent antitumoral effect), that can be used for therapeutic vaccinations [14].

   As an example, we used seq2HLA in conjunction with primary tumor RNA-Seq and somatic mutation data from TCGA in conjunction with HLA-binding prediction tools to approximate the number of likely presented neo-antigens on HLA class II in various tumor entities [12]. Seq2HLA is widely accepted by the scientific community and used by other to

Table 5
Non-classical HLA class I type in 4-digit digit resolution for (a) DAUDI (SRR387401) stored in SRR387401-ClassI-nonclass.HLAgenotype4digits and (b) JEG-3 (SRR3317028) stored in JEG3_SRR3317028-ClassI-nonclass.HLAgenotype4digits

	(a) DAUDI				(b) JEG-3			
Locus	**Allele 1**	**Confidence**	**Allele 2**	**Confidence**	**Allele 1**	**Confidence**	**Allele2**	**Confidence**
E	E*01:03’	NA	E*01:03	NA	E*01:01’	NA	E*01:01	NA
F	F*01:01	NA	F*01:01	NA	F*01:03’	NA	F*01:03	NA
G	No	NA	no	NA	G*01:01	NA	G*01:01	NA

Table 6
Expression of classical and non-classical HLA class I and HLA class II genes for (a) DAUDI (SRR387401) stored in DAUDI_SRR387401-ClassI-class.expression, DAUDI_SRR387401-ClassI-nonclass.expression, DAUDI_SRR387401-ClassII.expression and b) JEG-3(SRR3317028) stored in JEG3_SRR3317028-ClassI-class.expression, JEG3_SRR3317028-ClassI-nonclass.expression, JEG3_SRR3317028-ClassII.expression

	(a) DAUDI	(b) JEG-3
Locus	**Expression [RPKM]**	
Classical HLA class I		
A	53.2	0
B	325.52	0.58
C	149.15	35.11
Non-classical HLA class I		
E	25.59	20.44
F	8.27	0.04
G	0	11.32
HLA class II		
DQA1	76.94	0
DQB1	51.51	0
DRB1	197.1	0
DRA	439.4	0
DPA1	183.38	0
DPB1	53.92	0.08

assign HLA types from RNA-Seq data to identify neo-antigen landscapes in different tumor entities [28, 29].

2. Using seq2HLA and other bioinformatic tools in conjunction with publicly available raw RNA-Seq data, we determined the HLA type and abundance, identified expressed viruses, calculated gene expression, and predicted antigenic mutations of 1082 human cancer cell lines [30]. All results are integrated into web-based portal, which provides an interactive user interface with advanced search capabilities. It is freely accessible at http://celllines.tron-mainz.de. Also, this resource is widely used in the scientific community for the identification of cell lines based on a specific HLA type [31, 32] or predicted neo-antigens [33].

4 Notes

1. In cases of very low expression (smaller than 1 RPKM) it is worth looking into the logfile of bowtie. In this case xxxxbowtielog reports that only one read (out of 30 million) could be aligned to this allele, which strengthens the hypothesis that this typing is due to chance.
2. Seq2HLA works by aligning the RNA-Seq against a reference database of HLA alleles, which contains only the sequences of the exons encoding the peptide-binding site and thus contains most of the polymorphisms [21]. So, the actual length of the sub-transcripts contained in this reference dataset is: 694 nucleotides (nts) for class I alleles, 400, 421, 421, 405, 396, 415 nts for class II alleles (DQA1, DQB1, DRB1, DRA, DPA1, and DPB1 respectively), and 694, 1089, 694 nts for non-classical HLA class I alleles (E, F, and G respectively).
3. When determining non-classical HLA class I alleles (HLA-E, -F, -G), seq2HLA also measures HLA class I pseudogenes (HLA-H, -I, -K, –L, -P, -V) and stores them in the output files for non-classical HLA class I genes. As they are not producing proteins and their function is yet unknown [34], these genes are not mentioned in this protocol.

References

1. Dulberger CL, McMurtrey CP, Hölzemer A et al (2017) Human leukocyte antigen F presents peptides and regulates immunity through interactions with NK cell receptors. Immunity 46(6):1018–1029.e7. https://doi.org/10.1016/j.immuni.2017.06.002
2. Leinonen R, Sugawara H, Shumway M (2011) The sequence read archive. Nucleic Acids Res 39(Database):D19–D21. https://doi.org/10.1093/nar/gkq1019
3. Toribio AL, Alako B, Amid C et al (2017) European nucleotide archive in 2016. Nucleic Acids Res 45(D1):D32–D36. https://doi.org/10.1093/nar/gkw1106
4. Lappalainen I, Almeida-King J, Kumanduri V et al (2015) The European genome-phenome archive of human data consented for biomedical research. Nat Genet 47(7):692–695. https://doi.org/10.1038/ng.3312
5. Hudson TJ, Anderson W, Artez A et al (2010) International network of cancer genome projects. Nature 464(7291):993–998. https://doi.org/10.1038/nature08987
6. Tryka KA, Hao L, Sturcke A et al (2014) NCBI's database of genotypes and phenotypes: DbGaP. Nucleic Acids Res 42(Database issue):D975–D979. https://doi.org/10.1093/nar/gkt1211
7. Kovalevskaya NV, Whicher C, Richardson TD et al (2016) DNAdigest and repositive: connecting the world of genomic data. PLoS Biol 14(3):e1002418. https://doi.org/10.1371/journal.pbio.1002418
8. Bernstein MN, Doan A, Dewey CN (2017) MetaSRA: normalized human sample-specific metadata for the sequence read archive. Bioinformatics 33(18):2914–2923. https://doi.org/10.1093/bioinformatics/btx334
9. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45(6):580–585. https://doi.org/10.1038/ng.2653
10. Weinstein JN, Collisson EA, Mills GB et al (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45(10):1113–1120. https://doi.org/10.1038/ng.2764
11. Castle JC, Kreiter S, Diekmann J et al (2012) Exploiting the mutanome for tumor vaccination. Cancer Res 72(5):1081–1091. https://doi.org/10.1158/0008-5472.CAN-11-3722

12. Kreiter S, Vormehr M, van de Roemer N et al (2015) Mutant MHC class II epitopes drive therapeutic immune responses to cancer. Nature 520(7549):692–696. https://doi.org/10.1038/nature14426
13. Sahin U, Derhovanessian E, Miller M et al (2017) Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature 547(7662):222–226. https://doi.org/10.1038/nature23003
14. Vormehr M, Schrörs B, Boegel S et al (2015) Mutanome engineered RNA immunotherapy: towards patient-centered tumor vaccination. J Immunol Res 2015:595363. https://doi.org/10.1155/2015/595363
15. Algarra I, García-Lora A, Cabrera T et al (2004) The selection of tumor variants with altered expression of classical and nonclassical MHC class I molecules: implications for tumor immune escape. Cancer Immunol Immunother 53(10):904–910. https://doi.org/10.1007/s00262-004-0517-9
16. Grandis JR, Falkner DM, Melhem MF et al (2000) Human leukocyte antigen class I allelic and haplotype loss in squamous cell carcinoma of the head and neck: clinical and immunogenetic consequences. Clin Cancer Res 6(7):2794–2802
17. Cock PJA, Antao T, Chang JT et al (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics 25(11):1422–1423. https://doi.org/10.1093/bioinformatics/btp163
18. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25. https://doi.org/10.1186/gb-2009-10-3-r25
19. Schmitz R, Young RM, Ceribelli M et al (2012) Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. Nature 490(7418):116–120. https://doi.org/10.1038/nature11378
20. Ferreira LMR, Meissner TB, Mikkelsen TS et al (2016) A distant trophoblast-specific enhancer controls HLA-G expression at the maternal-fetal interface. Proc Natl Acad Sci U S A 113(19):5364–5369. https://doi.org/10.1073/pnas.1602886113
21. Boegel S, Löwer M, Schäfer M et al (2012) HLA typing from RNA-Seq sequence reads. Genome Med 4(12):102. https://doi.org/10.1186/gm403
22. Boegel S, Scholtalbers J, Löwer M et al (2015) In silico HLA typing using standard RNA-Seq sequence reads. Methods Mol Biol 1310: 247–258. https://doi.org/10.1007/978-1-4939-2690-9_20
23. Boegel S, Löwer M, Bukur T et al (2014) A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. Oncoimmunology 3(8):e954893. https://doi.org/10.4161/21624011.2014.954893
24. Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5(7):621–628. https://doi.org/10.1038/nmeth.1226
25. Apps R, Murphy SP, Fernando R et al (2009) Human leucocyte antigen (HLA) expression of primary trophoblast cells and placental cell lines, determined using single antigen beads to characterize allotype specificities of anti-HLA antibodies. Immunology 127(1):26–39. https://doi.org/10.1111/j.1365-2567.2008.03019.x
26. Couzin-Frankel J (2013) Breakthrough of the year 2013. Cancer immunotherapy. Science 342(6165):1432–1433. https://doi.org/10.1126/science.342.6165.1432
27. Vormehr M, Diken M, Boegel S et al (2016) Mutanome directed cancer immunotherapy. Curr Opin Immunol 39:14–22. https://doi.org/10.1016/j.coi.2015.12.001
28. Bueno R, Stawiski EW, Goldstein LD et al (2016) Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations. Nat Genet 48(4):407–416. https://doi.org/10.1038/ng.3520
29. Bailey P, Chang DK, Forget M-A et al (2016) Exploiting the neoantigen landscape for immunotherapy of pancreatic ductal adenocarcinoma. Sci Rep 6:35848. https://doi.org/10.1038/srep35848
30. Scholtalbers J, Boegel S, Bukur T et al (2015) TCLP: an online cancer cell line catalogue integrating HLA type, predicted neo-epitopes, virus and gene expression. Genome Med 7:118. https://doi.org/10.1186/s13073-015-0240-5
31. Gloger A, Ritz D, Fugmann T et al (2016) Mass spectrometric analysis of the HLA class I peptidome of melanoma cell lines as a promising tool for the identification of putative tumor-associated HLA epitopes. Cancer Immunol Immunother 65(11):1377–1393. https://doi.org/10.1007/s00262-016-1897-3
32. Ritz D, Gloger A, Weide B et al (2016) High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs

and the identification of peptides from cell lines and patients' sera. Proteomics 16(10):1570–1580. https://doi.org/10.1002/pmic.201500445

33. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ et al (2015) Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. Mol Cell Proteomics 14(3):658–673. https://doi.org/10.1074/mcp.M114.042812
34. Carlini F, Ferreira V, Buhler S et al (2016) Association of HLA-A and non-classical HLA class I alleles. PLoS One 11(10):e0163570. https://doi.org/10.1371/journal.pone.0163570

Chapter 13

PHLAT: Inference of High-Resolution HLA Types from RNA and Whole Exome Sequencing

Yu Bai, David Wang, and Wen Fury

Abstract

Inferring HLA types from genome-wide sequencing data has gained growing attention with the development of new cost-efficient sequencing technologies and the increasing need to integrate HLA types with transcriptomic or other genomic information for insights into immune-mediated diseases, vaccination, and cancer immunotherapy. PHLAT is a computational tool designed for high-resolution (4-digit) typing of the major class I and class II HLA genes using RNAseq or exome sequencing data as input. We illustrate here how PHLAT can be installed, configured, and executed. This document also provides guidance for how to read and interpret the output results. Finally, the best practices of using PHLAT are also discussed.

Key words RNAseq, Whole exome sequencing, Transcriptome sequencing, Human leukocyte antigen, High-resolution HLA typing, Hematopoietic transplant, Organ transplant, HLA matching, Autoimmune, Immuno-oncology, Cancer vaccine, Bayesian inference

1 Introduction

Accurate HLA typing at four-digit resolution or higher not only is critical to the success of hematopoietic stem cell and organ transplant, but also facilitates the understanding and treatment of autoimmune disorders, infectious diseases, and cancer. Upon the advent of next-generation sequencing technologies, HLA typing rapidly evolved from Sanger sequencing to high-throughput amplicon-sequencing that can target the HLA loci [1–4]. In recent years, HLA typing using transcriptome or exome profiling data has attracted increasing attention as a result of the advances in next-generation sequencing technology and interest in correlating HLA types with genome-wide information to enrich data interpretation [5, 6]. However, compared to the targeted amplicon methods, transcriptome and exome/genome sequencing apply methods that use reduced read length and coverage such that resolving the highly homologous HLA alleles becomes more difficult.

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_13,

Table 1
Summary of HLA typing programs for genome-wide sequencing data

Program	Contig assembly	MHC class	Data type benchmarked	Year	DOI reference
seq2HLA	No	Class I, II	RNA-seq	2012	10.1186/gm403
HLAforest	No	Class I, II	RNA-seq	2013	10.1371/journal.pone.0067885
HLAminer	Yes	Class I, II	RNA-Seq, WES,WGS	2012	10.1186/gm396
ATHLATES	Yes	Class I, II	WES	2013	10.1093/nar/gkt481
PHLAT	No	Class I, II	RNA-Seq, WES	2014	10.1186/1471-2164-15-325
OptiType	No	Class I	RNA-Seq, WES,WGS	2014	10.1093/bioinformatics /btu548
PolySolver	No	Class I	WES	2015	10.1038/nbt.3344

Developing computational algorithms to identify HLA types from genome-wide data remains an active area of research.

Several recent algorithms are summarized in Table 1 [7–14]. The general strategy is to map the reads, or the contigs assembled from reads, against a collection of allele sequences, followed by an optimization to infer the HLA types based on the number, quality, and sequence consistency of the alignments. The use of assembled contigs can greatly reduce the ambiguity of the allele prediction from given sequence information compared to using single reads. Nonetheless, it is computationally costly and the resulting contigs may not be error-free especially for short (i.e., <100 bp) and/or single-ended reads. Thus, both contig assembly and direct read mapping have been actively explored. High prediction accuracy has been reported for multiple tools using either approach [7, 10–12]. The deduction of the genotypes is often solved as an optimization problem via various computational methods such as greedy algorithms (e.g., seq2HLA), integer linear programming (e.g., Optitype), and Bayesian inference (e.g., PHLAT, PolySolver). Linear programming and the Bayesian framework appear to be more effective given the improved performance of PHLAT, Optitype, and PolySolver over most of their predecessors [7, 11, 12]. However, one should keep in mind that the improvement is often the result of a collaborative effect with other factors; for instance, more accurate HLA read mapping and careful allele prioritization [7, 11]. PHLAT is capable of typing both class I and class II HLAs using transcriptomic or exomic data over a wide range of read length (37–250 bp) [*see* **Note 1**]. Class II HLA typing is not available in the current release of Optitype and PolySolver.

ALTHLATES and PolySolver are designed for exome sequencing. Their performances on transcriptome data and short reads (<100 bp) remain to be further studied.

PHLAT employs a Bayesian likelihood model to discover the most probable pair of HLA alleles given observed data at four-digit resolution or higher. The likelihood is defined by taking into account the sequence consistency at individual exomic single nucleotide polymorphism (SNP) sites, the phase consistency across adjacent SNPs, as well as the prevalence of HLA alleles in the population. Moreover, it implements the following novel strategy to facilitate accurate HLA read extraction and alignment. The whole human genome, together with the collection of genomic sequences of the HLA alleles, is used as the mapping reference. Because the reads are originated genome-wide, it is rational to search for the best alignment over the entire genome instead of just the HLA loci. Furthermore, a hierarchical candidate allele prioritization method effectively eliminates a large number of the false alleles. The prioritization also makes it tractable to evaluate all pair-wise combinations of remaining alleles for the best pair, in contrast to some methods that infer the two alleles sequentially and thus may miss the optimal choice when both are considered simultaneously. This framework is suitable to support HLA typing up to the full digit resolution. Currently, PHLAT focuses on the polymorphism within the coding sequence (CDS) regions and thereby the four- to six-digit typing [*see* **Notes 1, 2**].

PHLAT is freely available for non-commercial users. The text below describes in detail how to set up and execute PHLAT, and how to interpret its outputs. Suggestions for appropriately and effectively using PHLAT to obtain the best results are discussed as well.

2 Materials

2.1 Hardware and Software Environment

PHLAT is designed to run on any Linux environment (Ubuntu preferred) with Python 2.7. Several external dependencies are also required including Bowtie2 [15] (version 2.0.0-beta7 is recommended as this version was used when downstream algorithms in PHLAT were optimized), as well as the Python modules *psyam* (version 0.8.3, avoid 0.9+ version to ensure compatibility) and *Cython*. Other required modules, *cPickle*, *sets*, *copy*, *math*, *re*, *os*, *sys*, and *subprocess*, are typically included in Python2.7. Python modules can be installed using

```
> pip install module_name
```

An overview of PHLAT workflow is illustrated in Fig. 1.

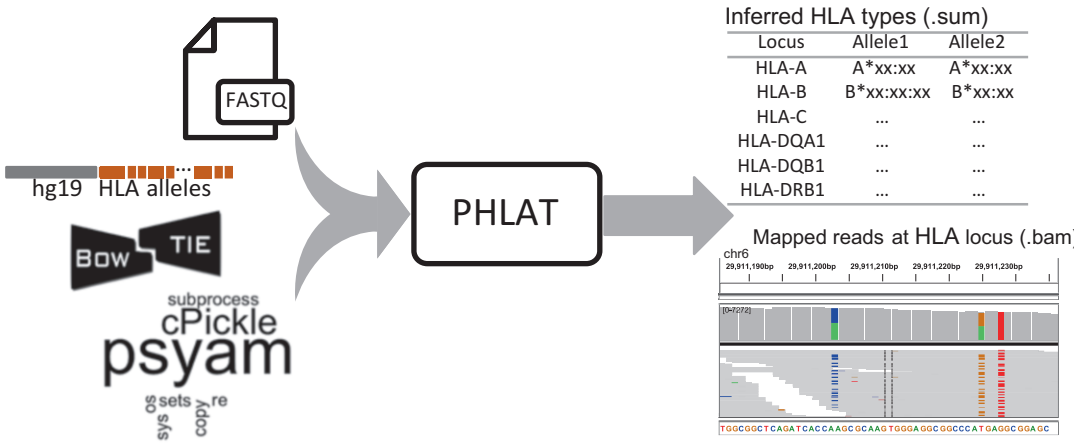


Fig. 1 Workflow of PHLAT. PHLAT takes input data in FASTQ format generated from RNAseq or whole exome sequencing, the reference genome that is composed of the whole human genome (hg19) and the collection of HLA alleles from IMGT. It also requires the installation of Bowtie2 and a series of indicated Python modules. The output from PHLAT is a text report of the most likely pairs of alleles at the major MHC class I and class II loci. Additionally, the user can choose to output the mapped reads in BAM format that support the genotype inference

2.2 Download and Installation

PHLAT has several dependencies that must be installed before use.

1. Python: PHLAT is compatible with Python 2.7. The latest version can be obtained from: https://www.python.org/downloads/.
2. Bowtie2: PHLAT uses Bowtie2 to map reads to the HLA reference. Version 2.0.0-beta7 is recommended and can be obtained from: https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.0.0-beta7/.
3. Download access for PHLAT can be requested at https://sites.google.com/site/phlatfortype. Unpack the tar file "phlat-release.tar.gz" into a root directory called "phlat-release." Create a sub-directory called "b2folder," and then unpack the contents of "b2folder.tar.gz" here.

2.3 Input Data

PHLAT accepts RNAseq, Whole Exome Seq, or Amplicon Seq data in fastq or gzipped fastq format sequenced using either single-end or pair-end reads. If single-ended, only one fastq file containing the reads is required. If pair-ended, PHLAT requires two fastq files, one for each set of reads.

Publicly available data in fastq format can be obtained from EBI. Search for a desired dataset by accession number/studyID and runID and download through either the direct link or "wget" with the ftp URL. An example of an acceptable dataset is accession number (or Study ID): ERP000101 and Run ID ERR009142.

The fastq file for each set of reads can be downloaded using the command below:

```
> wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR009/
ERR009142/ERR009142_1.fastq.gz
> wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR009/
ERR009142/ERR009142_2.fastq.gz
```

3 Methods

3.1 Inputs and Options

1. PHLAT can be run directly from the command line within the unzipped "phlat-release" folder:

   ```
   > python -O dist/PHLAT.py [inputs] [options]
   ```

   The "example.sh" file within the "phlat-release" folder demonstrates how to execute the program. This example can be executed using:

   ```
   > sh example.sh
   ```

2. The following inputs are required for execution:
   (a) **−1**: fastq file of the reads if single-end, or the first reads if paired-end.
   (b) **−2**: fastq file of the second reads if paired-end; ignore if single-end.
   (c) -**index**: url to the index files for Bowtie2 [default: b2folder subfolder in phlat-release directory].
   (d) -**b2url**:url to *Bowtie2* executable.
3. The options for the execution are:
   (a) -**orientation**: parameter to specify the relative orientation of the read mates as defined in Bowtie2 [default --fr]; this parameter is not used for single-end input.
   (b) -**tag**: prefix name label for the sample associated with the fastq files.
   (c) -**p**: number of threads for running Bowtie2 [default 8].
   (d) -**e**: url to the home folder of the "phlat-release" root directory.
   (e) -**o**: url to a directory where results shall be stored. A directory must be manually created before execution.
   (f) -**pe**: flag indicating whether the data shall be treated as paired-end (1) or single-end (0) [default 1].
   (g) -**tmp**: flag to either keep (1) or remove (0) the temporal file folder after the run [default 0].

3.2 Output

When running PHLAT, the expected messages are shown in Table 2. After execution, the HLA typing results are reported in a file with the ".sum" extension. Table 3 shows an example output

file produced by PHLAT upon successful analysis of the RNAseq data of subject NA11840 in run ERR009142 from study ERP000101. The fields in the output are explained below.

1. **Locus:** There are six HLA loci on human chromosome 6: HLA-A, HLA-B, HLA-C, HLA-DQA1, HLA-DQB1, and HLA-DRB1. Each of the loci has a pair of alleles that are defined in the columns "Allele1" and "Alllele2."
2. **Allele1:** One of the two alleles at the specified locus. The notation should be read as the HLA locus and digit resolution separated by a "*." PHLAT will always attempt to report 4-digit or higher typing as long as the resolution is permitted by the data. If the entries in "Allele1" and "Alllele2" are identical, the locus is homozygous. If they are different, the locus is heterozygous. In case more than one prediction are reported, this allele may be considered unresolved (or ambiguous) due to insufficient information in the data.
3. **Allele2:** The other allele at the specified locus.
4. **LLtot:** The log-likelihood score of the reported pair of alleles (Allele1 and Allele2). As described in previous work [7], it is the sum of the log-likelihood of the prediction supported by the data regarding individual SNPs and adjacent pairs of SNPs, as well as by the known prevalence of the alleles. The value measures the posterior probability of the reported pair of alleles given all available information.
5. **pval1:** The p-value of the allele is reported in "Allele1." It refers to the probability to observe an equal or higher number of reads mapped to that allele by chance. The value is comwputed as the one-tailed *z*-test *p*-value with respect to a Gaussian distribution whose mean and variance are approximated upon the read counts being mapped to all candidate alleles. In case

Table 2
PHLAT running messages and interpretation

Message	Meaning
Process Bowtie 2 mapping on example 9797 reads; of these: 9797 (100.00%) were paired; of these: 0 (0.00%) aligned concordantly 0 times 44 (0.45%) aligned concordantly exactly 1 time 9753 (99.55%) aligned concordantly >1 times 100.00% overall alignment rate	Bowtie2 execution succeeds
Prepare files of example for PHLAT Running PHLAT Done! Total PHLAT process time:0:01:08.573177	PHLAT execution succeeds

Table 3
The example output file of PHLAT

Locus	Allele1	Allele2	LLtot	pval1	pval2
HLA_A	A*02:01:01	A*02:01:01	−6140.19	4.30*E*−06	4.30*E*−06
HLA_B	B*27:05:03	B*57:01:01	−53014.24	1.50*E*−04	2.10*E*−02
HLA_C	C*02:02:02	C*06:02:01	−10939.56	1.60*E*−05	1.60*E*−02
HLA_DQA1	DQA1*02:01	DQA1*03:01:01	−3478.13	0	8.90*E*−03
HLA_DQB1	DQB1*03:02:01	DQB1*03:03:02	−565.28	1.50*E*−02	1.90*E*−04
HLA_DRB1	DRB1*04:04:01	DRB1*07:01:01	−15246.29	1.30*E*−04	2.90*E*−02

there is only one allele with reads mapped, the *p*-value is 0. Note that the PHLAT framework does not use *p*-value of individual alleles to determine the final HLA type. The *p*-value is nonetheless reported to help illustrate how significantly the most likely alleles surpass other alleles.

6. **pval2:** The *p*-value of the allele reported in "Allele2."

If the user chooses to keep the temporal file folder after PHLAT finishes, it can be found in the output folder defined by the option "-o," and its name is specified by the option "-tag" with a ".tmp" extension. The folder contains a BAM file that holds information of the reads mapped to the collection of HLA allele sequences. Note the coordinates of the reads are converted back to the corresponding genomic positions on chromosome 6. Users may visualize the file for details of the mapping and genotyping. An example of IGV visualization of the BAM file is shown in Fig. 2.

4 Notes

1. It is worth noting that although the framework of PHLAT is suitable for HLA typing with both RNA and DNA sequencing data, it currently utilizes the polymorphism solely within the coding sequence regions. Its performance with transcriptome data appears better than with exome sequencing (WES) data, as the WES protocol often captures partial intron regions flanking the exons such that reads originated therein may interfere with the mapping and subsequent analysis. In addition, sufficient read coverage in the data, as benchmarked in the previous publication [7], is critical to a high typing accuracy.

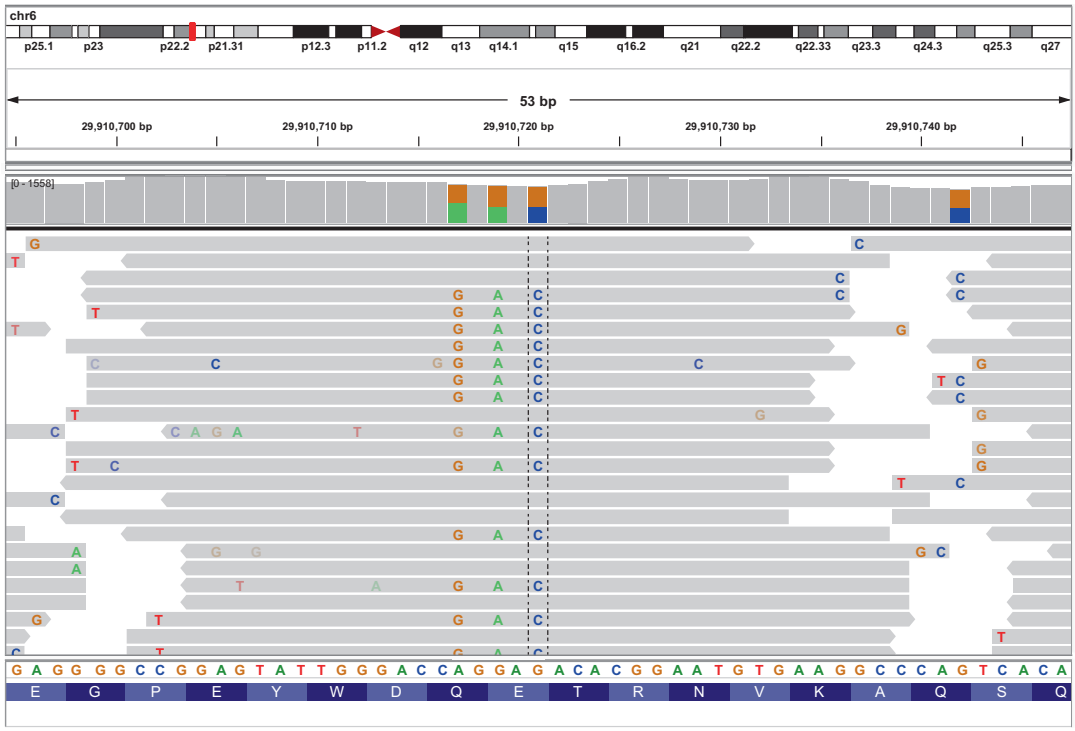


Fig. 2 The mapped reads that support the HLA typing displayed in the Integrative Genomics Viewer. The example is the output by PHLAT after analyzing the HapMap RNAseq data of the subject NA12761 (study ID:ERP000101, run ID: ERR009106). The viewer is zoomed into the genomic region chr6:29910700–29910750. The hg19 reference sequences of the HLA-A gene are shown at the bottom of the panel. The nucleotide bases A, C, G, and T are colored in green, blue, brown, and red, respectively. The bases in the reads, only if different from the reference sequence at the aligned positions, are highlighted and color coded. The pileup counts of bases at each position are shown in the top panel, from which the genotype at each position can be read out. For instance, the subject exhibits heterogeneous genotypes at locations chr6:29910716(A/G), chr6:29910718(A/G), chr6:29910720(G/C) and chr6:29910741(C/G)

2. When applied to MHC class II typing, special attention should be paid to HLA-DQA1*03:03, HLA-DQA1*05:05, and HLA-DQB1*02:02 alleles. A careful examination in our previous study suggested that for reads of 37–75 bp, these results are error-prone owing to possible misalignment of the reads from the minor HLA-DQA2 and DQB2 loci to their homologous major HLA-DQA1 and DQB1 loci, respectively. This is a limitation prevalent in most algorithms that provide class II HLA typing based on the IMGT database [7], because DQA2 and DQB2 loci are not included in IMGT due to limited knowledge. It is therefore difficult to recognize the actual origin of the reads and map them correctly. Until IMGT documentation is improved, we recommend that users use data with paired-end reads of 100 bp or longer to reduce the misalignment [7] and validate HLA-DQA1*03:03, HLA-

DQA1*05:05 and HLA-DQB1*02:02 alleles by targeted amplicon or Sanger sequencing, if they observe such predictions and are interested in their biological significance.

References

1. Danzer M, Niklas N, Stabentheiner S, Hofer K, Proll J, Stuckler C, Raml E, Polin H, Gabriel C (2013) Rapid, scalable and highly automated HLA genotyping using next-generation sequencing: a transition from research to diagnostics. BMC Genomics 14:221. https://doi.org/10.1186/1471-2164-14-221
2. Duke JL, Lind C, Mackiewicz K, Ferriola D, Papazoglou A, Gasiewski A, Heron S, Huynh A, McLaughlin L, Rogers M, Slavich L, Walker R, Monos DS (2016) Determining performance characteristics of an NGS-based HLA typing method for clinical applications. HLA 87(3):141–152. https://doi.org/10.1111/tan.12736
3. Erlich RL, Jia X, Anderson S, Banks E, Gao X, Carrington M, Gupta N, DePristo MA, Henn MR, Lennon NJ, de Bakker PI (2011) Next-generation sequencing for HLA typing of class I loci. BMC Genomics 12:42. https://doi.org/10.1186/1471-2164-12-42
4. Weimer ET, Montgomery M, Petraroia R, Crawford J, Schmitz JL (2016) Performance characteristics and validation of next-generation sequencing for human leucocyte antigen typing. J Mol Diagn 18(5):668–675. https://doi.org/10.1016/j.jmoldx.2016.03.009
5. Carapito R, Radosavljevic M, Bahram S (2016) Next-generation sequencing of the HLA locus: methods and impacts on HLA typing, population genetics and disease association studies. Hum Immunol 77(11):1016–1023. https://doi.org/10.1016/j.humimm.2016.04.002
6. Hosomichi K, Shiina T, Tajima A, Inoue I (2015) The impact of next-generation sequencing technologies on HLA research. J Hum Genet 60(11):665–673. https://doi.org/10.1038/jhg.2015.102
7. Bai Y, Ni M, Cooper B, Wei Y, Fury W (2014) Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. BMC Genomics 15:325. https://doi.org/10.1186/1471-2164-15-325
8. Boegel S, Lower M, Schafer M, Bukur T, de Graaf J, Boisguerin V, Tureci O, Diken M, Castle JC, Sahin U (2012) HLA typing from RNA-Seq sequence reads. Genome Med 4(12):102. https://doi.org/10.1186/gm403
9. Kim HJ, Pourmand N (2013) HLA typing from RNA-seq data using hierarchical read weighting. PLoS One 8(6):e67885. https://doi.org/10.1371/journal.pone.0067885
10. Liu C, Yang X, Duffy B, Mohanakumar T, Mitra RD, Zody MC, Pfeifer JD (2013) ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. Nucleic Acids Res 41(14):e142. https://doi.org/10.1093/nar/gkt481
11. Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, Stevens J, Lane WJ, Dellagatta JL, Steelman S, Sougnez C, Cibulskis K, Kiezun A, Hacohen N, Brusic V, Wu CJ, Getz G (2015) Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat Biotechnol 33(11):1152–1158. https://doi.org/10.1038/nbt.3344
12. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O (2014) OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics 30(23):3310–3316. https://doi.org/10.1093/bioinformatics/btu548
13. Wang C, Krishnakumar S, Wilhelmy J, Babrzadeh F, Stepanyan L, Su LF, Levinson D, Fernandez-Vina MA, Davis RW, Davis MM, Mindrinos M (2012) High-throughput, high-fidelity HLA genotyping with deep sequencing. Proc Natl Acad Sci U S A 109(22):8676–8681. https://doi.org/10.1073/pnas.1206614109
14. Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, Holt RA (2012) Derivation of HLA types from shotgun sequence datasets. Genome Med 4(12):95. https://doi.org/10.1186/gm396
15. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359. https://doi.org/10.1038/nmeth.1923

Chapter 14

Using Exome and Amplicon-Based Sequencing Data for High-Resolution HLA Typing with ATHLATES

Chang Liu and Xiao Yang

Abstract

ATHLATES (accurate typing of human leukocyte antigen through exome sequencing) was originally developed to analyze whole-exome sequencing (exome-seq) data from the Illumina platform and to predict the HLA genotype at 2-field or higher resolution. HLA locus-specific reads are first collected by stringent read mapping to the IMGT/HLA database. ATHLATES then performs read assembly, candidate allele identification, and genotype inference. Here, we describe the protocol of using ATHLATES for the above purpose and expand the application to analyze targeted sequencing data using amplicons of full HLA genes.

Key words Human leukocyte antigen (HLA), HLA typing, Whole-exome sequencing (exome-seq), Targeted sequencing, Polymerase chain reaction (PCR), Amplicons

1 Introduction

Since the initial success of using massively parallel pyrosequencing for high-resolution HLA typing [1], various next-generation sequencing (NGS) technologies have been rapidly adopted for the HLA typing of transplant recipients and donors [2]. The capability of NGS to phase densely packed sequence variants within HLA genes largely eliminated the problem of cis-trans ambiguity frequently encountered during Sanger sequencing. NGS also allows more extensive coverage of HLA genes, achieves higher throughput, and lowers the overall cost of HLA typing [3].

Most commercial or laboratory developed NGS assays for HLA typing utilize PCR to enrich target regions and take one of the following three approaches. First, shot-gun sequencing (short reads) of long-range amplicons of HLA genes on the Illumina or ion-torrent platforms [4, 5]. Second, single-molecule sequencing (long reads) of long-range amplicons on the PacBio or Nanopore platforms [6, 7]. Third, direct sequencing (short reads) of short-range amplicons of informative exons within target HLA genes on

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_14,

the Roche 454 or Illumina platforms [1, 8, 9]. These approaches, each with its pros and cons, offered many excellent options for clinical HLA typing and immunogenetic research.

In addition to amplicon-based methodologies, successful HLA typing can also be achieved by whole-genome sequencing, whole-exome sequencing (exome-seq), and RNA sequencing [3, 10–12]. We developed the ATHLATES program (Accurate Typing of HLA through Exome Sequencing) in 2012 with the consideration that exome-seq data should contain sufficient information for HLA typing. It was unclear at that time whether short reads from the Illumina platform were suitable for HLA typing, and few bioinformatic tools were available to accomplish such a task with high accuracy. ATHLATES takes locus-specific reads from the initial read mapping to all known HLA alleles in the IMGT/HLA database as input, and performs read assembly, candidate allele identification, and genotype inference (Fig. 1) [3]. The method was validated using high-coverage exome-seq datasets from the 1000 Genomes Project against Sanger sequencing of the corresponding DNA specimens. ATHLATES was subsequently used for HLA typing in landmark studies in cancer immunotherapy [13–15]. Since then, several excellent, open-access software packages have become available that implements various algorithms for HLA typing starting from different types of NGS data [10, 12, 16–19].

Here, we outline an abbreviated procedure for using ATHLATES to analyze standard exome-seq datasets generated on the Illumina platform to determine genotypes at HLA loci of your interest. We also demonstrate that the method is applicable to shot-gun sequencing data from long-range amplicons of selected HLA genes (Fig. 1).

2 Materials

2.1 Hardware

Stringent mapping of exome-seq reads from a large dataset requires significant resource. We requested 4 nodes, each with four cores at 2.4 GHz and 256 GB RAM per node, from a local IBM High Performance Computing system (chpc.wustl.edu) for this task, although this potentially could be scaled down. All processors run the latest version of the Redhat Linux operating system. Read mapping for amplicon-based sequencing data can be executed using one core. With the input bam file ready after the read mapping, ATHLATES can be run within a few minutes per locus with one core, or on a laptop computer (Intel® Core i7-6500 U CPU at 2.5 GHzx4, 15.6 GB RAM) running Ubuntu 16.04 LTS.

2.2 Software Dependencies

1. Bamtools [20] version 2.3.0 for processing of bam files: download bamtools-2.3.0.zip to your computer (https://github.com/pezmaster31/bamtools/archive/v2.3.0.zip). Unzip the

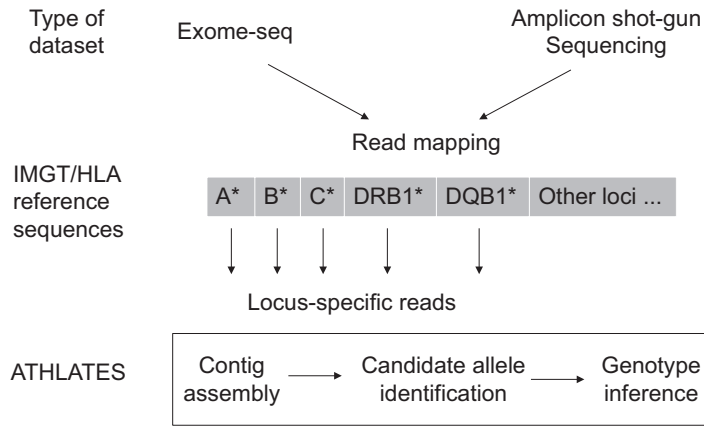


Fig. 1 Bioinformatic workflow

file, and enter the bamtools-2.3.0 folder. Create a "build" directory, move into the directory, compile the software using "cmake .." followed by "make."

2. Novoalign [21] version 3.03.00 for read mapping: this program was installed by the administrator of our computing cluster.
3. Samtools [22] version 1.3 for processing of sam and bam files: this program was installed by the administrator of our computing cluster. It can be installed on a Ubuntu laptop using "sudo apt-get install samtools."

2.3 Download and Compilation of ATHLATES

1. Download ATHLATES at the website of Broad Institute after filling out a registration form (https://www.broadinstitute.org/viral-genomics/viral-genomics-analysis-software-registration). Click on the "Download ATHLATES" link and save the compressed file named "athlates.zip." Unzip the file.
2. Enter the athlates folder, followed by "bash" and "export LD_LIBRARY_PATH = $LD_LIBRARY_PATH:/path/to/bamtools-2.3.0/lib."
3. Edit the makefile under athlates/src as follows using a text editor (e.g., gedit): "MYPATH=/path/to/bamtools-2.3.0/" and "COMPILER=/path/to/g++."
4. Go to the src directory and compile with "make."

2.4 Prepare Reference, Bed, and msa Files

1. Download all genomic and cDNA sequences of all HLA alleles from the IMGT/HLA database to a temporary folder "db-temp" using "wget ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/hla_gen.fasta" and "wget ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/hla_nuc.fasta". Combine the two files using "cat hla* > hla_all.fasta."

2. Download the script "hla_ref_clean.pl" from https://github.com/cliu32/athlates to the db-temp folder.
3. Remove redundant sequences and prepare the reference and bed files: run command "perl. /hla_ref_clean.pl -ref hla_all.fasta -oprefix hla_3.29.0." The command will report the total number of genes (39) and reference sequence files (20,148). The gene names will be listed in the end. Note that 3.29.0 is the current IMGT/HLA release version.
4. Transfer files to the correct folders using "cp hla_3.29.0.clean.fasta /path/to/athlates/db/ref" and "cp *.bed /path/to/athlates/db/bed."
5. Download msa (multi-sequence alignment) files for loci of you intend to type, example "A_nuc.txt" for HLA-A locus, from ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/alignments/. Rename the msa files by prefixing with the version number "hla_3.29.0." and transfer the files to "/path/to/athlates/db/msa." Note that the last block of C_nuc.txt contains an extended sequence at the 3′-end of C*04:09N, which must be deleted due to its interference with the function of ATHLATES. In addition, DRB_nuc.txt must be renamed to DRB1_nuc.txt to be compatible with the scripts mentioned in Subheading 2.5.

2.5 Download Related Scripts

1. Download "athlates_proto.sh" from https://github.com/cliu32/athlates to the "athlates" folder. This bash script will perform the initial read mapping and feed the sorted bam file to "runAthlates.pl."
2. Download "runAthlates.pl" from https://github.com/cliu32/athlates to the "athlates" folder. This is a wrapper that takes all HLA-specific reads in a sorted bam file to generate locus-specific reads and perform in silico HLA typing.

2.6 Sample Datasets

1. Exome-seq data: Download the paired fastq files for sample "HG01756" using "weget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR359/SRR359102/SRR359102_1.fastq.gz | weget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR359/SRR359102/SRR359102_2.fastq.gz" to your local data folder such as "/path/to/data_exome."
2. Amplicon shot-gun sequencing data: Download the paired fastq files for sample "HG01886" from https://github.com/cliu32/athlates to your local data folder such as "/path/to/data_amplicon." The data was generated using long-range amplicons of HLA-A, -B, -C, -DRB1, and -DQB1 genes. The amplicons from one sample were pooled for library preparation using the Nextera XT kit (Illumina) followed by sequencing.

3 Methods

3.1 Updating Paths and Folder Names in the "athlates_proto.sh" Script

1. Open the bash script in a text editor. Update the script by supplying the following full paths to replace the placeholders: (a) bamtools: /path/to/bamtools-2.3.0/lib; (b) athlates: /path/to/athlates; (c) data folder for sequence reads files: /path/to/data. A result folder will be created when running the script using the data folder name suffixed with "_rslt" to store intermediate files and final typing results.
2. Paired-end sequence reads in the data folder should be named as "<UniqueSampleName>_1.fastq.gz" and "<UniqueSampleName>_2.fastq.gz." Data from multiple samples can be stored in the data folder for batch analysis. Results generated in the result folder will be organized in subfolders named after the corresponding sample names. If the input files are fastq files, suffix1 and suffix2 in the bash script must be changed to "_1.fastq" and "_2.fastq," respectively.
3. Update the IMGT/HLA version if you prepared your reference, bed, and msa files using a version other than 3.29.0.
4. Modify the "#load required tools" section to allow access to novoalign and samtools (Subheading 2.2).

3.2 Selecting the Command Options for Read Mapping by Novoalign (V3.03.00)

1. For exome-seq datasets from 1000 Genomes Project, we used parameters "-r Random -i PE 100-1400 -H -t 30 -n 100." For amplicon-based sequencing using library prepared with Nextera XT, we used parameters "-r Random -i PE 100-1400 -H -t 30 -n 150."
2. Choose one of the two options, or it may be necessary to customize the parameters based on the fragment lengths (−i PE) and read length (−n) suitable for your library. [*see* **Note 1**]

3.3 Updating Paths in the "runAthlates.pl" Script

1. Open the perl script in a text editor. Identify the two lines of code as follows and supply the full paths to replace the placeholders: (a) my $samtool_cmd= "/path/to/samtools"; (b) my $athlates = "/path/to/athlates/bin/typing."

3.4 Executing the "athlates_proto.sh" Script

1. Save the updated bash script, and change the access permission with "chmod +x athaltes_proto.sh."
2. Execute the bash script by "./athlates_proto.sh." If you are working at a cluster, you may need to submit the job to requested computing cores for execution.
3. Once started, the program iterates over unique samples (sequence read file pairs) in the data folder in a for-loop. For each sample, novoalign maps sequence reads to the IMGT/HLA database to generate a sorted bam file. Next, the

"runAthlates.pl" script is executed to collect locus-specific reads, assemble contigs, identify possible and candidate alleles, and infer genotypes.

3.5 Interpreting the Output

1. The typing results for the two sample datasets are summarized in Table 1. The results are concordant with the consensus typing results at the 3-field resolution, except for the HLA-DQB1 locus of HG01756 where an additional allele DQB1*02:53Q could not be ruled out. The latter was discovered only recently with a "questionable" status, and it differs from DQB1*02:01:01 by a triple-base deletion at position 289–291 causing the loss of a lysine at position 65 (Fig. 2). The possible explanation for including DQB1*02:53Q is that the triplet deletion exists at the junction between two neighboring contigs, and there is zero mismatch between these contigs and DQB1*02:53Q.
2. A sample report for the HLA-A locus of HG01756 (exome-seq) is shown in Table 2. The report has three parts. *Part 1* reports a list of alleles supported, to various extents, by the assembled contigs. HD denotes the Hamming distance between each possible allele and its best hit in the contigs. When there is a difference between the alignment length (Aln_len) and cDNA length (cDNA_len), the gap indicates one or more exons that are (1) not captured and sequenced, or (2) excluded from calculation due to poor coverage or too many mismatches. For example, exon 8 of A*30:02:01:03 (first row) is absent in the contigs, as indicated by "(8, 5)." Perhaps this exon was not captured due to its small size, leaving a gap of five bases. For A*26:01:01:08 in the sixth row, the entire exon 2 (270 bp) was excluded from analysis in addition to the missing exon 8, leaving a gap of 275 bp. For alleles with partial reference sequences in the IMGT/HLA database, such as A*26:71N with reference sequence for exons 2–5 only, the exons with no reference sequence available are indicated as exon ID followed

Table 1
Typing results for sample datasets

Sample datasets (characteristics)	Alleles	HLA loci				
		A*	B*	C*	DRB1*	DQB1*
HG01756 (exome-seq)	Allele 1	30:02:01	18:01:01	05:01:01	03:01:01	02:01:01/02:53Q
	Allele 2	66:01:01	41:02:01	17:03:01	03:01:01	02:01:01/02:53Q
HG01886 (amplicon)	Allele 1	30:02:01	15:03:01	02:10:01	11:01:02	05:02:01
	Allele 2	74:01:01	57:03:01	07:01:02	13:02:01	06:09:01

```
cDNA                    110        120        130        140        150        160        170        180        190        200
DQB1*02:01:01             A GGATTTCGTG TACCAGTTTA AGGGCATGTG CTACTTCACC AACGGGACAG AGCGCGTGCG TCTTGTGAGC AGAAGCATCT ATAACCGAGA
DQB1*02:53Q               - ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------

cDNA                    210        220        230        240        250        260        270        280        290        300
DQB1*02:01:01       AGAGATCGTG CGCTTCGACA GCGACGTGGG GGAGTTCCGG GCGGTGACGC TGCTGGGGCT GCCTGCCGCC GAGTACTGGA ACAGCCAGAA GGACATCCTG
DQB1*02:53Q         ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- --------.. .---------

cDNA                    310        320        330        340        350        360        370
DQB1*02:01:01       GAGAGGAAAC GGGCGGCGGT GGACAGGGTG TGCAGACACA ACTACCAGTT GGAGCTCCGC ACGACCTTGC AGCGGCGAG
DQB1*02:53Q         ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------
```

Fig. 2 Comparison of DQB1*02:01:01 and DQB1*02:53Q. The alignment of exon 2 is generated using the online tool at IMGT/HLA (http://www.ebi.ac.uk/ipd/imgt/hla/align.html) and shown here. Three nucleotides at position 289-291 of DQB1*02:01:01 are deleted in DQB1*02:53Q, removing a lysine from the amino acid sequence position 65

by 0: "(1,0) … (6,0) (7,0) (8,0)." When the HD is above 0 and up to 2, each mismatch is reported with the exon ID and base position, such as "[3, 24]" and "[2, 209]" for A*30:10 and A*30:25, respectively. *Part 2* reports candidate alleles selected from part 1 to infer the final genotype. The candidate allele list excludes alleles with large exons (>25 bp) that find no hit in the contigs. For example, A*26:01:01:08 is not selected due to the no-hit exon 2. The candidate list also prioritizes alleles with an HD of zero. However, if all the 0-HD alleles encode the same protein sequence, additional alleles with HD of 1 are included in the candidate list. *Part 3* reports the final genotypes inferred from the candidate list. The genotypes must maximally represent the repertoire of sequences from the candidate alleles. Due to the lack of intron information from the exome-seq, ambiguities at the 4-field level cannot be eliminated. [*see* **Notes 2–4**]

4 Notes

1. Parameter optimization and aligner for read mapping. ATHLATES requires high-quality input reads with few or no mismatch. The "-t 30" option allows only one mismatch per read. For the read length option ("-n 100" or "-n 150"), although much longer reads can be generated with the current Illumina chemistry, it is important to trim the reads to exclude the part with a lower quality. We recommend that, when typing their own datasets, users optimize and validate the read mapping parameters using a few samples with known HLA genotypes to achieve the best results. In addition to novoalign, MOSAIK is also acceptable for read mapping in our hands [3].
2. Typing additional HLA genes. The program described in this protocol is capable of typing additional HLA genes such as DPB1, although this function has not been extensively validated. To do this, simply add DPB1 to the existing list of genes typed

Table 2
A sample ATHLATES output for the HLA-A locus of sample HG01756 (exome-seq)

Name	HD	Aln_len	cDNA_len	Similarity	Avg_cov	Missing_exons (ID,len); mismatches [ID,pos]
A*30:02:01:03	0	1093	1098	1	148.062	(8,5)
A*66:01:01:01	0	1093	1098	1	110.608	(8,5)
A*30:02:01:01	0	1093	1098	1	148.062	(8,5)
A*66:01:01:02	0	1093	1098	1	110.608	(8,5)
A*30:02:01:02	0	1093	1098	1	148.062	(8,5)
A*26:01:01:08	0	823	1098	1	132.245	(2270) (8,5)
A*26:01:01:09	0	823	1098	1	132.245	(2270) (8,5)
(skip 27 rows)						
A*30:04:01	0	817	1098	1	168.135	(3276) (8,5)
A*30:99	0	817	1098	1	168.135	(3276) (8,5)
A*26:71 N	0	669	939	1	158.179	(1,0) (2270) (6,0) (7,0) (8,0)
A*30:10	1	1093	1098	0.99909	148.062	(8,5) [3,24]
A*30:25	1	1093	1098	0.99909	148.062	(8,5) [2209]
(skip 48 rows)						
A*26:03:01	2	1093	1098	0.99817	110.608	(8,5) [2209] [2219]
A*30:03	2	1093	1098	0.99817	148.062	(8,5) [2165] [2170]
A*25:27:02	2	823	1098	0.99757	132.245	(2270) (8,5) [3273] [3275]
A*26:08	2	823	1098	0.99757	132.245	(2270) (8,5) [3195] [3196]
A*30:115	2	823	1093	0.99757	180.062	(2270) (8,0) [3183] [4,24]
(skip 8 rows)						
Candidate allelic pairs						
A*30:02:01:03	0	1093	1098	1	148.062	(8,5)
A*66:01:01:01	0	1093	1098	1	110.608	(8,5)
A*30:02:01:01	0	1093	1098	1	148.062	(8,5)
A*66:01:01:02	0	1093	1098	1	110.608	(8,5)
A*30:02:01:02	0	1093	1098	1	148.062	(8,5)
Inferred allelic pairs						
Allele1	*Allele2*			*HD*		
A*30:02:01:03	A*66:01:01:01			0		
A*30:02:01:03	A*66:01:01:02			0		

(continued)

Table 2
(continued)

Name	HD	Aln_len	cDNA_len	Similarity	Avg_cov	Missing_exons (ID,len); mismatches [ID,pos]
A*66:01:01:01	A*30:02:01:01			0		
A*66:01:01:01	A*30:02:01:02			0		
A*30:02:01:01	A*66:01:01:02			0		
A*66:01:01:02	A*30:02:01:02			0		

and listed in the "athlates_proto.sh" script, and add the msa file "hla_3.29.0.DPB1_nuc.txt" to the athlates/db/msa folder. We recommend that users validate the typing of additional genes using a few samples with known genotypes to achieve reliable results.

3. Limitation of ATHLATES. ATHLATES requires high-quality reads and stringent read mapping to achieve the optimal results. Although it performs well with high-coverage and low-error-rate datasets, it may not be as sensitive as programs like Optitype that has been shown to successfully type low-coverage datasets. ATHLATES also generates genotypes that maximally describe the information present in a dataset, not necessarily the most likely genotypes. Thus erroneous reads due to systematic sequencing errors, if present in large numbers and happen to match some rare alleles, may lead to the inclusion of these extra alleles in the candidate allele list and the final genotype. The average coverage (Avg_cov), although reported in the typing result, is not explicitly considered to identify the most likely genotype.
4. Pretest probability based on linkage disequilibrium and allele frequency. We recommend that, when reviewing the typing results, users take into consideration the common genetic associations for the HLA-B~HLA-C and HLA-DRB1~HLADQB1 haplotype blocks. If uncommon haplotypes or rare alleles outside of the common well-documented alleles are reported [23], possible typing error should be ruled out using additional information (e.g., coverage data) or through consultation with an HLA specialist.

References

1. Gabriel C, Danzer M, Hackl C, Kopal G, Hufnagl P, Hofer K, Polin H, Stabentheiner S, Proll J (2009) Rapid high-throughput human leukocyte antigen typing by massively parallel pyrosequencing for high-resolution allele identification. Hum Immunol 70(11):960–964. https://doi.org/10.1016/j.humimm.2009.08.009
2. Erlich HA (2015) HLA typing using next generation sequencing: An overview. Hum Immunol 76(12):887–890. https://doi.org/10.1016/j.humimm.2015.03.001
3. Liu C, Yang X, Duffy B, Mohanakumar T, Mitra RD, Zody MC, Pfeifer JD (2013) ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. Nucleic Acids Res 41(14):e142. https://doi.org/10.1093/nar/gkt481
4. Barone JC, Saito K, Beutner K, Campo M, Dong W, Goswami CP, Johnson ES, Wang ZX, Hsu S (2015) HLA-genotyping of clinical specimens using Ion Torrent-based NGS. Hum Immunol 76(12):903–909. https://doi.org/10.1016/j.humimm.2015.09.014
5. Duke JL, Lind C, Mackiewicz K, Ferriola D, Papazoglou A, Gasiewski A, Heron S, Huynh A, McLaughlin L, Rogers M, Slavich L, Walker R, Monos DS (2016) Determining performance characteristics of an NGS-based HLA typing method for clinical applications. HLA 87(3):141–152. https://doi.org/10.1111/tan.12736
6. Mayor NP, Robinson J, McWhinnie AJ, Ranade S, Eng K, Midwinter W, Bultitude WP, Chin CS, Bowman B, Marks P, Braund H, Madrigal JA, Latham K, Marsh SG (2015) HLA Typing for the Next Generation. PLoS One 10(5):e0127153. https://doi.org/10.1371/journal.pone.0127153
7. Ammar R, Paton TA, Torti D, Shlien A, Bader GD (2015) Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. F1000Research 4:17. https://doi.org/10.12688/f1000research.6037.1
8. Lange V, Bohme I, Hofmann J, Lang K, Sauter J, Schone B, Paul P, Albrecht V, Andreas JM, Baier DM, Nething J, Ehninger U, Schwarzelt C, Pingel J, Ehninger G, Schmidt AH (2014) Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics 15:63. https://doi.org/10.1186/1471-2164-15-63
9. Schofl G, Lang K, Quenzel P, Bohme I, Sauter J, Hofmann JA, Pingel J, Schmidt AH, Lange V (2017) 2.7 million samples genotyped for HLA by next generation sequencing: lessons learned. BMC Genomics 18(1):161. https://doi.org/10.1186/s12864-017-3575-z
10. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O (2014) OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics (Oxford) 30(23):3310–3316. https://doi.org/10.1093/bioinformatics/btu548
11. Xie C, Yeo ZX, Wong M, Piper J, Long T, Kirkness EF, Biggs WH, Bloom K, Spellman S, Vierra-Green C, Brady C, Scheuermann RH, Telenti A, Howard S, Brewerton S, Turpaz Y, Venter JC (2017) Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. Proc Natl Acad Sci U S A. https://doi.org/10.1073/pnas.1707945114
12. Boegel S, Lower M, Schafer M, Bukur T, de Graaf J, Boisguerin V, Tureci O, Diken M, Castle JC, Sahin U (2012) HLA typing from RNA-Seq sequence reads. Genome Med 4(12):102. https://doi.org/10.1186/gm403
13. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, Miller ML, Rekhtman N, Moreira AL, Ibrahim F, Bruggeman C, Gasmi B, Zappasodi R, Maeda Y, Sander C, Garon EB, Merghoub T, Wolchok JD, Schumacher TN, Chan TA (2015) Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 348(6230):124–128. https://doi.org/10.1126/science.aaa1348
14. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, Berent-Maoz B, Pang J, Chmielowski B, Cherry G, Seja E, Lomeli S, Kong X, Kelley MC, Sosman JA, Johnson DB, Ribas A, Lo RS (2016) Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. Cell 165(1):35–44. https://doi.org/10.1016/j.cell.2016.02.065
15. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, Walsh LA, Postow MA, Wong P, Ho TS, Hollmann TJ, Bruggeman C, Kannan K, Li Y, Elipenahli C, Liu C, Harbison CT, Wang L, Ribas A, Wolchok JD, Chan TA (2014) Genetic basis for clinical response to CTLA-4 blockade in melanoma. N Engl J Med 371(23):2189–2199. https://doi.org/10.1056/NEJMoa1406498
16. Boegel S, Scholtalbers J, Lower M, Sahin U, Castle JC (2015) In silico HLA typing using standard RNA-Seq sequence reads. Methods

Mol Biol 1310:247–258. https://doi.org/10.1007/978-1-4939-2690-9_20

17. Ka S, Lee S, Hong J, Cho Y, Sung J, Kim HN, Kim HL, Jung J (2017) HLAscan: genotyping of the HLA region using next-generation sequencing data. BMC Bioinformatics 18(1):258. https://doi.org/10.1186/s12859-017-1671-3
18. Huang Y, Yang J, Ying D, Zhang Y, Shotelersuk V, Hirankarn N, Sham PC, Lau YL, Yang W (2015) HLAreporter: a tool for HLA typing from next generation sequencing data. Genome Med 7(1):25. https://doi.org/10.1186/s13073-015-0145-3
19. Bai Y, Ni M, Cooper B, Wei Y, Fury W (2014) Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. BMC Genomics 15:325. https://doi.org/10.1186/1471-2164-15-325
20. Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT (2011) BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics (Oxford) 27(12):1691–1692. https://doi.org/10.1093/bioinformatics/btr174
21. Hatem A, Bozdag D, Toland AE, Catalyurek UV (2013) Benchmarking short sequence mapping tools. BMC Bioinformatics 14:184. https://doi.org/10.1186/1471-2105-14-184
22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352
23. Osoegawa K, Mack SJ, Udell J, Noonan DA, Ozanne S, Trachtenberg E, Prestegaard M (2016) HLA haplotype validator for quality assessments of HLA typing. Hum Immunol 77(3):273–282. https://doi.org/10.1016/j.humimm.2015.10.018

Chapter 15

HLA Typing from Short-Read Sequencing Data with OptiType

András Szolek

Abstract

The established standards for HLA genotyping rely on targeted DNA sequencing techniques. However, the increasing abundance of short-read sequencing data has prompted a demand for computational tools capable of HLA typing from general purpose sequencing data as well. OptiType is a software that performs HLA typing from short-read DNA and RNA sequencing data, and this chapter guides the user through its installation and usage.

Key words HLA class I typing, RNA-Seq, Whole-exome Seq, OptiType, Mapping, In silico, Bioinformatics

1 Introduction

HLA typing from next-generation sequencing (NGS) data is challenging, and the traditional genotyping pipeline (i.e., read mapping to a single reference genome followed by variant calling) is hopelessly inadequate for HLA genes, primarily due to their highly polymorphic nature. A number of purpose-built tools have been developed to tackle the problem using various algorithmic approaches, with different strengths and weaknesses according to independent benchmarks [1, 2].

OptiType performs HLA typing using a combinatorial optimization approach [3]. Reads are mapped to a reference panel consisting of HLA Class I allele sequences centered around their most polymorphic, and functionally most important region, exons 2 and 3. Reads are allowed to map to any number of alleles as long as they are co-optimal alignments (i.e., contain no more mismatches than any other possible alignment) and have at least 97% identity with the subject sequence.

From these alignments a binary hit matrix is constructed, whose rows and columns correspond to reads and alleles respectively. The binary value of a given cell reflects whether the read

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_15,

mapped to the allele co-optimally, in other words, whether the read can be explained by the allele's presence in the sample or not. Based on the assumption that the correct genotype explains more reads than any other genotype, OptiType uses an integer linear program (ILP) model to select up to two alleles per locus, such that they explain the largest possible number of reads together.

OptiType takes fastqfiles as input, and reports the predicted 4-digit HLA Class I genotype in a tab-separated file and an accompanying PDF document, containing detailed coverage plots, allowing a critical visual inspection of the results. Below, a step-by-step guide is provided to the installation and usage of OptiType, and guidelines to interpreting its output.

2 Materials

2.1 Software Dependencies

OptiType is a Linux software written in Python. Its external requirements are the read mapper RazerS3 [4] or Yara [5], an ILP solver such as CBC or CPLEX, the HDF5 library, and the Python packages NumPy [6], Pyomo [7], Pandas [8], Pysam [9], and Matplotlib [10]. As the installation process is slightly involved, some users may prefer the self-contained Docker image of OptiType including all dependencies, hosted on Docker Hub.

2.2 Installation from Source

1. OptiType is compatible with both Python 2 and 3 from versions 2.7.9 and 3.5 or above. Ensure a supported version is available on your system.
2. Download and install the read mapper RazerS3 by following the instructions on https://github.com/seqan/seqan/tree/master/apps/razers3.
3. Install an integer programming solver software. Any solver supported by the Python package Pyomo should be supported by OptiType as well, but we suggest either the open-source CBC (https://projects.coin-or.org/Cbc) or CPLEX, which is a commercial software with a free academic license (http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/).
4. Verify that the solver is globally accessible in the computing environment. If executing `cbc` or `cplex` from the command line does not start them, include their install destination in the `PATH` environment variable.
5. On most command-line interfaces this can be done with

   ```
   export PATH=/path/to/ilp/cbc:$PATH
   ```

   To make it permanent, include this line in your shell's startup script (like `.bashrc` for bash).

6. Download and extract the current HDF5 shared library from https://support.hdfgroup.org/ftp/HDF5/current18/bin/linux-centos7-x86_64-gcc485/.
7. Include the extracted `lib` folder in the `LD_LIBRARY_PATH` environment variable.
8. Create a temporary environment variable `HDF5_DIR` containing the path to the HDF5 base folder with

```
export HDF5_DIR=/path/to/hdf5-1.8.xx-linux-centos7-x86_64-gcc485-shared/
```

9. Install the Python packages numpy, pyomo, pysam, matplotlib, tables, pandas and future with Python's package manager pip by running `pip install <package-name>` for all of them. While not necessary, it is good practice to create an isolated Python environment for the above packages instead of installing them system-wide. If using Python 2, consider installing the package virtualenv before the above, and consult https://virtualenv.pypa.io/en/stable/userguide/ how to create and active such an environment. With Python 3, use the built-in `venv` module to achieve the same (https://docs.python.org/3.6/library/venv.html).
10. Fetch OptiType from https://github.com/FRED-2/OptiType. Either download and extract the ZIP file or use git to clone the repository.
11. Create a `config.ini` file in the OptiType directory by adapting `config.ini.example`.
12. Provide the full path to the RazerS3 executable, and set the number of threads to be used.
13. Specify the solver that has been installed, such as `cbc` or `cplex`. Unlike in the previous step, this should not be a path, just a simple keyword.

2.3 Installing the Docker Version

1. Ensure Docker (https://www.docker.com/) is installed on the system.
2. Download the Docker image of OptiType with the command docker pull fred2/optitype.

3 Methods

3.1 Running OptiType (Python Version)

1. If the Python dependencies had been installed in a virtual environment, activate the environment first.
2. Assuming the base directory of OptiType is `/path/to/OptiType`, it can be launched with

```
python /path/to/OptiType/OptiTypePipeline.py -i <reads1> [<reads2>] (-d|-r) -o <outdir> [-p <prefix>] [-v] [-b <beta>] [-e <n>] [-c <configfile>]
```

3. `reads1` and `reads2` are the input fastq or fastq.gz files. If `reads2` is provided, the data will be treated as paired-end (*see* **Note 1**).
4. The mutually exclusive `-d` and `-r` flags tell OptiType whether the input is DNA or RNA sequencing data.
5. `outdir` is the output directory. OptiType will create this directory and write the results into a timestamped folder inside. If outdir exists (for example due to a previous run with different settings) the timestamp will make sure that the previous results are not overwritten. If the optional -p prefix argument is provided, the output files will be written directly into outdir, with filenames starting with that prefix. This is useful when processing an entire batch of samples.
6. The flag `-v` enables verbose mode with diagnostic information printed on the standard output.
7. `beta` is an optional floating point parameter that controls OptiType's propensity to call homozygous genotypes over heterozygous. The default value of 0.009 has been validated to be optimal in most circumstances.
8. The `-e <n>` option will report n−1 suboptimal solutions alongside the optimal one. This is strictly a debugging feature and the suboptimal solutions shall not be taken at face value.
9. The optional `configfile` allows the user to override `config.ini` in OptiType's base directory. The `[behavior]` section of the default configuration file contains several further settings with accompanying explanations of their effect.
10. OptiType is shipped with two small test datasets found in the `test/exome` and `test/rna` subfolders of the installation directory. It is advised to run OptiType on these test samples first, to ensure that everything is working correctly. Enter the installation directory and run

```
python OptiTypePipeline.py -i test/exome/NA11995_
SRR766010_1_fished.fastq test/exome/NA11995_
SRR766010_2_fished.fastq -d -v -o /path/to/outdir
```

See Subheading 3.4 regarding the content of the output files.

3.2 Running OptiType (Docker Version)

1. OptiType can be run with

```
docker run -v /path/to/my_data:/data/ -t fred2/op-
titype -i <reads1> [<reads2>] (-d|-r) -o <outdir>
[-v] [-b <beta>] [-e <n>] [-c <configfile>]
```

2. `/path/to/my_data` is a directory on the host system that will be mounted at `/data/` inside the Docker container.
3. Regarding OptiType's command line options, *see* **steps 3** and onward of Subheading 3.1.

4. The path of the input files and output directory need to be referred relative to the Docker container's `/data/` directory (which is the default working directory inside the Docker container). If the input data is located at `/path/to/my_data/sample345/1.fastq` on the host system, the `reads1` argument can either be `sample345/1.fastq` or `/data/sample345/1.fastq` but not `/path/to/my_data/sample345/1.fastq`. The same applies to `outdir` and `configfile`.
5. If the desired output directory on the host system is outside `/path/to/my_data`, one can mount it with an additional `-v /path/to/desired_out/:/out/` option and set the `outdir` argument to `/out`.
6. To check whether everything is set up correctly, download the two test fastq files from https://github.com/FRED-2/OptiType/tree/master/test/exome into `/path/to/my_data` and execute

```
docker run -v /path/to/my_data:/data/ -t fred2/optitype -i NA11995_SRR766010_1_fished.fastq NA11995_SRR766010_2_fished.fastq -d -v -o outdir
```

   The results should soon appear in `/path/to/my_data/outdir`.

3.3 Read Mapping with Alternative Read Alignment Software

The read mapping strategy of RazerS3 is unusual in that it involves indexing reads instead of the reference, therefore its memory consumption is proportional to the input size. While this is rarely an issue on the high-memory machines OptiType is typically ran on, if a user is restricted to a desktop computer, they may seek a more memory-efficient alternative solution. By filtering for HLA reads before passing them to OptiType, the size of the input fastq files can be dramatically reduced. This can be achieved by mapping reads against OptiType's custom reference sequence panel, and channeling the captured reads into smaller fastq files. Any established read mapper should be suitable for the purpose. However, if we can guarantee that the read mapper is fully sensitive, exact, and capable of finding all co-optimal alignments, OptiType can take the resulting BAM files as input directly, bypassing RazerS3 in the alignment process altogether. Yara [5] is a read mapper that fulfills these criteria and will be used for an example.

1. Download and install Yara from https://github.com/seqan/seqan/tree/master/apps/yara.
2. Download and install SAMtools [9] from http://www.htslib.org/.
3. Index OptiType's custom reference panel with

```
yara_indexer -o /path/to/OptiType/data/yara_dna /path/to/OptiType/data/hla_reference_dna.fasta
```

   For the RNA reference obviously replace `dna` with `rna`.

4. Map reads against the reference with

```
yara_mapper -e 3 -f bam -u /path/to/OptiType/data/
yara_dna <reads1> | samtools view -h -F 4 -b1 - >
mapped_1.bam
```

5. Repeat step 4 for the second end replacing `<reads1>` and `mapped_1.bam` with `<reads2>` and `mapped_2.bam`
6. Run OptiType with `mapped_1.bam` and `mapped_2.bam` as input instead of the usual fastq files (*see* Subheadings 3.1 or 3.2).

3.4 Interpreting the Results

A successful OptiType run generates two files in the output directory: a `tsv` and a `pdf` file.

3.4.1 Interpreting the tsv File

The `tsv` file contains eight columns:

1. `A1, A2, B1, B2, C1, C2` contain the predicted 4-digit HLA genotype. Homozygous loci have two identical entries in both columns. Note: 1 and 2 are not meant to reflect large-scale maternal or paternal haplotypes.
2. `Reads` contains the total number of reads that the predicted HLA genotype could explain.
3. `Objective` is a closely related value to `Reads` that entails the zygosity regularization factor set by the beta parameter in the ILP model. It is of little interest to the typical user.
4. If OptiType was called with the setting `-e <n>` for suboptimal solution enumeration, the table should contain n rows instead of just one, in decreasing order of optimality.

3.4.2 Interpreting the pdf File

The pdf file contains the coverage plots of the predicted alleles in a 3 × 2 arrangement, each row standing for a locus. Homozygous loci have only one coverage plot in their row.

The *x*-axis of each plot shows position in sequence, starting from the beginning of intron 1 and ending at the end of intron 3 for DNA data, and starting at the beginning of exon 2 and ending at the end of exon 3 for RNA data. The gray bands in the background show the location of exons 2 and 3. The logarithmically scaled *y*-axis shows coverage depth at the corresponding position, which is presented as a stacked area chart consisting of up to eight different bands.

The chart uses four different colors, each in a vivid, saturated, and a lighter, less saturated tone. The saturated bands consist of reads that aligned to the allele better than every other allele in the predicted genotype, whereas light color bands represent reads that aligned to multiple alleles (usually two alleles of the same locus) equally well. The meaning of the four colors is as follows:

1. The green area consists of paired-end reads where both ends aligned to the allele sequence without any mismatches.

2. The yellow area consists of reads where only one end could be mapped, but it did so without mismatches. This is typically present toward the beginning and the end of the plot due to the other read end falling outside the inspected region.
3. The red area shows paired-end reads where both ends aligned to the allele sequence but did so with at least one mismatch. The most common reason for mismatches is sequencing errors. Usually, this manifests as a thin red band on top of the larger green hills.
4. Finally, blue stands for single-end alignments with mismatches.

3.4.3 Example

Let us take a closer look at the coverage plots of the test exome dataset, as seen in Fig. 1.

1. The HLA-A locus is homozygous, B and C are heterozygous.
2. Exome enrichment is apparent: coverage is high on both exons on all loci, and drops at the exon boundaries. Curiously the end of intron 3 is highly covered as well.
3. Most read pairs could be aligned without mismatches, and importantly, they cover the entirety of the exons for every

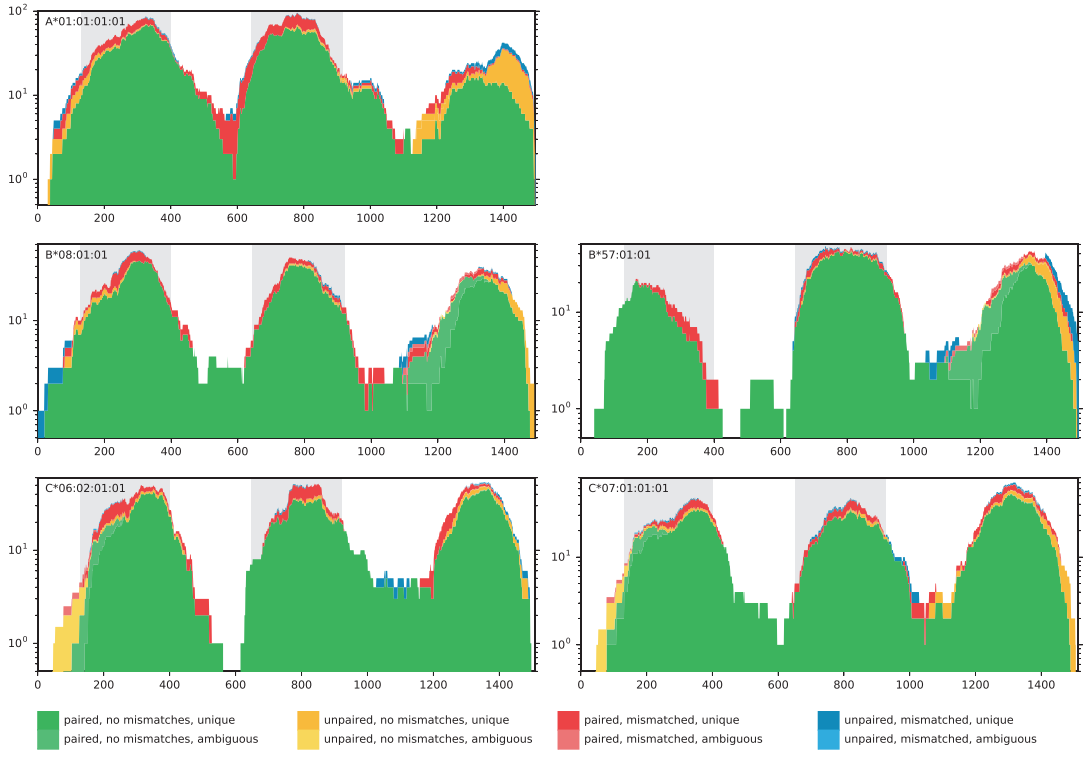


Fig. 1 Coverage plots of the HLA alleles predicted in the 1000 Genomes exome sequencing dataset NA11995/SRR766010

allele. The plots appear conclusive and leave no doubt about the presence of the predicted alleles in the sample.

4. C*06:02 and C*07:01 share a light green/yellow band at the beginning of exon 2. Due to the fact that these two alleles have perfect sequence identity spanning 245 bases around the intron 1–exon 2 junction, read pairs falling entirely in that window cannot be assigned to either allele uniquely, and are hence presented with a light tone. The same can be observed between B*08:01 and B*57:01 in the middle of intron 3.
5. Toward the end of intron 2 of A*01:01 one can observe a dip in perfectly matching (green) reads in favor of mismatched (red) reads. Although coverage is low at that position, it may hint at a potential variant in one of the two A*01:01 alleles, which further manual analysis could verify or refute.

4 Notes

1. If the original FASTQ files are not available, but a BAM file with the reads aligned to a reference genome is, the user can still extract the necessary reads to use OptiType. The genomic location of the HLA superlocus is on chromosome 6, between megabase coordinates 29.5 and 33.1. However, extracting reads mapped to this region may not be sufficient, due to the imperfect sensitivity of some read mappers when aligning HLA reads that originate from alleles dissimilar to the reference sequence [11]. Therefore, if possible, unmapped reads should be extracted from the BAM file as well. The below command extracts HLA region and unmapped reads from a paired-end, sorted and indexed `in.bam` and outputs the individual read ends into `out_1.fastq` and `out_2.fastq`:

```
samtools view -h in.bam -F 256 chr6:29,500,000-33,200,000 "*" | samtools bam2fq -1 out_1.fq -2 out_2.fq -
```

   It is important to keep in mind that the command may require customization for certain reference genome builds. If the reference genome contains alternate chromosome 6 contigs, or additional contigs for a set of HLA alleles, reads should be extracted from those regions as well, since their purpose is to capture HLA reads. The presence of alternate contigs can be determined by inspecting the BAM file's header with

```
samtools view -H in.bam
```

   and looking for identifiers such as `chr6_GL000250v2_alt` or `HLA-A*01:01`. If they are present, all such identifiers

should be placed at the end of the `samtools view` subcommand in a space-separated fashion, without any coordinate qualifiers. Once reads have been extracted, OptiType can be run with `out_1.fq` and `out_2.fq` as the input.

References

1. Bauer DC, Zadoorian A, Wilson LOW et al (2016) Evaluation of computational programs to predict 5HLA6 genotypes from genomic sequencing data. Brief Bioinform pii:bbw097. https://doi.org/10.1093/bib/bbw097
2. Kiyotani K, Mai TH, Nakamura Y (2016) Comparison of exome-based 5HLA6 class I genotyping tools: identification of platform-specific genotyping errors. J Human Genet 62(3):397–405. https://doi.org/10.1038/jhg.2016.141
3. Szolek A, Schubert B, Mohr C et al (2014) OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics 30(23):3310–3316. https://doi.org/10.1093/bioinformatics/btu548
4. Weese D, Holtgrewe M, Reinert K (2012) 5RazerS6 3: faster, fully sensitive read mapping. Bioinformatics 28(20):2592–2599. https://doi.org/10.1093/bioinformatics/bts505
5. Siragusa E, Weese D, Reinert K (2013) Fast and accurate read mapping with approximate seeds and multiple backtracking. Nucleic Acids Res 41(7):e78–e78. https://doi.org/10.1093/nar/gkt005
6. van der Walt S, Colbert SC, Varoquaux G (2011) The 5NumPy6 array: a structure for efficient numerical computation. Comput Sci Eng 13(2):22–30. https://doi.org/10.1109/mcse.2011.37
7. Hart WE, Watson J-P, Woodruff DL (2011) Pyomo: modeling and solving mathematical programs in Python. Math Progr Comput 3(3):219–260. https://doi.org/10.1007/s12532-011-0026-8
8. McKinney W (2010) Data structures for statistical computing in python. In: van der Walt S, Millman J (eds) Proceedings of the 9th python in science conference. Creative Commons, Austin, TX, pp 51–56
9. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and 5SAMtools6. Bioinformatics 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352
10. Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9(3):90–95. https://doi.org/10.1109/mcse.2007.55
11. Brandt DYC, Aguiar VRC, Bitarello BD et al (2015) Mapping bias overestimates reference allele frequencies at the 5HLA6 genes in the 1000 Genomes Project phase I data. G3 Genet 5(5):931–941. https://doi.org/10.1534/g3.114.015784

Chapter 16

Comprehensive HLA Typing from a Current Allele Database Using Next-Generation Sequencing Data

Shuji Kawaguchi, Koichiro Higasa, Ryo Yamada, and Fumihiko Matsuda

Abstract

HLA allele information is essential for a variety of medical applications, such as genomic studies of multifactorial diseases, including immune system and inflammation-related disorders, and donor selection in organ transplantation and regenerative medicine. To obtain this information, an accurate *HLA* typing method that is applicable for any allele registered in *HLA* allele databases is needed. Here, we describe a method for determining alleles from a current *HLA* database using next-generation sequencing (NGS) results.

Key words *HLA* typing, NGS, Software, Database, Bioinformatics

1 Introduction

The accurate genotyping of HLA alleles is essential in developing novel therapeutic strategies for many diseases and the development of therapeutic strategies. Next-generation sequencing (NGS) technologies have enabled dramatic breakthroughs in *HLA* typing in terms of high-throughput read generation derived from a haploid genome [1–3]. This advance has made it possible to determine the complete sequence of any *HLA* allele, which is difficult using conventional PCR-based methods, such as the PCR-sequence-specific oligonucleotide probes (PCR-SSOP) represented by Luminex® and sequencing-based typing (PCR-SBT) [4, 5].

Nevertheless, complete and an accurate *HLA* typing from NGS data is still challenging, because of the high polymorphism of the HLA genes. The IPD-IMGT/HLA, the world's database of *HLA* alleles [6], has identified more than seventeen thousand HLA alleles to date (July, 2017), the number of which doubled in just five years. Unfortunately, however, a large number of the *HLA* alleles are registered as only a partial nucleotide sequence consisting of one or two exons encoding the functionally critical G-DOMAIN [6]. It is impossible to identify a specific *HLA* allele

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_16,

without considering sequences outside the G-DOMAIN, because multiple alleles can share the same G-DOMAIN sequence.

Several *HLA* typing methods aimed at enhancing the accuracy of allelic determination have been implemented and published. These methods can be classified broadly into two categories, according to whether the applicable alleles are restricted or not (Table 1). HLAforest [7] and PHLAT [8] are assigned to the unrestricted group because they use information about all of the exons for *HLA* typing. On the other hand, HLAreporter [9], OptiType [10], PCR-SBT, and PCR-SSOP are assigned to the restricted group, because the exons, primers, or probes used are limited. OptiType eliminates rare alleles and imputes unknown intron sequences based on phylogenetic information about the other alleles. To improve accuracy, HLAreporter uses conservative parameter settings so that the method does not give alleles unless the coverage of the reads is sufficient [9]. The true positive rate of major alleles is improved by these decision schemes in exchange for the ability to correctly type minor alleles.

We have developed a new method, called HLA-HD (*HLA* typing from a High-quality Dictionary) [11], which types *HLA* alleles with 6-digit precision from NGS data. The method creates an *HLA* allele dictionary from the current allele information to increase the completeness of applicable alleles, and types the allele

Table 1
Existing *HLA* typing algorithms and methods

Method	Technology	Applicable *HLA* genes	Applicable alleles	Typing resolution
HLA-HD	NGS	All *HLA* genes	All alleles registered in the IPD-IMGT/HLA	Six digits
HLAforest	NGS	All *HLA* genes	All alleles registered in the IPD-IMGT/HLA	Up to eight digits
PHLAT	NGS	*A, B, C, DQA1, DQB1, DRB1*	All alleles registered in the IPD-IMGT/HLA	Six digits
HLAreporter	NGS	*A, B, C, DPB1, DQA1, DQB1, DRB1,3,4,5*	Limited to alleles that can be distinguished by the exons used	Six digits
OptiType	NGS	*A, B, C*	Limited to alleles that can be distinguished by the exons used	Four digits
PCR-SBT	PCR	Preset genes	Limited to alleles that can be distinguished by the exons used	Four digits in most cases
PCR-SSOP	PCR	Preset genes	Preset alleles	Four digits in most cases

with high accuracy using an unbiased comparison algorithm. Therefore, HLA-HD belongs in the "not restricted" category, because it uses all of the *HLA* genes and alleles recorded in the IPD-IMGT/HLA database (Table 1). In this chapter, we describe how to use HLA-HD with the allele data of the IPD-IMGT/HLA database and demonstrate an actual *HLA* typing from whole exome data (WES).

2 Materials

HLA-HD is freely available for academic use and research purposes upon registration and requires additional mapping software, allele data, and NGS sequence data. Before typing, you must construct a database of allele sequences separated into exon or intron units, called the "HLA allele dictionary," from information in the IPD-IMGT/HLA database [6].

2.1 Hardware

HLA-HD was coded in basic C++ language, so the software can work on almost any operating system, including Linux, Mac OS, or Windows.

2.2 Software Required

HLA-HD requires bowtie2 [12] to align the NGS reads to the HLA allele dictionary. We checked the operation in version 2.8.3. After installing bowtie2, you must create a direct pass to "bowtie2" in your computer environment. For example, if you are using bash, add "export PATH = $PATH:/path_to_bowtie2" to .bashrc or .bash_profile.

2.3 Installing HLA-HD

1. Download the latest version of HLA-HD onto your computer from the web site (https://www.genome.med.kyoto-u.ac.jp/HLA-HD/).
2. Uncompress the downloaded tar.gz file using "tar -zxvf hlahd.latest-version.tar.gz."
3. Move to the uncompressed directory and type "sh install.sh." For the installation, the g++ compiler by the GNU Compiler Collection (https://gcc.gnu.org/) must be installed on your computer (*see* **Note 1**).
4. Add the installed directory to PATH in your computer environment; i.e., change the .bashrc to "export PATH = $PATH:/path_to_HLA-HD_install_directory/bin."
5. If the hlahd.sh file in /path_to_HLA-HD_install_directory/bin/ is not in an executable format, change the file permission by the command, "chmod +x hlahd.sh."

2.4 Updating the HLA Dictionary

You can update the *HLA* allele dictionary to the current release of the IPD-IMGT/HLA database by the command, "sh update.dictionary.sh." The database update feature was added to HLA-HD after the version 1.1.0 was published. Wget command (https://www.gnu.org/software/wget/) is required for the database update (*see* **Note 2**).

2.5 Input Data Set

HLA-HD was developed to enable HLA typing of a whole genome, a whole exome, or a target sequence data of *HLA* genes. RNA-Seq data can also be applied without any setting changes (*see* **Note 3**). Both single and paired end short reads produced by NGS can be applied to this software.

3 Methods

3.1 Running

1. Check the open files value on your computer by the command: ulimit -Sa. If the open files value is less than 1024, type "ulimit -n 1024" or change the environment file of your computer with a command such as "/etc./security/limits.conf."
2. Run the HLA-HD: For paired-end short read data, the input command is hlahd.sh [−m <int>] [−t <int>] [−c 0 to 1.0] [−f /path/to/freq/data] <fastq data 1> <fastq data 2> <hla gene split file> /path/to/dictionary <result name> /output/directory. If the NGS data are from a single end short read, input the same file name with <fastq data 1> to <fastq data 2>.
3. The *HLA* genes you want to type must be included in <hla gene split file>. All of the classical *HLA* genes and other major genes are listed in the HLA_gene.split.txt file in the installed directory (*see* **Note 4**).
4. -m: minimum length in input reads. A read whose length is shorter than this parameter is ignored. Default size is 100.
5. -t: number of cores used to execute the program. This parameter refers to both mapping and typing.
6. -c: if a match sequence is not found in the dictionary, trim the read until some sequence is matched to or reaches this ratio (Fig. 1). Default is 1.0 (*see* **Note 5**).
7. -f: this option enables allele determination by allele frequencies from allele count data in cases where an allele pair is not uniquely determined by the end of the run (*see* **Note 6**). The default data exist in the installed directory (/hlahd.version/freq_data).
8. Using the 1000 Genomes Project [13] WES data with accession number SRR360288, HLA-HD can be executed by typing the command: hlahd.sh -t 10 -m 100 -c 0.95 -f freq_data/SRR360288_1.fastq SRR360288_2.fastq HLA_gene.split.txt dictionary/ SRR360288 results.

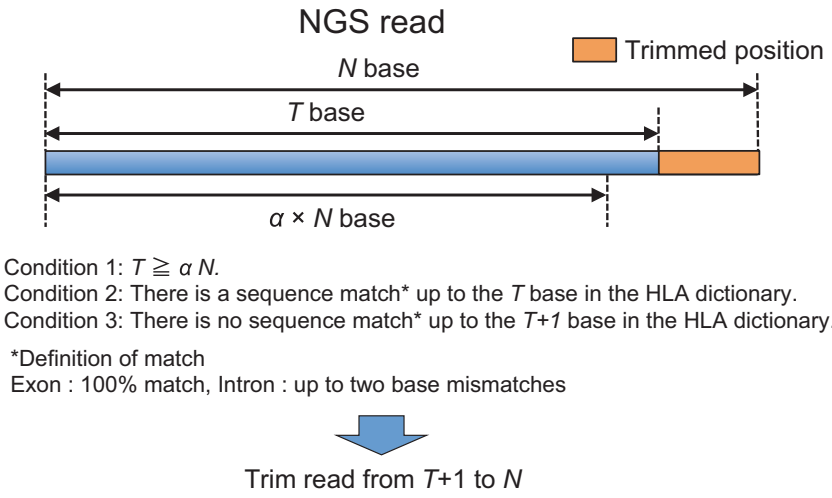


Fig. 1 Trimming process of the NGS read. If the read does not match any sequence in the HLA allele dictionary, even though there is a sequence match up to the ***T*** (≧α***N***) base of the read, the method trims the read from ***T*** + 1 to ***N***

3.2 Interpretation of the Typing Results

1. An execution result of HLA-HD is put into the ./output/directory/sampleID/result/ directory (e.g., /results/SRR360288/result/). If you want to keep the disk space, remove other directories.
2. The typed *HLA* alleles of input genes listed in <hla gene split file> are recorded in sampleID_final.result.txt at 6-digit resolution as tab-separated text (Table 2). One allele is represented by a hyphen if the gene was typed as having homozygous alleles. If a candidate has not been determined for an allele pair by the end of the run, multiple candidates are listed in parallel. On the other hand, the allele pair is recorded as "Not typed" if no candidates were obtained.
3. Details of the typing results for each gene are recorded to sampleID_gene.est.txt (e.g., SRR360288_G.est.txt). Many pairs at 8-digit resolution are recorded below the "#Best allele pair." After these allele pairs, information (coverage average and number of uncovered bases) about the mapped reads on common exons set in <hla gene split file> are recorded (Fig. 2). If "#Other ambiguous pair" is shown in the output file, allele pairs such as "ambiguous allele combinations," which are ambiguous allele pair conjugates between exons [6, 14] and are discarded from candidates in the iterative score maximization, are listed [11].
4. The information about all of the mapped reads for typed alleles is recorded to sampleID_gene.read.txt (e.g., SRR360288_DRB1.read.txt). The mapped exon with the position and the weight used for calculating a mapping score is recorded (Fig. 3) (*see* **Note 7**).

Table 2
Typed alleles from the WES SRR360288 data by HLA-HD

Gene	Allele 1	Allele 2	Allele 1 (second candidate)	Allele 2 (second candidate)
A	*HLA-A*02:11:01*	*HLA-A*02:01:01*		
B	*HLA-B*35:05:01*	*HLA-B*15:04:01*		
C	*HLA-C*01:02:01*	*HLA-C*04:01:01*		
DRB1	*HLA-DRB1*04:11:01*	*HLA-DRB1*09:01:02*		
DQA1	*HLA-DQA1*03:01:01*	*HLA-DQA1*03:02:01*		
DQB1	*HLA-DQB1*03:03:02*	*HLA-DQB1*03:02:01*		
DPA1	*HLA-DPA1*02:01:01*	–		
DPB1	*HLA-DPB1*14:01:01*	–		
DRB3	Not typed	Not typed		
DRB4	*HLA-DRB4*01:03:01*	*HLA-DRB4*01:03:02*		
E	*HLA-E*01:01:01*	–		
F	*HLA-F*01:01:01*	–		
G	*HLA-G*01:01:01*	*HLA-G*01:01:02*		
L	*HLA-L*01:02*	*HLA-L*01:01:01*	*HLA-L*01:02*	*HLA-L*01:01:02*

Listed genes are a portion of the entire typing results

4 Notes

1. You can also compile programs according to your computer environment by replacing the compile commands in install.sh.
2. If you want to use previous release data, access the github site "https://github.com/ANHIG/IMGTHLA" and obtain the hla.dat file of the appropriate release. The HLA allele dictionary can then be updated by executing the shell command "update.dictionary.sh" by deleting the line of the first wget command. We checked the HLA-HD using release 3.29.0.1.
3. We checked the typing results of RNA-Seq data using only a few samples. However, with the HLA-HD algorithm, there is no disadvantage to using RNA-Seq data compared with DNA sequence data.
4. If you want to include a new gene, add the gene name, the number of exons, and the Boolean value for a common exon

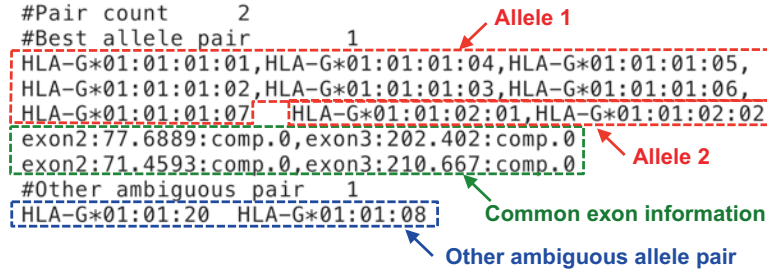


Fig. 2 Detailed typing results for ***HLA-G*** from the WES SRR360288 data. All allele pairs having the same score are listed by 8-digit values (boxed by red dotted lines). Information about common exons is provided next (boxed by green dotted line). The term "comp" means that the exon is completely covered by reads. If an uncovered nucleotide exists, "incomp" appears with the number of uncovered bases after a colon. If there is an ambiguous allele pair with partial exons, the pair is also listed (boxed by blue dotted line). The allele pair ***HLA-G*01:01:08*** and ***HLA-G*01:01:20*** is recorded as an ambiguous typing combination of ***HLA-G*01:01:01*** and ***HLA-G*01:01:02*** over exons 2 and 3 in the IPD/IMGT/HLA database (release 3.29.0.1)

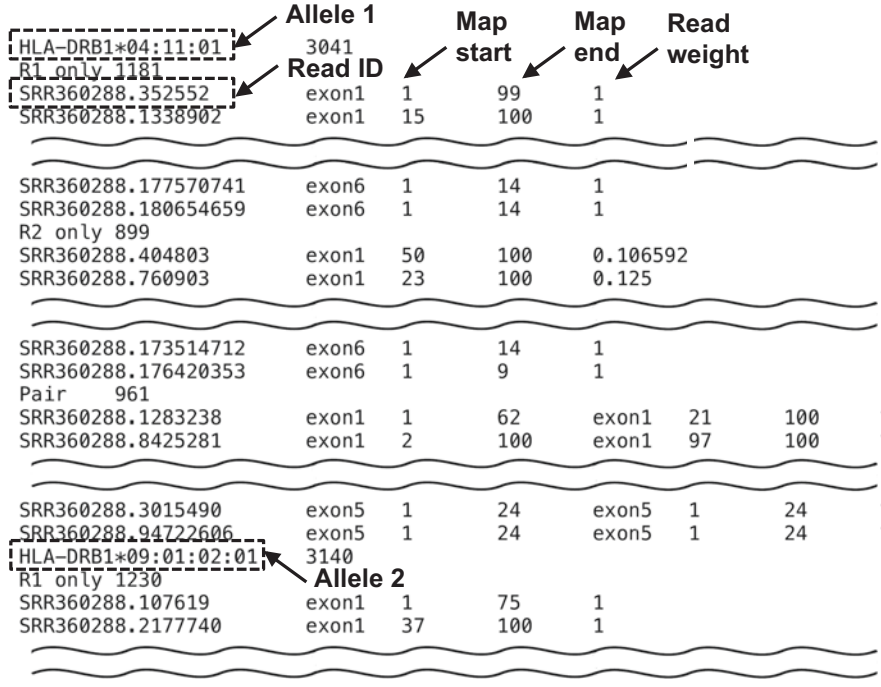


Fig. 3 Details of the mapped reads for ***HLA-DRB1*** in the typing result from the WES SRR360288 sample. All of the hit reads with the mapped position and calculated weight are recorded

to <hla gene split file> by a tab-separated value line (e.g., DRB1<tab>6<tab>010000). The common exon means that any allele registered in the IPD-IMGT/HLA database has sequence information for this exon. The common exons are set to 1 in the file and mostly correspond to exons 2 and 3 for class I genes or exon 2 for class II genes.

5. Trimming was not used in the original publication [11]. It was implemented later to adopt NGS data whose read length is longer than 100 bases. We recommend setting –c to 0.5 for 300 × 2 base-paired end-read data such as the data observed with MiSeq.
6. If a pair is not uniquely determined even after the typing algorithm is finished, the allele pair (A, B) with the highest P(A) P(B) is selected. In this case, P(Z) is the frequency of allele Z in the Allele Frequency Net Database [15].
7. HLA-HD selects the most suitable allele pair from the read counts weighted according to the number of hit alleles among all of the *HLA* genes [11]. Therefore, there are reads that do not entirely contribute to the typing result even if these are matched with the typed allele. These reads may be assigned to other genes.

References

1. Gabriel C, Danzer M, Hackl C et al (2009) Rapid high-throughput human leukocyte antigen typing by massively parallel pyrosequencing for high-resolution allele identification. Hum Immunol 70:960–964. https://doi.org/10.1016/j.humimm.2009.08.009
2. Gabriel C, Fürst D, Faé I et al (2014) HLA typing by next-generation sequencing - getting closer to reality. Tissue Antigens 83:65–75. https://doi.org/10.1111/tan.12298
3. Lind C, Ferriola D, Mackiewicz K et al (2010) Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum Immunol 71:1033–1042. https://doi.org/10.1016/j.humimm.2010.06.016
4. Saiki RK, Walsh PS, Levenson CH, Erlich HA (1989) Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. Proc Natl Acad Sci U S A 86:6230–6234
5. Santamaria P, Lindstrom AL, Boyce-Jacino MT et al (1993) HLA class I sequence-based typing. Hum Immunol 37:39–50
6. Robinson J, Halliwell JA, Hayhurst JD et al (2015) The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 43:D423–D431. https://doi.org/10.1093/nar/gku1161
7. Kim HJ, Pourmand N (2013) HLA typing from RNA-seq data using hierarchical read weighting [corrected]. PLoS One 8:e67885. https://doi.org/10.1371/journal.pone.0067885
8. Bai Y, Ni M, Cooper B et al (2014) Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. BMC Genomics 15:325. https://doi.org/10.1186/1471-2164-15-325
9. Huang Y, Yang J, Ying D et al (2015) HLAreporter: a tool for HLA typing from next generation sequencing data. Genome Med 7:25. https://doi.org/10.1186/s13073-015-0145-3
10. Szolek A, Schubert B, Mohr C et al (2014) OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics (Oxford) 30:3310–3316. https://doi.org/10.1093/bioinformatics/btu548
11. Kawaguchi S, Higasa K, Shimizu M et al (2017) HLA-HD: an accurate HLA typing algorithm for next-generation sequencing data. Hum Mutat 38:788–797. https://doi.org/10.1002/humu.23230

12. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
13. 1000 Genomes Project Consortium, Abecasis GR, Auton A et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65. https://doi.org/10.1038/nature11632
14. Adams SD, Barracchini KC, Chen D et al (2004) Ambiguous allele combinations in HLA class I and class II sequence-based typing: when precise nucleotide sequencing leads to imprecise allele identification. J Transl Med 2:30. https://doi.org/10.1186/1479-5876-2-30
15. González-Galarza FF, Takeshita LYC, Santos EJM et al (2015) Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. Nucleic Acids Res 43:D784–D788. https://doi.org/10.1093/nar/gku1166

Chapter 17

Accurate Assembly and Typing of HLA using a Graph-Guided Assembler *Kourami*

Heewook Lee and Carl Kingsford

Abstract

Accurate typing of human leukocyte antigen (HLA) is essential for successful organ transplantation and HLA genes are heavily associated with various diseases. Widely used typing assays often involve a set of specially designed primers or probes requiring additional experiments. With the maturing of high-throughput sequencing (HTS) technologies, whole genome sequencing (WGS) as well as other HTS assays are becoming more accessible even in the clinical settings. We describe various computational methods capable of directly typing HLA genes using HTS data including *Kourami*, our HLA assembler. *Kourami* is the first HLA assembler capable of discovering novel alleles. *Kourami* assembles full-length sequences across the peptide-binding regions of HLA genes. Here, we focus on how a user would use *Kourami* on a new sample. We demonstrate the application by typing HLA alleles from a recently published WGS data with validated HLA types using *Kourami*.

Key words Whole genome sequencing, WGS, HLA, Assembly, High-throughput, Bioinformatics in silico

1 Introduction

Human leukocyte antigen (HLA) genes encode for the major histocompatibility complex (MHC) consisting of cell surface proteins that control immune responses, and they are essential to regulation of the immune system. HLA genes also are associated with various infectious and autoimmune diseases, and play a large role in transplant rejection, thus making HLA typing an indispensable assay. For these reasons, combined with the rapidly increasing availability of high-throughput sequencing (HTS), accurate computational HLA typing from high-throughput sequencing data is an important problem for clinical application as well as for understanding the underlying biology and evolution of highly polymorphic HLA sequences.

Chief among the challenges in computational typing of HLA alleles is the extremely high level of polymorphisms found in the MHC region. The unusually high polymorphism results in a huge

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_17,

Fig. 1 Hierarchical nomenclature of HLA types and an example of 'G' grouping allele. An example HLA type is shown with annotation for each part of HLA nomenclature in (**a**). An allele shown in (**b**) has a different number assigned for the exon part compared to the allele in (**a**) meaning that they differ in one or more exons. Two alleles shown in (**a**) and (**b**) share identical sequence across exon 2 and 3, therefore they are both assigned to the same 'G' grouping type shown in (**c**)

number of HLA haplotypes (over 15,000 known alleles just for 6 HLA genes -HLA-A, -B, -C, -DQA1, -DQB1, -DRB1 -that are routinely typed). Among the known HLA-A, -B, and -C alleles, any pair of alleles is found to differ by at most 70 nucleotides [1] and many differ only by 1 nucleotide. Most of the nucleotide differences are densely located in the exons that are responsible for encoding the peptide-binding domain of HLA genes (exons 2 and 3 for HLA class I loci and exon 2 for HLA class II loci). Having many alleles with large nucleotide differences makes typing difficult because accurately aligning sequences to a single reference genome becomes hard. On the other hand, having many alleles that are nearly identical makes it challenging for typing algorithms to accurately pinpoint true alleles from those alleles with high sequence similarities.

In order to name HLA alleles systematically, a four-level, hierarchical number system is currently used to assign types to alleles (Fig. 1a). The first level denotes allele groups (2-digit resolution) and the second level denotes protein sequence (4-digit resolution). The last two levels denote exon sequence (6-digit resolution) and intron sequence (8-digit resolution) respectively. HLA typing is often carried out at 4- or 6-digit resolution. Furthermore, typing is often limited to the exons encoding peptide-binding regions. When using just these exons, there can be ambiguous alleles that have identical sequences across the typing exons but differ only in other regions. 'G'-grouping at 6-digit resolution is used to categorize ambiguous alleles. An example is shown in Fig. 1.

Several computational methods that are capable of typing HLA alleles using enrichment-free HTS data have been developed. Different methods use different types of HTS data ranging from WGS, whole exome sequencing (WES), to transcriptome sequencing (RNA-seq). Previously developed HLA typing algorithms use alignment-based or assembly-based approaches in

Table 1
Partial list of currently available HLA typing software

Software	Approach	Typing resolution	Input
HLAminer [18]	de novo assembly + alignment	4-digit	WGS, WES, RNA-seq
seq2HLA [1]	Alignment	4-digit	RNA-seq, WES[a]
HLAforest [19]	Alignment	4-digit	RNA-seq
PHLAT [20]	Alignment	4- or 6-digit	WGS, WES, RNA-seq
HLAreporter [21]	de novo assembly + alignment	4-digit	WGS, WES
HLA-VBSeq [22]	de novo assembly + alignment	8-digit	WGS, WES
HLA[a]*PRG* [23]	Alignment (graph)	6-digit 'G'	WGS, WES[a]
Kourami [8]	Reference-based assembly (graph)	6-digit 'G'	WGS, WES[a]

[a]Partial support

order to type HLA genes. All alignment-based approaches directly take advantage of the large collection of known HLA alleles available from the IMGT/HLA database [2]. They first align reads to the known alleles and try to select 2 best alleles per HLA gene based on varying criteria. Assembly approaches first assemble HLA reads into longer contigs and compare the contigs against known alleles to determine HLA types. A partial list of currently available HLA typing algorithms is shown in Table 1.

One major problem of the previously developed HLA typing algorithms is that they are unable to discover novel alleles. The capability to discover novel alleles is important because the number of known alleles is still increasing rapidly, indicating that there are many alleles yet to be discovered. To address this problem, a large amount of time and effort is being invested by immunogenetic communities to study rare and novel alleles. The previously developed computational methods are database-matching approaches that are designed to find the best matching alleles among the known alleles in the existing database. The typing capabilities of these methods are fundamentally limited by the incompleteness of the current database, preventing them from discovering novel alleles.

Various sequence-based typing (SBT) assays are popular because of their high accuracy. Recently, SBT combined with HTS techniques has improved the assay by speeding it up drastically. However, SBT is a HLA-specific assay using specially designed probes to target HLA genes to amplify and sequence amplicons, adding an additional cost and time, especially if there already is

high-coverage whole genome sequencing (WGS) of a test individual. Continually decreasing cost and increasing throughput of HTS have opened the door to the era of personal genomes and precision medicine. Just a merely decade ago, the first HTS-based human genome was published [3], and now large-scale high-coverage WGS studies [4–6] or government-funded projects such as Trans-Omics for Precision Medicine (TOPMed) [7] of National Institutes of Health's National Heart, Lung, and Blood Institute (NHLBI) are becoming the norm. The number of WGS samples in each study ranges from thousands to tens of thousands.

With the widely increasing popularity of WGS and the inability to discover novel alleles of database-matching approaches, we developed *Kourami* [8] an algorithm that directly takes WGS reads as input to accurately assemble and type HLA alleles at 6-digit 'G'-resolution. *Kourami* represents known HLA alleles as a directed acyclic graph, where each path through the graph is a potential HLA allele. Once read alignments to known HLA alleles are projected onto the graph, *Kourami* carries out systematic modifications to the graph, encoding differences between reads and known alleles. This process effectively captures the differences that potential novel alleles may possess. Once the graph is modified and read counts are assigned as weights to edges, *Kourami* determines the best pair of paths that jointly maximizes the amount of phasing information as well as the calibrated coverage using base quality scores. *Kourami* is highly accurate (>99%), comparable to SBT-based methods. Also, *Kourami* takes only a fraction of the computing time required by other state-of-the-art methods with comparable typing accuracy [8].

Kourami was first reported in [8], and more details about its algorithm can be found there. Here, we focus on how a user would use *Kourami* on a new sample to assemble and type HLA genes. As an example, we demonstrate *Kourami's* ability to accurately assemble and type HLA alleles using a publicly available high-coverage WGS data.

2 Materials

In order for *Kourami* to assemble and type HLA alleles, *Kourami* along with additional software and dependencies must be installed. Once all required software and dependencies listed in this chapter are successfully installed, *Kourami* can be run on paired-end WGS data of an individual (Fig. 2). WGS data can either be obtained from one of the popular publicly available repositories such as NCBI SRA (https://www.ncbi.nlm.nih.gov/sra) and EBI ENA (https://www.ebi.ac.uk/ena) or generated from your own experiments.

The distribution of *Kourami* consists of a set of utility scripts and the main program. The utility scripts are used to (1) prepare

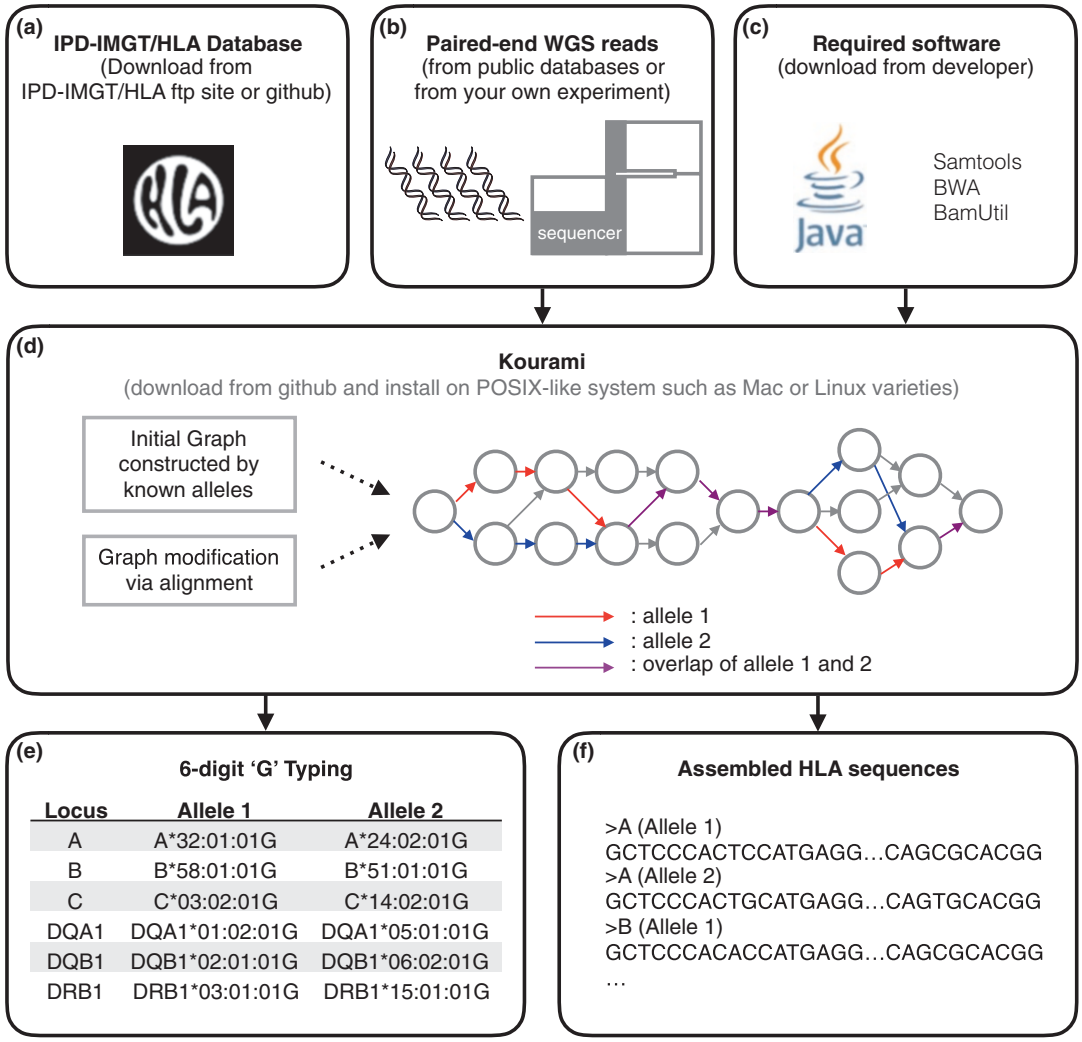


Fig. 2 Computational workflow of *Kourami*. *Kourami* requires multiple sequence alignments of known alleles downloadable from IPD-IMGT/HLA database (**a**) and high-coverage paired-end WGS data (**b**) as input. Several other software that are readily available online are required for *Kourami* to run properly (**c**). *Kourami* uses the known alleles to construct an initial directed acyclic graph and further modifies the graph to encode both the known alleles and other possible novel alleles. Simultaneous assembly of two alleles is obtained by finding two paths through the graph (**d**). *Kourami* outputs 6-digit 'G' types (**e**) as well as the assembled sequences (**f**)

Kourami reference panel from an IMGT/HLA database release or a custom set of known HLA allelic sequences, (2) download various flavors of GRCh38 human reference genome, and (3) extract the reads that are likely be originated from HLA regions and realign them to *Kourami* reference panel. The main program takes the extracted reads aligned to *Kourami* reference panel to assemble each HLA locus separately and provide typing results.

2.1 Hardware

`Kourami` was developed on Ubuntu 14.04 and the main program can be compiled and run on various operating systems; however, the utility scripts require a POSIX-like environment such as Linux or Mac OS X.

2.2 Software Dependencies

`Kourami` is mainly written in Java and the utility scripts included in each distribution are written in bash. In each release, a pre-compiled executable jar as well as the source code are available for download. Users do not need to install Apache Maven or the third party Java libraries listed below, when using a pre-compiled executable. In order to compile the main program, the Java SE Development Kit (JDK) must be installed and Apache Maven 3.3 or higher installation is recommended. In order to execute `Kourami`, having just Java SE Runtime Environment (JRE) is sufficient. The utility scripts use the alignment software *bwa* [9] and use *Samtools* [10] and *BamUtil* (https://genome.sph.umich.edu/wiki/BamUtil) for SAM/BAM file processing. The following software must be installed to successfully compile and run `Kourami` (the libraries used for the main program are listed as well):

1. Java (https://www.oracle.com/javadownload): `Kourami` was developed using Java 8. Java 8 or newer is required.
2. Apache Maven (https://maven.apache.org): Apache Maven 3.3 or newer is strongly recommended for compiling `Kourami` (not necessary when using a precompiled executable).
3. Third party Java libraries: *JGraphT* (http://www.jgrapht.org), *HTSJDK* (http://samtools.github.io/htsjdk), The Apache Commons CLI (https://commons.apache.org/proper/commons-cli/), and *fastutil* (http://fastutil.di.unimi.it) are used by `Kourami`, and they are automatically downloaded by Apache Maven during compilation (not necessary when using a pre-compiled executable).
4. Sequence alignment software: Although it is possible to substitute other HTS aligners supporting SAM/BAM output, the utility scripts are specifically written to use *bwa*.
5. SAM/BAM processing software: *Samtools* and *BamUtil* are used by the utility scripts to process SAM/BAM files.

2.3 Download and Install `Kourami`

1. Download the latest version of `Kourami` from https://github.com/Kingsford-Group/`kourami`/releases/latest.
2. We recommend downloading the source code in tar.gz format from the latest release page.
3. Unzip and untar the downloaded file using *tar* to a desired location.

4. A copy of pre-formatted IMGT/HLA database can be automatically downloaded and indexed by running the following command:

```
$ cd /path/to/Kourami
$ scripts/download_panel.sh
```

5. Build an executable JAR file using either *Apache Maven* (Recommended for Automatic Download of Dependencies) by running the following command from the installation directory of `Kourami`:

```
$ mvn install
```

2.4 Download the Human Reference Genome

The current version of the human reference genome (GRCh38) comes in multiple components and they are:

1. Primary assembly: This includes chromosomes, unplaced contigs, unlocalized contigs, and the Epstein-Barr virus (EBV) genome.
2. Decoy sequences: This is a set of sequences that are not part of the human genome reference but are included to capture sequences that are often part of sequencing data.
3. Alternate sequences (ALT): ALT sequences are alternate haplotype sequences, including 8 MHC haplotype sequences.
4. HLA sequences: A collection of HLA sequences are included in GRCh38, and it is maintained by Heng Li from Broad Institute.

Depending on the combination of the components you choose, there can be differences in downstream analysis. We recommend using the full reference (hs38DH) including all of the components. The full reference is also used by the Genome Analysis Toolkit (GATK) [11], a widely used variant calling method, as well as the 1000 Genomes Project [12]. Additionally for HLA read extraction, `Kourami` uses another variant of the reference (hs38NoAltDH), which includes all of the components except ALT. The full reference as well as the variant version can be downloaded by running the following command from the installation directory of `Kourami`:

```
$ cd /path/to/Kourami
$ scripts/download_grch38.sh hs38DH
$ scripts/download_grch38.sh hs38NoAltDH
```

Upon successfully running both commands, the reference fasta files (hs38DH.fa and hs38NoAltDH.fa) is placed under the "resources" directory and index files for *bwa* aligner are created. If there already is a downloaded copy of the human reference genome, the fasta file must be separately indexed for *bwa*. Based on our testing, downloading both genomes should take only about 3 to 5 min.

2.5 Download Whole Genome Sequencing Data from Public Repositories

Kourami requires high-coverage paired-end WGS data. The minimum recommended coverage is 30X and to ensure high-quality assembly of alleles 40-50X coverage is desirable. When downloading WGS data from publicly available repositories such as the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) or the European Nucleotide Archive (ENA) of the European Bioinformatics Institute (EMBL-EBI), data often comes either in fastq files (raw sequencing reads) or BAM files (aligned reads to a reference genome). Here, we show a simple method to download sequencing data from ENA as it provides direct ftp links for fastq files.

We specifically want to choose a WGS dataset with validated HLA types to asses the accuracy of *Kourami*. Here, we choose WGS data from the recently published diploid genome assembly of the Korean individual AK1 [13] as it provides both the high-coverage sequencing data as well as the validated HLA results. Since a BAM submission is not available for this data, we download fastq files.

The run accession number for the data is SRR3602759. All fastq files under a run accession number is available from ftp://ftp.sra.ebi.ac.uk/vol1/fastq/<dir1>[/<dir2>]/<run accession>, where <*dir1*> is the first 6 characters of the run accession and <*dir2*> is only used for accession number with 7 or more digits. For an accession number with 7 digits, <*dir2*> is '00' followed by the last digit of the accession. Since the run accession we want has 7 digits, all fastq files can be accessed from ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR360/009/SRR3602759. The following simple commands create a data directory (*data/AK1*) and download all files ending with fastq.gz:

```
$ cd /path/to/Kourami
$ mkdir -p data/AK1
$ cd data/AK1
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/
SRR360/009/SRR3602759/*fastq.gz
```

3 Methods

In this section, we provide detailed instructions for running *Kourami*, from preparing data, running the method, and interpreting the results. We use the AK1 WGS data throughout the chapter to demonstrate the application of our method.

3.1 Aligning WGS Reads to the Human Reference Genome

Reads must be aligned to the reference genome prior to running *Kourami*. Although, simply just running *bwa* may be sufficient, it may be more useful to follow well-established and widely used alignment/variant-calling protocols as other downstream analysis can be easily carried out using the same BAM file. We recommend

following the GRCh38 data processing protocols [14] (a detailed document is available online [17]) of 1000 Genomes project to obtain a BAM file. Since read alignment and variant calling is often performed automatically upon sequencing, pre-computed BAM files may be available. In this case, the pre-computed BAM files can be directly used by `Kourami`.

3.2 HLA Read Extraction and Realignment

Once a BAM file is obtained, an extraction script provided in `Kourami` first must be executed to extract reads that are likely from HLA loci. Since we use the full reference genome (*see* Subheading 2.4) to align WGS data, we use *alignAndExtract_hs38DH.sh* script located under scripts directory. In case of working with a BAM aligned to a different variant of the reference genome, additional scripts included. The script works by first extracting all the reads that are aligned to the known HLA loci in the reference including ALT haplotypes. Then these extracted reads are aligned to the `Kourami` reference panel to obtain a compact BAM containing just the alignments to known HLA alleles.

1. Let *SRR3602759.bam* be a bam file containing read alignments of paired-end fastq files from Subheading 2.5 and *SRR3602759.bam.bai* be the corresponding BAM index file. We assume the location of the BAM file and its index are located under *data/AK1* along with the fastq files.
2. The script takes two mandatory parameters: a sample name and a sorted and indexed bam file (aligned to the reference). It also takes up to two optional parameters.
3. The first optional parameter flagged by *-d* is the location of the `Kourami` panel (the default location is *db* under your `Kourami` installation directory.
4. The second optional parameter flagged by *-r* is the location of the variant reference genome (hs38NoAltDH) in case it is not in the default location (the default location is *resources/hs38NoAltDH.fa*).
5. The extraction script outputs a BAM file (*SRR3602759_on_Kourami.bam*) upon running the following command successfully:

```
$ cd /path/to/Kourami/data/AK1
$ ../../scripts/alignAndExtract_hs38DH.
sh -d ../../db -r ../../resources/hs38NoAltDH.fa
SRR3602759 SRR3602759.bam
```

3.3 Running Kourami

Now that a BAM file (*SRR3602759_on_Kourami.bam*) containing HLA reads aligned to the `Kourami` panel is obtained, the main program of `Kourami` can be invoked to assemble and type HLA genes.

1. If *maven* was used to compile and build, the executable jar (*Kourami.jar*) is located under *target* directory. The standard usage is:

```
$ java -jar /path/to/Kourami.jar [-a] -d
<path_to_panel> -o <sample_name> <bam>
```

2. *-a*: When this optional flag is passed, *Kourami* assembles 13 additional HLA loci (*see* **Notes**) other than 6 default loci.
3. *-d*: This is a required parameter and <*path_to_panel*> must be either a relative or absolute path to the *Kourami* panel directory.
4. -o: This is a required parameter and <*sample_name*> can be any string with no whitespace. *Kourami* will use this string as a prefix for all of the output files. Using a sample identifier or name can be a good choice.
5. <*bam*> should be the path to the BAM obtained by running the extraction script explained in Subheading 3.2.
6. For the AK1 WGS data we are using, the following command will trigger *Kourami* to assemble and output both the allele sequences and typing results:

   $ cd /path/to/Kourami/data/AK1

   $ java -jar ../../target/Kourami.jar -d ../../db -o SRR3602759 SRR3602759_on_Kourami.bam

3.4 Interpreting the Output

A minimal amount of logging is output to the running console screen in order to show the progress through the major steps of *Kourami* and an example output screen is shown in Fig. 3. There are several output files that *Kourami* writes to the running directory and they are:

1. <*sample_name*>_<*locus*>*.typed.fa*: This is a multi-fasta file containing a pair of assembled sequences.
2. <*sample_name*>*.result*: This file contains a 6-digit 'G' typing result for all HLA loci.
3. <*sample_name*>*.log*: This is a log file and it can be used when troubleshooting a problem.

```
kourami@kourami:~/kourami/data/AK1$ java -jar ../../target/Kourami.jar
-d ../../db -o SRR3602759 SRR3602759_on_KouramiPanel.bam
----------------REF GRAPH CONSTRUCTION--------------
----------------     READ LOADING     --------------
----------------    GRAPH CLEANING    --------------
Bubble Processing and Path Assembly for:        A
Bubble Processing and Path Assembly for:        B
Bubble Processing and Path Assembly for:        C
Bubble Processing and Path Assembly for:        DQA1
Bubble Processing and Path Assembly for:        DQB1
Bubble Processing and Path Assembly for:        DRB1
kourami@kourami:~/kourami/data/AK1$
```

Fig. 3 An example terminal output of a successful running of *Kourami*

3.4.1 Assembled HLA Gene Sequences

Unlike database-matching HLA typing algorithms, `Kourami` outputs a pair of assembled sequences that fully cover the lengths of the typing exons for each HLA gene and they are written out to a separate file for each locus. For example, *SRR3602759_A.typed.fa* contains two fasta formatted sequences for HLA-A locus for the AK1 WGS data that is used throughout this chapter. Even in the case of having homozygous alleles for a locus, two identical sequences are written out to a file.

3.4.2 HLA Typing Result

In order to provide HLA types to assembled sequences, they are aligned to IMGT/HLA database to find the closest matching sequences. We define the closest matching sequence as the known allele with the minimum edit distance. Often time, there will be many known alleles that are equally close and this is mostly because of ambiguous alleles that share identical sequences across the typing exons but differ elsewhere. For this reason, we use 6-digit 'G' type whenever it is available. 'G'-grouping types are maintained also by IMGT/HLA database and can be accessed from their website (http://hla.alleles.org/alleles/g_groups.html).

3.4.3 Evaluation of HLA Typing of AK1 Genome

The HLA types for selected loci (9 HLA genes) of AK1 have been previously validated by targeted sequencing approaches as well as long-read assays and the validated types are reported in [13]. We downloaded high-coverage paired-end WGS data (151 bp × 2, SRA accession SRR3602759) and used `Kourami` to assemble and type HLA alleles (the detailed steps described in Subheading 3. The HLA alleles `Kourami` assembled had identical types to the validated alleles across all of the 9 loci. The 6-digit "G" HLA types of AK1 are shown in Table 2.

4 Notes

1. In its default setting, `Kourami` assembles exonic sequences of the peptide binding region of HLA genes for 6 commonly typed HLA loci (HLA-A, -B, -C, –DQA1, -DQB1, and -DRB1. When a flag *-a* is passed when running `Kourami`, 13 additional HLA loci can also be assembled and typed (HLA-DOA, -DMA, -DMB, -DPA1, -DPB1, -DRA, -DRB3, -DRB5, -F, -G, -H, -J, and -L).
2. Each `Kourami` distribution includes a version of pre-formatted reference panel sequences. In the case that you need a specific release of IMGT/HLA-Database as the panel reference, you can download multiple sequence alignment (MSA) flat files for HLA loci and run a formatting script (*formatIMGT.sh*) under *scripts* directory. For each IMGT/HLA-database release, the MSA flat files can be downloaded either from https://github.com/ANHIG/IMGTHLA or ftp://ftp.ebi.ac.uk/pub/databases/ipd/imgt/hla/. They are located under *alignments*

Table 2
HLA typing of AK1 by Kourami

Locus	Allele 1	Allele 2
A	A*32:01:01G	A*24:02:01G
B	B*58:01:01G	B*51:01:01G
C	C*03:02:01G	C*14:02:01G
DQA1	DQA1*01:02:01G	DQA1*05:01:01G
DQB1	DQB1*02:01:01G	DQB1*06:02:01G
DPA1[a]	DPA1*01:03:01G	DPA1*02:02:02G
DPB1[a]	DPB1*02:01:02G	DPB1*05:01:01G
DRB1	DRB1*03:01:01G	DRB1*15:01:01G
DRB3[a]	DRB3*02:02:01G	DRB3*02:02:01G

[a]Additional locus `Kourami` typed

directory. The ambiguous allele 'G'-grouping file (*hla_nom_g.txt*) is provided in each IMGT/HLA database release and it must be placed inside the downloaded *alignments* directory.

3. Typically, the matched alleles are identical to assembled sequences (zero edit distance), and this is the case with AK1 data used in this chapter. However, if typing an individual from an understudied ethnic population such as African population, assembled sequences may not be identical to database alleles more often. `Kourami`'s ability to discover novel alleles can be useful when studying such population as well as typing individuals harboring novel alleles.
4. Largely due to the cost benefit, WES is extremely popular. `Kourami` can assemble HLA allele using WES data, although it was developed to work with WGS data. Based on the previous testing on HapMap WES data [8], typing accuracy is slightly lower (94.7%) when compared to WGS typing accuracy (>99%). A caution must be used when using WES as `Kourami` may skip HLA genes if there is no coverage in parts of the typing exons. Known biases in WES such as reference allele bias, GC bias and coverage fluctuations [15, 16] may also have an effect on typing performance.FundingThis research was funded in part by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through Grant GBMF4554 to C.K., by the US National Science Foundation (CCF-1256087, CCF-1319998) and by the US National Institute of Health (R01HG007104, R01GM122935).

References

1. Boegel S, Löwer M, Schäfer M et al (2012) HLA typing from RNA-Seq sequence reads. Genome Med 4(12):102
2. Robinson J, Halliwell JA, Hayhurst JD et al (2015) The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 43(Database issue):D423–D431. https://doi.org/10.1093/nar/gku1161
3. Wheeler DA, Srinivasan M, Egholm M et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452(7189):872–876
4. Telenti A, Pierce LCT, Biggs WH et al (2016) Deep sequencing of 10,000 human genomes. Proc Natl Acad Sci U S A 113(42): 11901–11906
5. Nagasaki M, Yasuda J, Katsuoka F et al (2015) Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun 6:8018
6. Gudbjartsson DF, Helgason H, Gudjonsson SA et al (2015) Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 47(5):435–444
7. National Heart, Lung and Blood Institute (2017) Trans-Omics for Precision Medicine (TOPMed) Program. https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed/. Accessed 29 Nov 2017
8. Lee H, Kingsford C (2018) Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. Genome Biology 19:16
9. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25(14):1754–1760
10. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16): 2078–2079
11. McKenna A, Hanna M, Banks E et al (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9): 1297–1303
12. Consortium T1GP (2015) A global reference for human genetic variation. Nature 526(7571):68–74
13. Seo J-S, Rhie A, Kim J et al (2016) De novo assembly and phasing of a Korean human genome. Nature 538(7624):243–247
14. Zheng-Bradley X, Streeter I, Fairley S et al (2017) Alignment of 1000 genomes project reads to reference assembly GRCh38. GigaScience 6(7):1–8
15. Meienberg J, Bruggmann R, Oexle K et al (2016) Clinical sequencing: is WGS the better WES? Hum Genet 135(3):359–362
16. Asan, Xu Y, Jiang H et al (2011) Comprehensive comparison of three commercial human whole-exome capture platforms. Genome Biol 12(9):R95
17. 1000 Genomes This README explains the alignment pipeline used to remap all the 1000 Genomes Project Phase 3 reads to GRCh38DH. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/README.1000genomes.GRCh38DH.alignment. Accessed 29
18. Warren RL, Choe G, Freeman DJ et al (2012) Derivation of HLA types from shotgun sequence datasets. Genome Med 4(12):95
19. Kim HJ, Pourmand N (2013) HLA haplotyping from RNA-seq data using hierarchical read weighting. PLoS One 8(6):e67885
20. Bai Y, Ni M, Cooper B et al (2014) Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. BMC Genomics 15:325
21. Huang Y, Yang J, Ying D et al (2015) HLAreporter: a tool for HLA typing from next generation sequencing data. Genome Med 7(1):25
22. Nariai N, Kojima K, Saito S et al (2015) HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data. BMC Genomics 16(Suppl 2):S7
23. Dilthey AT, Gourraud P-A, Mentzer AJ et al (2016) High-accuracy HLA type inference from whole-genome sequencing data using population reference graphs. PLoS Comput Biol 12(10):e1005151

Chapter 18

AmpliSAS and AmpliHLA: Web Server Tools for MHC Typing of Non-Model Species and Human Using NGS Data

Alvaro Sebastian, Magdalena Migalska, and Aleksandra Biedrzycka

Abstract

AmpliSAS and AmpliHLA are web server tools for automatic genotyping of MHC genes from high-throughput sequencing data. AmpliSAS is designed specifically to analyze amplicon sequencing data from non-model species and it is able to perform de-novo genotyping without any previous knowledge of the reference alleles. AmpliHLA is a human-specific version, it performs HLA typing by comparing sequenced variants against human reference alleles from the IMGT/HLA database. Here we describe four genotyping protocols: the first two use amplicon sequencing data to genotype the MHC genes of a passerine bird and human respectively; the third and fourth present the HLA typing of a human cell line starting from RNA and exome sequencing data respectively.

Key words Bioinformatics, Next-generation sequencing, NGS, Amplicon sequencing, RNA-Seq, WES, WXS, MHC, HLA, Genotyping, Alleles, Haplotypes, IMGT, AmpliSAS, AmpliHLA

1 Introduction

The major histocompatibility complex (MHC) encodes a family of genes of central importance in vertebrate adaptive immunity. In human, the MHC is commonly named as human leukocyte antigen system (HLA). MHC molecules are responsible for binding and presenting antigens to the immune system T-cells. There are two classes of classical MHC genes involved in adaptive immunity: class I which encode molecules that present peptides from the intracellular environment (e.g., viruses) to T cells; and class II that encode molecules to present peptides from the extracellular environment (e.g., bacteria) [1]. MHC genes are the most polymorphic genes currently characterized in vertebrates [2]. This polymorphism is believed to be primarily driven by co-evolving pathogens and through mate choice, and maintained in the populations by balancing selection [3–7]. The number of MHC genes differs greatly between species (Fig. 1), and in some taxa the

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_18,

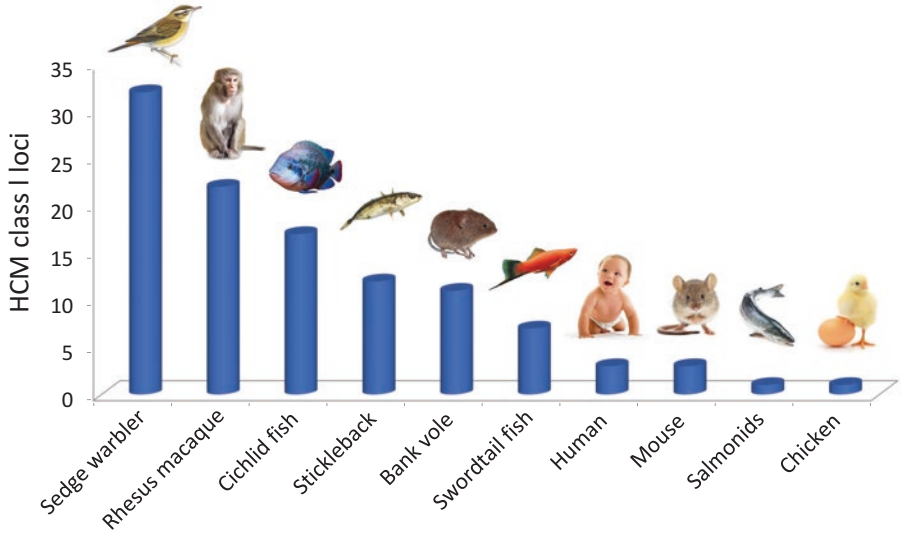


Fig. 1 Maximum number of MHC class I loci extracted from the literature for some representative species up to date

number of MHC genes differs also between individuals of the same species [8–19].

As a simplification, we can say that humans have a fix number of classical, functional major HLA loci: A, B, and C for class I and DP, DQ, and DR for class II (Fig. 1) [14]. Some of the HLA loci are extremely polymorphic (Table 1), for example, currently 4828 alleles have been described for HLA-B locus (IMGT/HLA database 3.29 release, July 2017) [20]. In birds there is a great variability in the complexity of the MHC regions, the number of MHC genes and their variabilities (Fig. 1). Chickens have a minimal essential MHC with two classical class I genes but only one is expressed at a high level and its diversity is considerably lower than human [17, 21, 22]. In contrast, birds of the order Passeriformes have much more complex MHC systems with dozens of genes highly duplicated and polymorphic [8, 23–25].

MHC genotyping is especially demanding in non-model species with highly duplicated MHC genes and where limited genomic information does not allow us to design locus-specific probes or primers. In the past, expensive and time-consuming methods were required for genotyping [26], but nowadays, next-generation sequencing (NGS) of MHC amplicons has become the method of choice for non-model species [24, 27–30] and human [31–33]. Amplicon sequencing technique is based on sequencing multiple PCR products (amplicons) at once by means of NGS technologies (Fig. 2). With a single experiment it is possible to accurately genotype hundreds of individuals with complex MHC systems [34, 35]. The **Note 1** describes briefly the technique workflow and the **Note 2** presents the benefits introduced by NGS.

Table 1
Number of alleles for each HLA gene as registered at the 3.29 release of the IMGT/HLA database from the European Bioinformatics Institute (EBI) at July 2017

Class I		Class II			
Gene	Alleles	Gene	A alleles	B alleles	A × B
HLA-A	3968	DR	7	2376	16,632
HLA-B	4828	DQ	94	1142	107,348
HLA-C	3579	DP	53	894	47,382

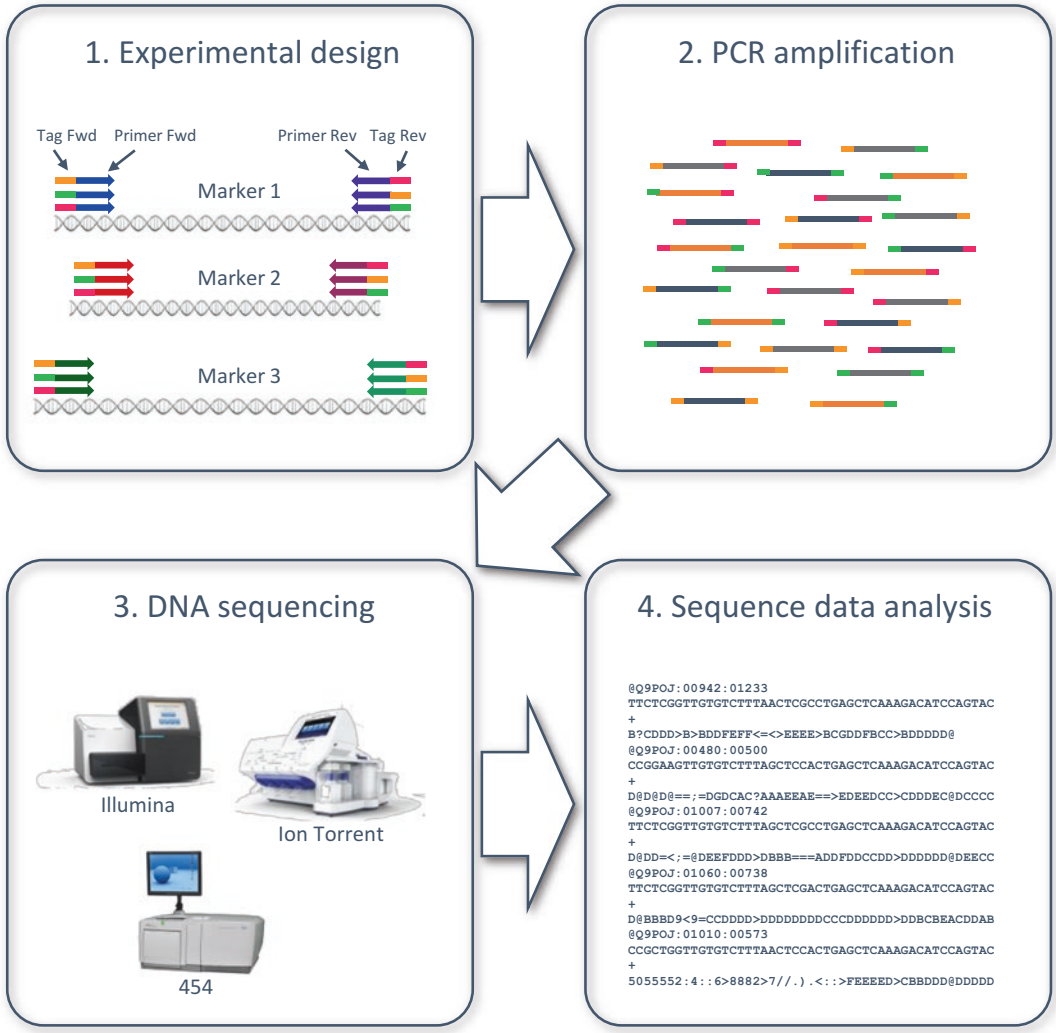


Fig. 2 Amplicon sequencing workflow schema: (**1**) experimental design (marker regions, primers, and tags), (**2**) PCR amplification, (**3**) DNA sequencing, and (**4**) sequence data analysis

However, relatively high error rates in the amplicon sequences, stemming both from intrinsic sequencing error rates of NGS technologies and PCR errors, such as chimera formation, arise new genotyping challenges (*see* **Note 3**). Different strategies have been proposed to detect sequence errors and correct them without altering genotypes [36, 37]. AmpliSAS is an error correction strategy that clusters real alleles with low frequency similar variants based on the particular error-rate of the NGS technology used [38]. In a recent benchmark, AmpliSAS produced reliable genotypes for the sedge warbler (*Acrocephalus schoenobaenus*), a passerine bird with a highly complex MHC system composed of dozens of loci and thousands of alleles (**Note 4**) [37]. Furthermore, AmpliSAS has been successfully validated in other non-model species as bank vole, guppy, three-spine stickleback, blue petrel, or black-tailed godwit [12, 38–40].

AmpliHLA is an adaptation of the AmpliSAS algorithm for human HLA typing, it combines the genotypes from multiple markers of the same locus and compares them with the deposited HLA alleles from the IMGT/HLA database [20]. As a result, it retrieves the HLA types with the highest possible resolution. AmpliSAS algorithm has previously been tested with human amplicon data retrieving accurate genotypes [38]. Recently, AmpliHLA has been expanded with an adaptation of the Seq2HLA algorithm to be able to analyse RNA-Seq and whole exome sequencing (WES or WXS) data [41].

Both, AmpliSAS and AmpliHLA, are available as ready-to-use web server tools at: http://evobiolab.biol.amu.edu.pl/amplisat/index.php.

Here, we present four MHC genotyping protocols. The first describes how to process amplicon data with AmpliSAS to obtain the MHC class I genotypes of five sedge warblers (passerine birds) that possess up to 56 MHC class I alleles per individual. The second presents the HLA typing with AmpliHLA of five human cell lines whose alleles were previously characterized by Sanger sequencing. The third and fourth protocols use also AmpliHLA to type of one of the previous human cell lines using RNA-Seq and WES data instead of amplicons.

2 Materials

The only resources required to replicate the analysis described in this chapter are an Internet connection and a web browser. These will suffice to learn how to use AmpliSAS, AmpliHLA, and other tools from the AmpliSAT suite (Amplicon Sequencing Analysis Tools) which are presented here.

2.1 AmpliSAS

Amplicon Sequence Assignment tool (AmpliSAS) is a web server tool designed to analyse NGS amplicon data and perform automatic genotyping of complex MHC systems (http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplisas, Fig. 3a) [38]. AmpliSAS workflow is divided into three main steps (Fig. 4): (1) sequence de-multiplexing, (2) sequence clustering, and (3) artifact

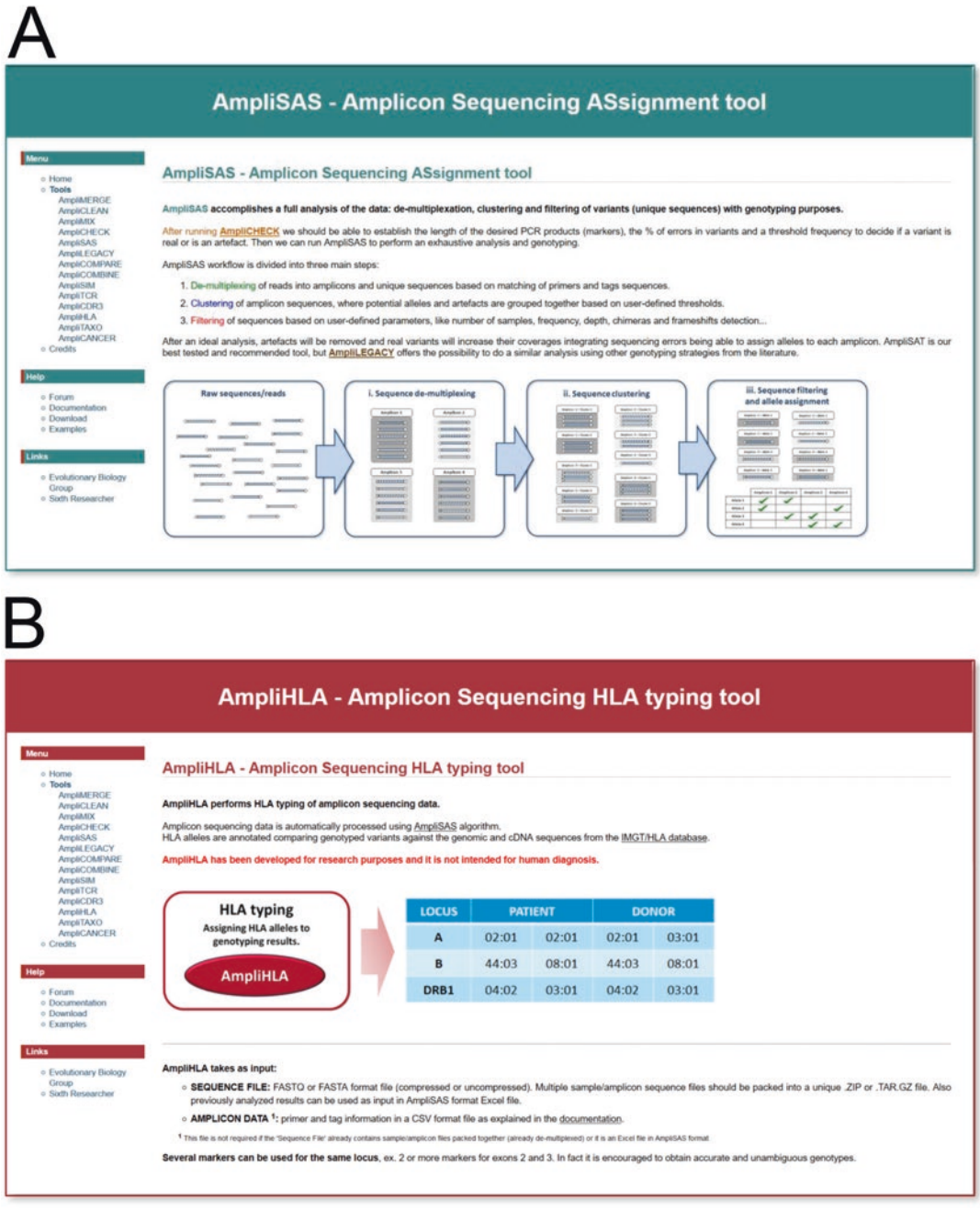


Fig. 3 AmpliSAS (**a**) and AmpliHLA (**b**) web interfaces

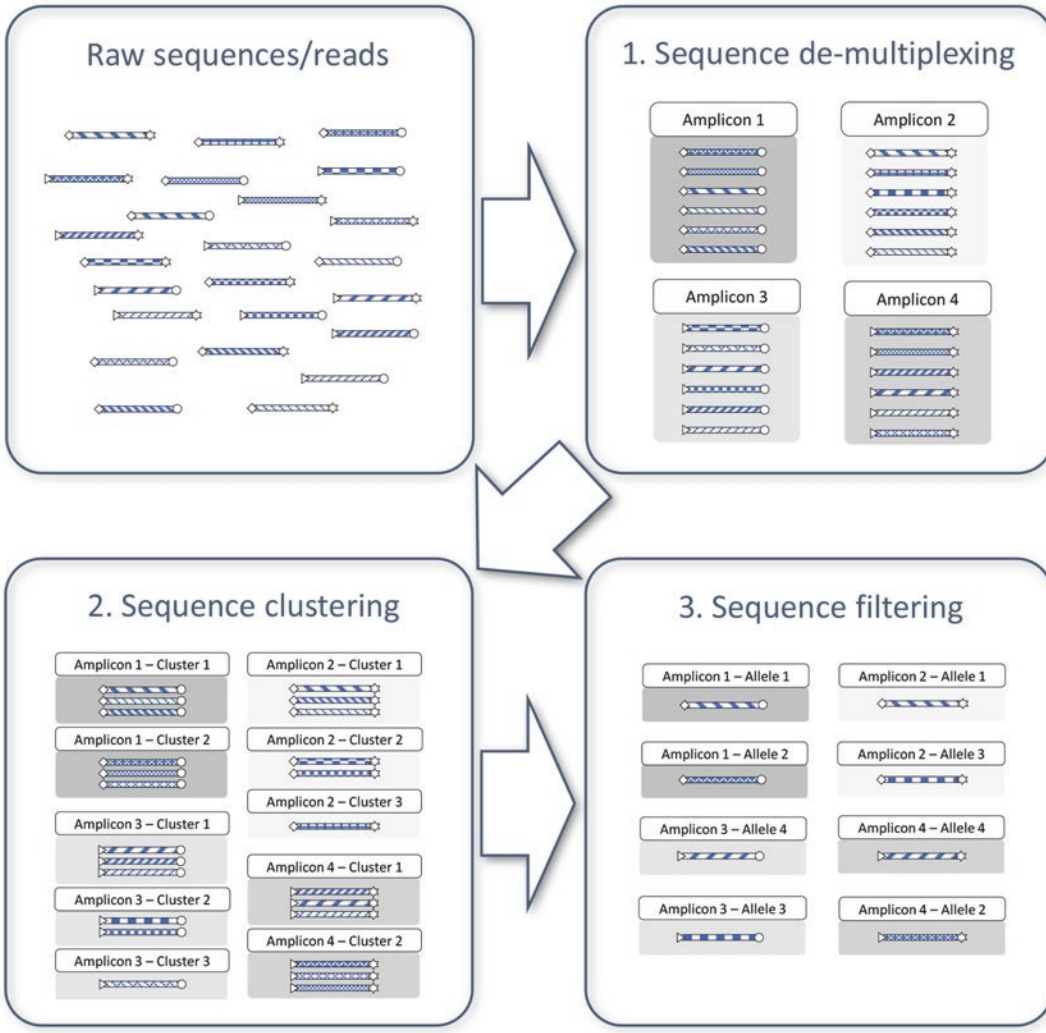


Fig. 4 AmpliSAS workflow schema: (**1**) sequence de-multiplexing, (**2**) clustering, and (**3**) filtering and allele assignment

filtering. In summary, first the reads are de-replicated and classified into amplicons. Second, during clustering, variants are aligned to each other to find sequencing errors, these erroneous variants are removed and their coverages added to the true ones. Third, remaining low frequency variants are inspected to remove artifacts and chimeric PCR products. Finally, allele sequences and frequencies for each amplicon are retrieved in an Excel spreadsheet format, making them easy to interpret. Definitions of amplicon, variant and other useful terms are listed in Table 2. Complete details about AmpliSAS algorithm can be found in [38], and a comparison of the performance of AmpliSAS against other MHC genotyping methods in [37].

Table 2
Definitions of commonly used terms in the amplicon sequencing technique: they can slightly differ between authors

Term	Definition
Marker	A DNA region to be amplified
Sample	A single genetic material to be sequenced (usually from an individual of the study organism)
Tag	A unique short DNA sequence that identifies unambiguously a sample. Tags are usually ligated after PCR amplification or directly included in one or both primers
Read	Each individual sequence retrieved by a sequencing run. A sequence run will retrieve thousands/millions of reads
Amplicon	A set of reads derived from a single PCR (one marker, one sample); may comprise products of several co-amplifying loci
Amplicon depth	Number of reads per amplicon
Variant	Unique sequence retrieved by a sequencing run. Usually multiple reads correspond to one variant (= one sequence)
Variant depth/coverage	Number of reads per variant
Per amplicon frequency	Number of reads per sequence divided by the total number of reads in a single amplicon
True variant/allele	Sequence that matches a real allele or real sequence in the sample genome
Artifact	Variant resulting from experimental/technical errors: Sequencing errors, polymerase errors, nonspecific amplifications (paralogs, pseudogenes), contaminants, PCR chimeras, etc

2.2 AmpliHLA

Amplicon Sequencing HLA typing tool (AmpliHLA) is a web server tool designed to retrieve automatically HLA haplotypes from amplicon, RNA-Seq, or WES data (http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplisas, Fig. 3b). AmpliHLA implements the AmpliSAS algorithm explained in the previous section to analyze amplicon data [38], but additionally it is able to combine the information from several amplified regions of a single locus and compare their sequences with the thousands of HLA alleles annotated in the IMGT/HLA database [20].

In the analysis of RNA-Seq and WES data, AmpliHLA uses a modified version of the Seq2HLA algorithm [41]. First, the reads are mapped with BOWTIE [42] against a curated dataset of HLA variable regions (exons 2 and 3) extracted from the IMGT/HLA database [20, 41]. Then AmpliHLA analyses mapping results in a locus-specific manner: (1) the allele with the maximum number of mapped reads is selected and these reads are substracted from the remaining allele mappings, (2) step (1) is successively repeated and the corrected mapped read numbers are annotated until there are no more alleles with mapped reads left, (3) one or two alleles with the highest corrected numbers of mapped reads are selected based on the drop of their values as in [29] and their HLA types are printed.

2.3 AmpliCOMPARE

AmpliCOMPARE is another tool from the AmpliSAT suite. It is available at: http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplicompare

AmpliCOMPARE sets side by side the genotyping results of two experimental or technical replicates, or between two different genotyping strategies. It accepts as input two Excel files with the same format as the AmpliSAS/AmpliHLA output one. With this tool it is easy to detect genotyping discrepancies highlighted in the comparison output Excel file.

3 Methods

3.1 MHC Class I Genotyping in a Passerine Bird

As previously explained, AmpliSAS algorithm is designed to genotype complex MHC gene families, such as those in the sedge warbler, a passerine bird with MHC class I copy number variation, and dozens of MHC class I loci in a single individual [8].

In the present protocol we will analyze data from a previous sedge warbler MHC class I genotyping study (accession PRJEB11775 at the European Nucleotide Archive—ENA). The purpose of the study was to use ultra-deep Illumina sequencing to resolve genotypes at exon 3 of MHC class I genes in the sedge warbler [37]. We will use a pre-processed and compressed FASTQ file with already merged and cleaned Illumina paired-end reads (*see* **Note 5**).

1. Open the AmpliSAS online submission form (Fig. 5a):http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplisas.
2. Enter a name for the run and, optionally, an email address if you desire to receive the results by email.
3. Copy and paste the following link into the 'Sequence file URL' field: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR113/008/ERR1136308/ERR1136308.fastq.gz.

   Optionally, you can download the compressed FASTQ file in your computer and upload it with the 'Browse' button (slow). If reads have been separated into single amplicon files

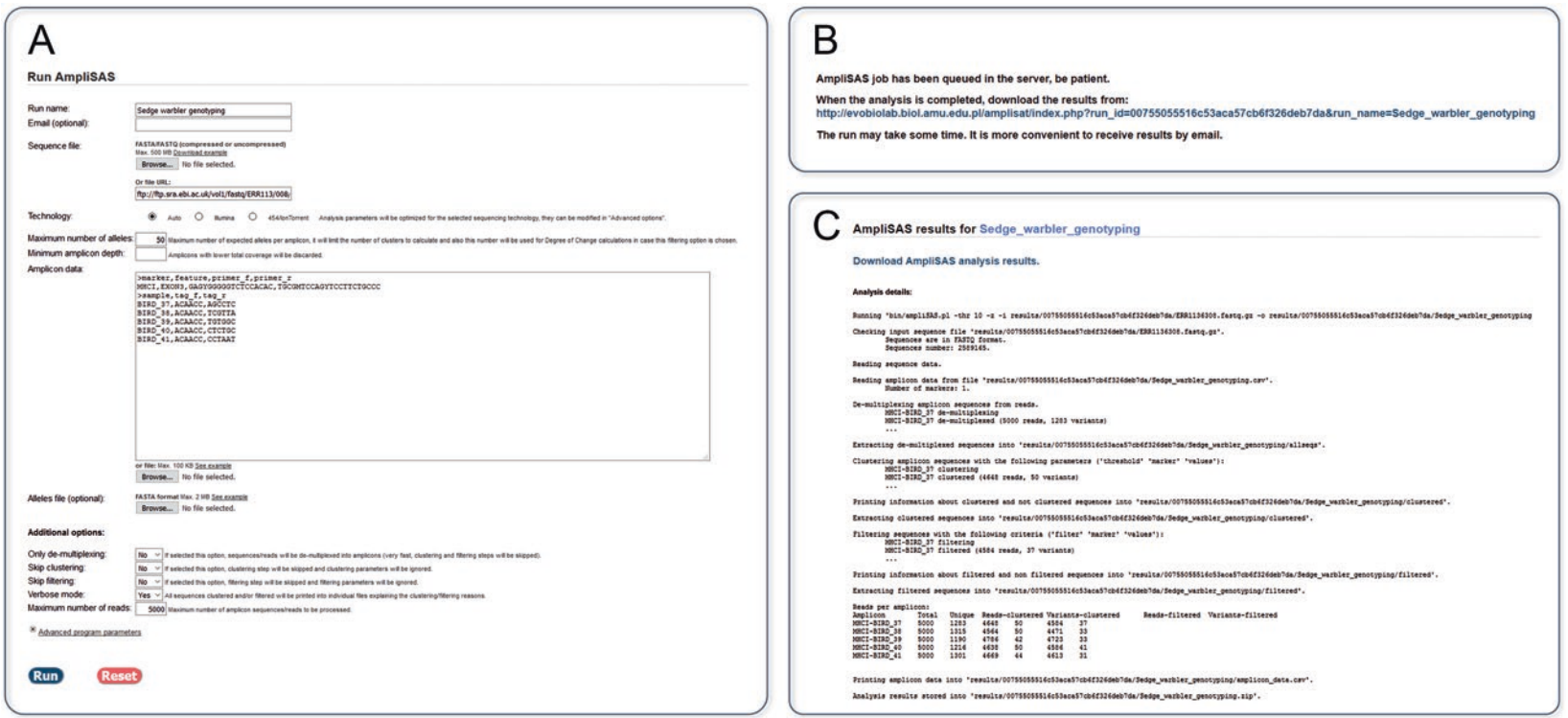


Fig. 5 (**a**) AmpliSAS input form. (**b**) Message with the link to the results page. (**c**) Output summary after analysis completion and link to the results ZIP file

after sequencing, they can be packed into a single ZIP or TGZ format file and used as input (*see* **Note 6**).

4. Select 'Auto' from the 'Technology' options to ask the program to automatically detect the sequencing technology used.
5. Copy and paste into the 'Amplicon data' field the following information about primers used to amplify the MHC class I exon 3 region and DNA tags included at the end of the primers to identify the first five individuals from the experiment:

```
>marker,feature,primer_f,primer_r
MHCI,EXON3,GAGYGGGGGTCTCCACAC,TGCGMTCCAGYTCCTTCTGCCC
>sample,tag_f,tag_r
BIRD_37,ACAACC,AGCCTC
BIRD_38,ACAACC,TCGTTA
BIRD_39,ACAACC,TGTGGC
BIRD_40,ACAACC,CTCTGC
BIRD_41,ACAACC,CCTAAT
```

   The information for all the individuals is available at the ENA experiment page: https://www.ebi.ac.uk/ena/data/view/ERX1215174

   If the input is a compressed file containing single amplicon files for each individual, then 'Amplicon data' must be left empty (*see* **Note 6**).
6. Adjust the following parameters in the submission form: 'Maximum number of alleles per amplicon': 50 and 'Maximum number of reads per amplicon': 5000. Keep the rest of parameters with default values.
7. Click the 'Run' button, and after a while (first the server has to upload the reads file), it will display the message 'AmpliSAS job has been queued in the server, be patient' together with a link to access the results when the analysis will be completed (Fig. 5b).
8. Click the link to the results page and reload it until the analysis is completed. The analysis should be finished in a few minutes, but the waiting time may depend on the current load on the server and the size of the uploaded files. Keep the results page open (Fig. 5c), in the Subheading 3.3 we will learn how to interpret them.

3.2 Customizing the MHC Class I Genotyping

In the previous section, we performed a complex MHC class I genotyping with mostly automatic and default options. Now, we will learn how to customize the analysis adjusting manually some program parameters (*see* **Note 7**).

1. Open a new tab in the web browser and repeat **steps 1–3** as explained in Subheading 3.1.
2. Select 'Illumina' from the 'Technology' options. Scroll down to the 'Advanced program parameters' section, you will notice

that some clustering and filtering parameters have been set to recommended values for Illumina data.

3. Copy and paste the following data in the 'amplicon data' field:

```
>marker,feature,length,primer_f,primer_r
MHCI,EXON3,235 238 241,GAGYGGGGGTCTCCACAC,TGCGMTCCAGYT
CCTTCTGCCC
>sample,tag_f,tag_r
BIRD_37,ACAACC,AGCCTC
BIRD_38,ACAACC,TCGTTA
BIRD_39,ACAACC,TGTGGC
BIRD_40,ACAACC,CTCTGC
BIRD_41,ACAACC,CCTAAT
```

In the previous analysis we did not specify expected allele lengths, so the program automatically set them to 241 (*see* **Note 8**).

4. Adjust the same parameters as in **step 6** from Subheading 3.1. Go to 'Advanced program parameters – Clustering parameters' section and set 'Exact length required': yes, 'Minimum dominant frequency': 10%. Go to 'Advanced program parameters—Filtering parameters' section and set 'Minimum amplicon frequency': 0.4%. These additional parameters were defined by the authors after manual inspection of the data in the original article (*see* **Note 8**) [37]. Keep the rest of parameters with the default values.

5. Repeat **steps 7** and **8** listed in Subheading 3.1.

3.3 Interpreting the Genotyping Results

1. If we have correctly followed the steps from Subheading 3.1 or 3.2, the output of AmpliSAS should look like this (Fig. 5c):

```
AmpliSAS results
Download AmpliSAS analysis results.
Analysis details:
Running 'bin/ampliSAS ...
Checking input sequence file ...
    Sequences are in FASTQ format.
    Sequences number: 2589165.
Reading sequence data.
Reading amplicon data from file ...
    Number of markers: 1.
De-multiplexing amplicon sequences from reads.
    MHCI-37 de-multiplexing
    MHCI-37 de-multiplexed (5000 reads, 1283 vari-
    ants)
    ...
Extracting de-multiplexed sequences into ...
Checking data and setting marker lengths.
    Marker 'MHCI' lengths: 235,238,241 (manual)
Clustering amplicon sequences with the following pa-
rameters
('threshold' 'marker' 'values'):
    substitution_threshold              all 1
```

```
    indel_threshold                    all 0.001
    cluster_exact_length               all 1
    min_dominant_frequency_threshold   all 10
    MHCI-BIRD_37 clustering
   MHCI-BIRD_37 clustered (4383 reads, 47 variants)
    ...
Printing information about clustered and not clus-
tered sequences into ...
Filtering sequences with the following criteria
('filter' 'marker' 'values'):
    min_amplicon_depth          all 100
    min_amplicon_seq_frequency  all 0.4
    min_chimera_length          all 10
    max_allele_number           all 50
    MHCI-BIRD_37 filtering
    MHCI-BIRD_37 filtered (4344 reads, 36 variants)
    ...
Printing information about filtered and non-filtered
sequences into ...
Reads per amplicon:
Amplicon    Total    Unique    Reads-clustered    Var-
iants-clustered    Reads-filtered    Variants-filtered
MHCI-BIRD_37 5000 1283 4383 47 4344 36
MHCI-BIRD_38 5000 1315 4513 50 4450 33
MHCI-BIRD_39 5000 1190 4744 35 4723 33
MHCI-BIRD_40 5000 1216 4600 47 4586 41
MHCI-BIRD_41 5000 1301 4597 38 4573 30
Printing amplicon data into ...
Analysis results stored into ...
```

2. Click the 'Download AmpliSAS analysis results' link to download a ZIP compressed file with the following contents:
   - 'results.xlsx': Excel file with the final genotyping results. *See* **step 3**.
   - 'allseqs', 'clustered' and 'filtered' folders: contain single amplicon FASTA files with variants recovered after every analysis step: de-multiplexing, clustering and filtering respectively. Each variant has annotated in the FASTA header its sequencing depth and frequency. All this data is also included into an Excel file per folder.
   - 'amplicon_data.csv': comma-separated values format file including the amplicon data and analysis parameters.
   - 'summary.txt': a tab-delimited file with the number of variants and associated reads retrieved after every analysis step.
3. The most informative file is 'results.xlsx', samples (individuals) are shown in columns and variants (alleles) in rows, the numeric values represent the variants' depths within each amplicon (Fig. 6). For example, the variant MHCI-0001 is present in the five individuals, having in the BIRD_39 the maximum depth (914 reads) and the variant MHCI-0002 is present in four individuals but not in the BIRD_39. The columns at the left include additional information about the variants: DNA sequence, length, sum of the depths in all the samples,

	BIRD_37	BIRD_38	BIRD_39	BIRD_40	BIRD_41
MHCI-0001	381	375	914	601	411
MHCI-0002	190	525		138	180
MHCI-0003	144		170	256	392
MHCI-0006		398	298		
MHCI-0004	172	186		132	203
MHCI-0007		184	236		266
MHCI-0005	177			291	206

Fig. 6 AmpliSAS Excel output file example. The numeric values show the variants' depths within each amplicon

number of samples containing the variant and mean, maximum and minimum frequencies along all the samples.

3.4 Comparing Two Genotyping Result Files

In the previous Subheadings 3.1 and 3.2 we obtained slightly different genotyping results due to the auto vs. manual adjustment of the genotyping parameters (*see* **Note 8**), here we will learn how to compare them with the tool AmpliCOMPARE.

1. Open the AmpliCOMPARE online submission form: http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplicompare
2. Upload the 'results.xlsx' file obtained in the Subheading 3.1 as 'First results file' and the file from the Subheading 3.2 as 'Second results file'.
3. Click the 'Run' button and follow the link to the results page.
4. The output of the comparison process will be like this:

```
AmpliCOMPARE results
Download AmpliCOMPARE analysis results.
Analysis details:
Running 'bin/ampliCOMPARE.pl ...
Reading File ...
MARKER 'MHCI':
Total unique samples: 5 (file1: 5, file2: 5)
Total seqs: 117 (file1: 117, file2: 114)
Compared samples: 5 (excluded from file1: 0, from file2: 0)
Compared seqs: 117 (missing in file1: 0, in file2: 3)
Total assignments: 176 (missing in file1: 1, missing in
file2: 3)
Comparison results written into ...
```

5. Click the link 'Download AmpliCOMPARE analysis results' to retrieve an Excel file with the differences between both genotypes.
6. Open the Excel file, variants (alleles) retrieved in the first analysis but not in the second are marked in cyan color and the opposite in magenta (Fig. 7). Common variants in both analysis remain un-formatted with their depths separated by a slash

	BIRD_37	BIRD_38	BIRD_39	BIRD_40	BIRD_41
MHCI-0001	364/381	366/375	887/914	577/601	396/411
MHCI-0002	177/190	511/525		130/138	168/180
MHCI-0003	137/144		161/170	250/256	383/392
...					
MHCI-0022	112/122				130/133
MHCI-0026	234				
MHCI-0028			220/228		
...					
MHCI-0048				134/136	
MHCI-0040	43/45	25		36/39	53/55
MHCI-0052			126/131		

Fig. 7 AmpliCOMPARE Excel output file example. The numeric values show the variants' depth in the compared files. Cyan color marks a variant that is present in the first file and not in the second, the opposite is marked in magenta

(Fig. 7). There should be three variants marked in cyan retrieved with automatic parameters (Subheading 3.1) and not retrieved with manual ones (Subheading 3.2). If we check the lengths of the different variants, they are 230 and 236 bp long, consequently they are not in-frame with the real allele lengths (235, 238 and 241) that we specified manually in the Subheading 3.2 analysis. Two of three variants have low depths, so most probably they are PCR or sequencing artifacts derived from other, higher frequency alleles. The third variant could be a product of non-specific amplification, or a pseudo-gene, rather than a technical artifact. As conclusion, both genotyping strategies perform well, but the manual adjustment of AmpliSAS parameters retrieves higher quality genotypes.

3.5 HLA Typing with Amplicon Sequencing Data

To show the functionality of AmpliHLA we will use a targeted amplicon sequencing dataset that consists of genomic sequences from exon 2 and exon 3 regions from HLA-A and HLA-B loci in five human cell lines sequenced with Illumina MiSeq [33] (ENA study accession: PRJEB4744). The data has been preprocessed for simplicity (*see* **Note 9**).

1. Open the AmpliHLA online submission form (Fig. 8a): http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplihla.
2. Enter a name for the run and, optionally, an email address if you desire to receive the results by email.
3. Choose 'Amplicons' as the analysis 'Data type'.
4. Copy and paste the following link into the 'Sequence file URL' field: http://evobiolab.biol.amu.edu.pl/amplisat/bin/examples/amplihla_example.fq.gz.

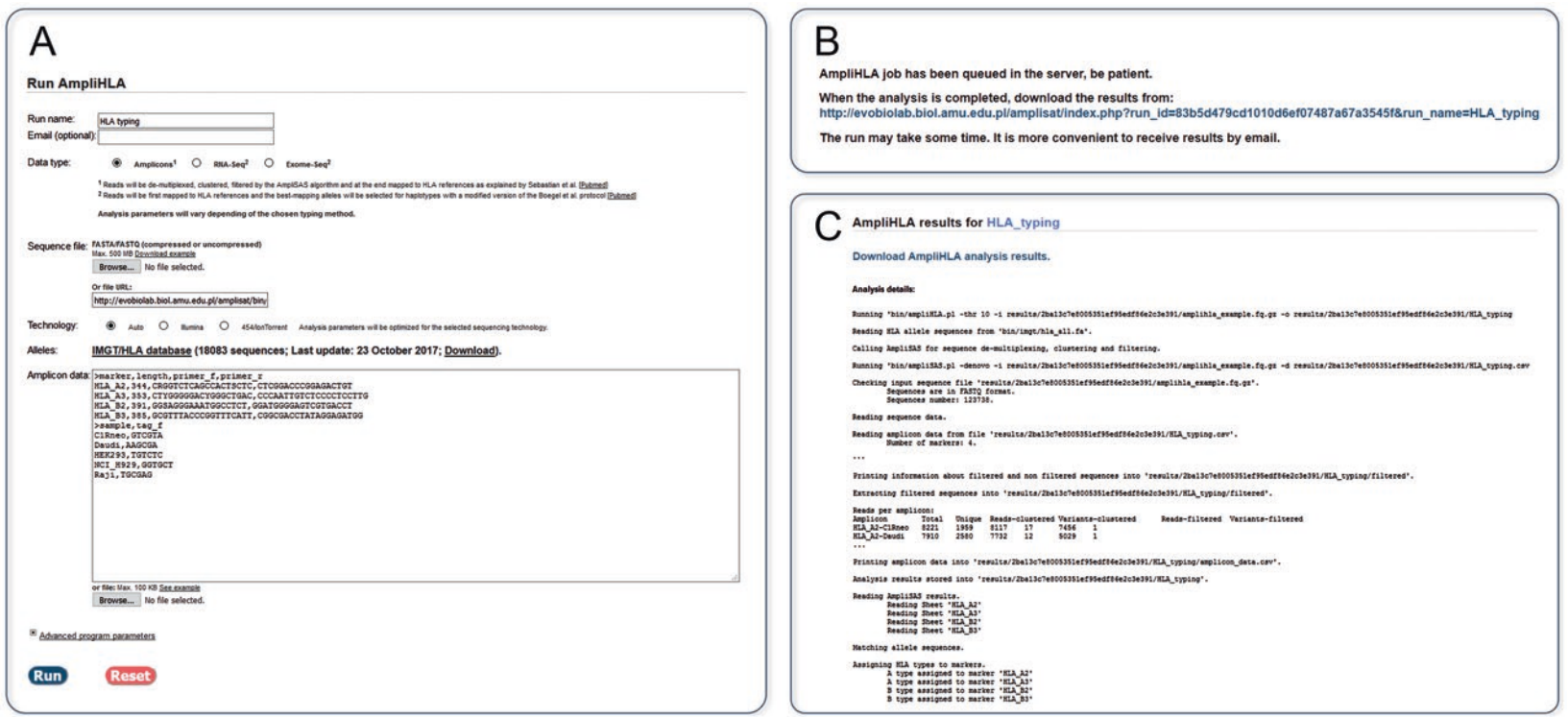


Fig. 8 (**a**) AmpliHLA input form. (**b**) Message with the link to the results page. (**c**) Output summary after analysis completion and link to the results ZIP file

Optionally, you can download the compressed FASTQ file in your computer and upload it with the 'Browse' button.

5. Select 'Auto' from the 'Technology' options, the program will automatically detect the sequencing technology used to adjust the variant clustering parameters and it will calculate the optimum frequency thresholds for each marker to filter artifacts.
6. The field 'Alleles' shows information about the reference sequences used by AmpliHLA to identify the HLA alleles.
7. Copy and paste in the 'Amplicon data' field the following information about primers and tags for the 5 human cell lines sequenced (*see* **Note 6**):

```
>marker,length,primer_f,primer_r
HLA_A2,344,CRGGTCTCAGCCACTSCTC,CTCGGACCCGGAGACTGT
HLA_A3,353,CTYGGGGGACYGGGCTGAC,CCCAATTGTCTCCCCTCCTTG
HLA_B2,391,GGSAGGGAAATGGCCTCT,GGATGGGGAGTCGTGACCT
HLA_B3,385,GCGTTTACCCGGTTTCATT,CGGCGACCTATAGGAGATGG
>sample,tag_f
C1Rneo,GTCGTA
Daudi,AAGCGA
HEK293,TGTCTC
NCI_H929,GGTGCT
Raji,TGCGAG
```

8. Click the 'Run' button and wait until the analysis is completed as explained in **steps 7** and **8** of Subheading 3.1 (Fig. 8b). Keep the results page open, in the next section we will learn how to interpret them.

3.6 Interpreting the HLA Typing Results

1. If we have correctly followed the steps from Subheading 3.5, the AmpliHLA output should look like this (Fig. 8c):

```
AmpliHLA results
Download AmpliHLA analysis results.
Analysis details:
Running 'bin/ampliHLA.pl ...
Reading HLA allele sequences ...
Calling AmpliSAS for sequence de-multiplexing,
clustering and filtering.
Running 'bin/ampliSAS.pl ...
... AMPLISAS OUTPUT ...
Reading AmpliSAS results.
     Reading Sheet 'HLA_A2'
     Reading Sheet 'HLA_A3'
     Reading Sheet 'HLA_B2'
     Reading Sheet 'HLA_B3'
Matching allele sequences.
Assigning HLA types to markers.
     A type assigned to marker 'HLA_A2'
     A type assigned to marker 'HLA_A3'
     B type assigned to marker 'HLA_B2'
     B type assigned to marker 'HLA_B3'
```

2. Click the 'Download AmpliHLA analysis results' link to download a ZIP compressed file including several files and folders as explained in Subheading 3.3, **step 2**.

	C1Rneo	Daudi	HEK293	NCI_H929	Raji
A*03:01			0,4	0,34	0,89
A*02	0,9				
A*01:02		0,54			
A*24:02				0,51	
A*02:01			0,44		
A*66:01 | ...		0,17			

ALLELE	AMBIGUITIES
A*02	A*02:01, A*02:03, A*02:04, A*02:07, A*02:16, A*02:17, A*02:22, A*02:24, A*02:2
A*66:01 | ...	A*25:01, A*26:01, A*26:03, A*43:01, A*66:01

Fig. 9 AmpliHLA Excel output file example. Every sample has assigned one or two HLA alleles, numeric values show the amplicon frequencies of the alleles

3. Open the 'results.xlsx' file to check the assigned genotypes, individuals are shown in columns and alleles in rows, the numeric values represent the average frequencies of the alleles within the amplicons (Fig. 9). Genotypes are given with the highest resolution that can be achieved by the program (maximum 4-digits), it will depend on the number and length of HLA regions (markers) sequenced in the experiment (in the example two exonic regions per locus). Sometimes, a variant shares the same identity with several alleles (e.g., Daudi A*66:01) or the type cannot be resolved with 4-digit resolution (e.g., C1Rneo A*02), then a list of allele ambiguities is listed below the table (Fig. 9). At the left column 'SEQUENCES' there is a list of the variant sequences that match a particular allele.
4. If we compare AmpliHLA typing results with the expected ones in Table 3 validated by Sanger sequencing [33], we observe that AmpliHLA has a 95% of accuracy in assigning genotypes with 2-digit resolution for both loci (the only error is the ambiguous assignment of the A*66:01 Daudi allele, Fig. 9) and 70% of accuracy with 4-digit resolution. Nevertheless, resolution could be improved by sequencing additional regions of the loci or selecting from the ambiguities the most frequent human alleles for the studied population.

3.7 HLA Typing with RNA-Seq Data

AmpliHLA functionality is not restricted to NGS amplicon data, in the present protocol we will analyze an RNA-Seq experiment from the Daudi cell line (ENA run accession SRR387401 from the study SRP009316) [43].

1. Open the AmpliHLA online submission form (Fig. 8a): http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplihla.
2. Enter a name for the run and, optionally, an email address if you desire to receive the results by email.

3. Select 'RNA-Seq' as the 'Data type'.
4. Copy and paste the following link into the 'Sequence/reads file URL' field: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR387/SRR387401/SRR387401_1.fastq.gz and into the 'Paired-end reads file URL' field: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR387/SRR387401/SRR387401_2.fastq.gz.
5. The field 'Alleles' shows information about the reference sequences used by AmpliHLA to identify the HLA alleles.
6. Click the 'Run' button and wait until the analysis is completed as explained in **steps 7** and **8** of Subheading 3.1. The output should look like this:

```
AmpliHLA results
Analysis details:
Running 'bin/ampliHLA.pl ...
Retrieving allele information from BAM data.
Parsing reference file ...
Parsing alignment file ...
74488 reads mapped to 5702 reference alleles.
5702 alleles are kept after filtering artefacts.
RESULTS:
LOCI    ALLELE      SCORE  READS   DISAMBIGUATION
A       A*66:01     48.4    4038
        A*01:02     36.4    3041
B       B*58:01     57.9   16704
        B*58:02     22.4    6466
C       C*03:02     35.2    8417
        C*06:02     33.8    8074
DQA1    DQA1*01:03  94.5    3473
DQB1    DQB1*06:13  50.4    1997
DRB1    DRB1*13:02  62.3    7820
```

7. Typing results are printed in five columns with the following information: HLA locus, assigned alleles, confidence score (allele-specific reads divided by the total number of mapped reads in the locus), corrected allele coverage (see explanation in Subheading 2.2), and allele ambiguities in case the allele has not been unequivocally assigned.
8. Comparing AmpliHLA results with laboratory validated HLA types for the Daudi cell line (Table 3), the genotyping accuracy is of 100% for the MHC class I loci with 4-digit resolution (HLA-A, B and C). Instead, MHC class II loci (HLA-DQA1, DQB1 and DRB1) are typed at 33% of accuracy with 4-digit resolution. If we look at the 2-digit resolution genotypes, all of them have an accuracy of 100%. The difference in accuracy between the class I and class II genotypes is explained because class II genes have only one variable exon and the high similarity among allele sequences makes their genotyping more problematic.

   The previous analysis can be replicated for any pair of FASTQ files from the ENA project SRP009316 (https://

Table 3
Correspondence of human cell lines and HLA types determined by Sanger sequencing

HLA Class I			HLA Class II		
C1Rneo	A*02:01	B*35:03			
Daudi	A*01:02 A*66:01	B*58:01 B*58:02	DQA1*01:02 DQA1*01:03	DQB1*06:02 DQB1*06:04	DRB1*13:01 DRB1*13:02
HEK293	A*02:01 A*03:01	B*07:02			
NCI-H929	A*03:01 A*24:02	B*07:02 B*18:01			
Raji	A*03:01	B*15:10	DQA1*01:01 DQA1*05:01	DQB1*02:01 DQB1*05:01	DRB1*03:01 DRB1*10:01

www.ebi.ac.uk/ena/data/view/SRP009316). For example, the RNA-Seq data from the Raji cell line (ENA run accession SRR387394).

9. Repeat **steps 1**–**7** replacing 'SRR387401' with 'SRR387394' in the reads file URLs at **step 4**. Analysis results will look like this:

```
LOCI   ALLELE      SCORE   READS   DISAMBIGUATION
A      A*03:01     69.3    2024
B      B*15:10     92.5    8318
C      C*04:01     47.9    3248
       C*03:04     29.3    1983
DQA1   DQA1*05:01  63.7    1911
       DQA1*01     36.1    1083    DQA1*01:01,DQA1*01:04
DQB1   DQB1*05:01  55.4    691
       DQB1*02     44.6    556     DQB1*02:01,DQB1*02:02
DRB1   DRB1*10:01  48.6    3700
       DRB1*03:01  48.5    3687
```

10. Comparing retrieved genotypes with Raji expected ones (Table 3), we obtain a 100% of accuracy for the 4-digit class I genotypes and the same for the 2-digit class II ones as previously stated by the Seq2HLA authors after the analysis of the same dataset [44].

3.8 HLA Typing with Exome Sequencing Data

In this final protocol, HLA typing will be performed with whole exome sequencing (WES) data from a Daudi cell line (*see* **Note 10**) [45].

1. Open the AmpliHLA online submission form (Fig. 8a): http://evobiolab.biol.amu.edu.pl/amplisat/index.php?amplihla.
2. Enter a name for the run and, optionally, an email address if you desire to receive the results by email.
3. Select 'Exome-Seq' as the 'Data type'.

4. Copy and paste the following link into the 'Sequence/reads file URL' field: http://evobiolab.biol.amu.edu.pl/amplisat/examples/DAUDI_R1.fq.gz and into the 'Paired-end reads file URL' field: http://evobiolab.biol.amu.edu.pl/amplisat/examples/DAUDI_R2.fq.gz.

   Optionally, you can download the compressed FASTQ file in your computer and upload it with the 'Browse' button.

5. The field 'Alleles' shows information about the reference sequences used by AmpliHLA to identify the HLA alleles.

6. Click the 'Run' button and wait until the analysis is completed as explained in **steps 7** and **8** of Subheading 3.1. The output should look like this:

```
AmpliHLA results
Analysis details:
Running 'bin/ampliHLA.pl ...
Retrieving allele information from BAM data.
Parsing reference file ...
Parsing alignment file ...
543 reads mapped to 1111 reference alleles.
1111 alleles are kept after filtering artifacts.
RESULTS:
LOCI    ALLELE      SCORE  READS    DISAMBIGUATION
A       A*66:01      51.5     69
        A*01:02      47.8     64
B       B*58:01      96.3     78
C       C*06         51.0     49    C*06:02,C*06:04
DQA1    DQA1*01:02   89.2    149
DQB1    DQB1*06:03   90.6     29
DRB1    DRB1*13      91.7     22    DRB1*13:01,DRB1*13:02
```

7. Typing results are printed in five columns as explained in **step 7** at Subheading 3.7.

8. Comparing AmpliHLA results with laboratory validated HLA types for the Daudi cell line (Table 3), the genotypes for the six MHC class I and class II loci evaluated have an accuracy of 92% and 33% for 2-digit and 4-digit resolutions respectively. Accuracy is noticeably lower than in Subheading 3.7 because there are far fewer WES reads mapping to HLA references (dozens) than RNA-Seq reads in the previous protocol (thousands).

4 Notes

1. Basically, there are four main steps in the NGS amplicon sequencing workflow (Fig. 2):

   (a) Design of the primers to amplify the desired gene regions (markers).

(b) Library preparation by PCR amplification of the selected regions, addition of sample-specific DNA tags and of platform-specific sequencing adaptors.

(c) NGS sequencing of the amplification products. The most commonly used platforms are: Illumina, Ion Torrent, and previously 454.

(d) Bioinformatic analysis of the sequencing data. The analysis should include: classification of reads into amplicons, sequencing error correction, filtering of spurious and contaminant reads, and final displaying of results in a human readable way, e.g., an Excel spreadsheet.

(e) For a list of definitions of commonly used terms in amplicon sequencing, see Table 2.

In the following link you will find a video explaining the amplicon sequencing process using NGS in a metagenomics experiment:

http://www.jove.com/video/51709/next-generation-sequencing-of-16s-ribosomal-rna-gene-amplicons.

2. Before NGS technologies were available, PCR products were Sanger sequenced individually. Sanger sequencing is only able to resolve one DNA sequence (allele) per sample. In special cases, a mix of two alleles is also possible (if they differ only by one nucleotide position (i.e., heterozygous individual at given locus). If a primer pair amplifies more than one locus (as it is often the case in MHC genotyping of non-model organisms) the only way to distinct multiple, mixed sequences was to clone sequences to bacterial vectors and further isolation, amplification and sequencing of individual clones. Nevertheless, bacterial cloning is a time-consuming and error prone approach that is only feasible with few dozens of sequences.

   Fortunately, NGS techniques are able to sequence millions of sequences with individual resolution. The combination of amplicon sequencing with NGS allows us to genotype hundreds/thousands of samples in a single experiment. The only requirement is to include different DNA tags to identify the individuals/samples in the experiment. A DNA tag is a short and unique sequence of nucleotides (e.g., ACGGTA) that is either ligated to a PCR product or attached at the end of one of the PCR primers (Fig. 2). Tags have to be unique for each sample/individual to enable assignment of the reads back to the original amplicon (individual or sample) [34, 35].

   However, the NGS techniques have some limitations: the lengths of the sequences are shorter than in Sanger sequencing and frequent sequencing errors result in a high number of artifacts. To alleviate those shortcomings, long sequences can be fragmented and assembled together later by computer and

increasing the depth/coverage ("reading" more times the same sequence) can correct random sequencing errors.

3. Homopolymer regions are a major issue for pyrosequencing and ion semiconductor NGS technologies (454 and Ion Torrent, respectively), where erroneous indels are introduced in high rates. Technology based on reversible dye-terminators (Illumina) suffers from a high number of mostly random substitutions [46–51]. PCR products also incorporate polymerase substitution errors and chimeras (sequences formed from two different sequences due to incomplete primer extension) [52].
4. Four different genotyping approaches were quantitatively evaluated for removing artifacts from NGS amplicon data and assigning MHC class I alleles in a set of sedge warbler individuals [37]. Among the four methods considered, AmpliSAS retrieved accurate, repeatable genotypes requiring lower coverages than the others. Furthermore, AmpliSAS supports different NGS platforms data and it is available as a web server.
5. Usually, an amplicon sequencing experiment sequenced with Illumina technology produces paired-end reads that should be cleaned and merged/overlapped before further processing. In the presented example, both steps were skipped for simplicity and reads are ready-to-use. Paired-end read overlapping and read cleaning were performed with the tools AmpliMERGE and AmpliCLEAN respectively, both are part of the AmpliSAT suite (*see* Subheading 2).
6. If reads have been separated into multiple files after sequencing, one file per amplicon, they can be packed into a single ZIP or TGZ format file and used as input. In such a case the 'amplicon data' field should be empty, AmpliSAS will use the folder and filenames to name the markers and amplicons respectively (Example of organization of reads files into the packaged file:. /MARKER/SAMPLENAME.FASTQ).
7. Adjusting analysis parameters is important because error profiles are affected by many factors: the sequencing platform, length of the amplicon, number of co-amplifying alleles, amplification bias introduced by each set of primers, etc. Specifically, frequency thresholds that separate genuine alleles and technical artifacts may vary between experimental setups, and should be carefully adjusted by the researcher.
8. Sedge warbler MHC class I amplicons have a major length of 241, but there are minor variants of 238 and 235 bp experimentally validated [8]. If we do not specify the lengths, AmpliSAS automatically sets the value to 241 bp and it will not detect the variants with 3 and 6 bp in-frame deletions.
9. The amplicon sequencing data from the five human cell lines that will be used in this example has been pre-processed for

simplicity: paired-end reads have been merged and sample-specific DNA tags have been artificially attached to the forward primers.

10. WES data has been kindly provided by R. Siebert, A. Franke and G. Hemmrich-Stanisak from their original article [45]. To save time in the analysis, the reads have been previously aligned to HLA genomic references with BOWTIE [42] and the mapped reads extracted with 'SamToFastq' command from Picard Tools suite [53]. As a result, only few hundreds of paired-end reads from the initial 34 millions have been saved into two FASTQ files that are used as input in the AmpliHLA procotol.

References

1. Murphy KM, Travers P, Walport M (2007) Janeway's immunobiology, 7th edn. Garland Science, New York
2. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE (2015) The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 43:D423–D431. https://doi.org/10.1093/nar/gku1161
3. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F (2005) Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol 15:1022–1027. https://doi.org/10.1016/j.cub.2005.04.050
4. Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc Biol Sci 277:979–988. https://doi.org/10.1098/rspb.2009.2084
5. Sommer S (2005) The importance of immune gene variability (MHC) in evolutionary ecology and conservation. Front Zool 2:16. https://doi.org/10.1186/1742-9994-2-16
6. Apanius V, Penn D, Slev PR, Ruff LR, Potts WK (1997) The nature of selection on the major histocompatibility complex. Crit Rev Immunol 17:179–224
7. Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J Evol Biol 16:363–377. https://doi.org/10.1046/j.1420-9101.2003.00531.x
8. Biedrzycka A, O'Connor E, Migalska M, Radwan J, Zając T, Bielański W, Solarz W, Ćmiel A, Westerdahl H (2017) Extreme MHC class I diversity in the sedge warbler (Acrocephalus schoenobaenus); selection patterns and allelic divergence suggest that different genes have different functions. BMC Evol Biol 17:159. https://doi.org/10.1186/s12862-017-0997-9
9. Wiseman RW, Karl JA, Bohn PS, Nimityongskul FA, Starrett GJ, O'Connor DH (2013) Haplessly hoping: macaque major histocompatibility complex made easy. ILAR J 54:196–210. https://doi.org/10.1093/ilar/ilt036
10. Sato A, Dongak R, Hao L, Takezaki N, Shintani S, Aoki T, Klein J (2006) Mhc class I genes of the cichlid fish Oreochromis niloticus. Immunogenetics 58:917–928. https://doi.org/10.1007/s00251-006-0151-0
11. Stutz WE, Bolnick DI (2014) Stepwise threshold clustering: a new method for genotyping MHC loci using next-generation sequencing technology. PLoS One 9:e100587. https://doi.org/10.1371/journal.pone.0100587
12. Migalska M, Sebastian A, Konczal M, Kotlík P, Radwan J, Kotlik P, Radwan J (2017) De novo transcriptome assembly facilitates characterisation of fast-evolving gene families, MHC class I in the bank vole (Myodes glareolus). Heredity (Edinb) 118:348–357. https://doi.org/10.1038/hdy.2016.105
13. Figueroa F, Mayer W, Sato A, Zaleska-Rutczynska Z, Hess B, Tichy H, Klein J (2001) Mhc class I genes of swordtail fishes, Xiphophorus: variation in the number of loci and existence of ancient gene families. Immunogenetics 53:695–708. https://doi.org/10.1007/s00251-001-0378-8
14. Mehra NK (2001) Histocompatibility antigens. Encycl Life Sci
15. Trowsdale J, Campbell RD (2001) Mouse MHC genes and products. In: Current protocols in immunology. Wiley, Hoboken, NJ, p Appendix 1L
16. Lukacs MF, Harstad H, Grimholt U, Beetz-Sargent M, Cooper GA, Reid L, Bakke HG, Phillips RB, Miller KM, Davidson WS, Koop BF (2007) Genomic organization of duplicated major histocompatibility complex class I regions in Atlantic salmon (Salmo salar). BMC Genomics 8:251. https://doi.org/10.1186/1471-2164-8-251

17. Kaufman J, Milne S, Göbel TW, Walker BA, Jacob JP, Auffray C, Zoorob R, Beck S (1999) The chicken B locus is a minimal essential major histocompatibility complex. Nature 401:923–925. https://doi.org/10.1038/44856
18. Kelley J, Walter L, Trowsdale J (2005) Comparative genomics of major histocompatibility complexes. Immunogenetics 56:683–695. https://doi.org/10.1007/s00251-004-0717-7
19. Kulski JK, Shiina T, Anzai T, Kohara S, Inoko H (2002) Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. Immunol Rev 190:95–122. doi: imr19008 [pii]
20. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SGE (2013) The IMGT/HLA database. Nucleic Acids Res 41:D1222–D1227. https://doi.org/10.1093/nar/gks949
21. Wallny H-J, Avila D, Hunt LG, Powell TJ, Riegert P, Salomonsen J, Skjødt K, Vainio O, Vilbois F, Wiles MV, Kaufman J (2006) Peptide motifs of the single dominantly expressed class I molecule explain the striking MHC-determined response to Rous sarcoma virus in chickens. Proc Natl Acad Sci U S A 103:1434–1439. https://doi.org/10.1073/pnas.0507386103
22. Livant EJ, Brigati JR, Ewald SJ (2004) Diversity and locus specificity of chicken MHC B class I sequences. Anim Genet 35:18–27
23. Westerdahl H, Wittzell H, von Schantz T, Bensch S (2004) MHC class I typing in a songbird with numerous loci and high polymorphism using motif-specific PCR and DGGE. Heredity (Edinb) 92:534–542. https://doi.org/10.1038/sj.hdy.6800450
24. Sepil I, Moghadam HK, Huchard E, Sheldon BC (2012) Characterization and 454 pyrosequencing of major histocompatibility complex class I genes in the great tit reveal complexity in a passerine system. BMC Evol Biol 12:68. https://doi.org/10.1186/1471-2148-12-68
25. O'Connor EA, Strandh M, Hasselquist D, Nilsson JÅ, Westerdahl H (2016) The evolution of highly variable immunity genes across a passerine bird radiation. Mol Ecol 25:977–989. https://doi.org/10.1111/mec.13530
26. Babik W (2010) Methods for MHC genotyping in non-model vertebrates. Mol Ecol Resour 10:237–251. https://doi.org/10.1111/j.1755-0998.2009.02788.x
27. Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009) New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. Mol Ecol Resour 9:713–719. https://doi.org/10.1111/j.1755-0998.2009.02622.x
28. Radwan J, Zagalska-Neubauer M, Cichoń M, Sendecka J, Kulma K, Gustafsson L, Babik W (2012) MHC diversity, malaria and lifetime reproductive success in collared flycatchers. Mol Ecol 21:2469–2479. https://doi.org/10.1111/j.1365-294X.2012.05547.x
29. Lighten J, van Oosterhout C, Paterson IG, McMullan M, Bentzen P (2014) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (Poecilia reticulata). Mol Ecol Resour 14:753–767. https://doi.org/10.1111/1755-0998.12225
30. Sommer S, Courtiol A, Mazzoni CJ (2013) MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts and allelic dropout. BMC Genomics 14:542. https://doi.org/10.1186/1471-2164-14-542
31. Moonsamy PV, Williams T, Bonella P, Holcomb CL, Höglund BN, Hillman G, Goodridge D, Turenchalk GS, Blake LA, D a D, Simen BB, Hamilton A, May AP, Erlich HA (2013) High throughput HLA genotyping using 454 sequencing and the Fluidigm access Array™ system for simplified amplicon library preparation. Tissue Antigens 81:141–149. https://doi.org/10.1111/tan.12071
32. Erlich RL, Jia X, Anderson S, Banks E, Gao X, Carrington M, Gupta N, DePristo MA, Henn MR, Lennon NJ, de Bakker PIW (2011) Next-generation sequencing for HLA typing of class I loci. BMC Genomics 12:42. https://doi.org/10.1186/1471-2164-12-42
33. Bai Y, Ni M, Cooper B, Wei Y, Fury W (2014) Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. BMC Genomics 15:325. https://doi.org/10.1186/1471-2164-15-325
34. Binladen J, Gilbert MTP, Bollback JP, Panitz F, Bendixen C, Nielsen R, Willerslev E (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS One 2:e197. https://doi.org/10.1371/journal.pone.0000197
35. Meyer M, Stenzel U, Myles S, Prüfer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res 35:e97. https://doi.org/10.1093/nar/gkm566
36. Lighten J, van Oosterhout C, Bentzen P (2014) Critical review of NGS analyses for de novo genotyping multigene families. Mol Ecol 23:3957–3972. https://doi.org/10.1111/mec.12843
37. Biedrzycka A, Sebastian A, Migalska M, Westerdahl H, Radwan J (2017) Testing geno-

typing strategies for ultra-deep sequencing of a co-amplifying gene family: MHC class I in a passerine bird. Mol Ecol Resour 17:642–655. https://doi.org/10.1111/1755-0998.12612

38. Sebastian A, Herdegen M, Migalska M, Radwan J (2016) Amplisas: a web server for multilocus genotyping using next-generation amplicon sequencing data. Mol Ecol Resour 16:498–510. https://doi.org/10.1111/1755-0998.12453
39. Leclaire S, Strandh M, Mardon J, Westerdahl H, Bonadonna F (2017) Odour-based discrimination of similarity at the major histocompatibility complex in birds. Proceedings Biol Sci 284:20162466. https://doi.org/10.1098/rspb.2016.2466
40. Pardal S, Drews A, Alves JA, Ramos JA, Westerdahl H (2017) Characterization of MHC class I in a long distance migratory wader, the Icelandic black-tailed godwit. Immunogenetics 69:463–478. https://doi.org/10.1007/s00251-017-0993-7
41. Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J, Boisguérin V, Türeci O, Diken M, Castle JC, Sahin U (2013) HLA typing from RNA-seq sequence reads. Genome Med 4:102. https://doi.org/10.1186/gm403
42. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. https://doi.org/10.1186/gb-2009-10-3-r25
43. Boegel S, Löwer M, Bukur T, Sahin U, Castle JC (2014) A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. Oncoimmunology 3:e954893. https://doi.org/10.4161/21624011.2014.954893
44. Boegel S, Scholtalbers J, Löwer M, Sahin U, Castle JC (2015) In silico HLA typing using standard RNA-seq sequence reads. In: Bugert P (ed) Molecular typing of blood cell antigens, Methods in molecular biology. Springer, New York, pp 115–121
45. Kreck B, Richter J, Ammerpohl O, Barann M, Esser D, Petersen BS, Vater I, Murga Penas EM, Bormann Chung CA, Seisenberger S, Lee Boyd V, Smallwood S, Drexler HG, Macleod RAF, Hummel M, Krueger F, Häsler R, Schreiber S, Rosenstiel P, Franke A, Siebert R (2013) Base-pair resolution DNA methylome of the EBV-positive endemic Burkitt lymphoma cell line DAUDI determined by SOLiD bisulfite-sequencing. Leukemia 27:1751–1753. https://doi.org/10.1038/leu.2013.4
46. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB (2013) Characterizing and measuring bias in sequence data. Genome Biol 14:R51. https://doi.org/10.1186/gb-2013-14-5-r51
47. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012:251364. https://doi.org/10.1155/2012/251364
48. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30:434–439. https://doi.org/10.1038/nbt.2198
49. Vandenbroucke I, Van Marck H, Verhasselt P, Thys K, Mostmans W, Dumont S, Van Eygen V, Coen K, Tuefferd M, Aerssens J (2011) Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications. BioTechniques 51:167–177. https://doi.org/10.2144/000113733
50. Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin J-F (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 12:245. https://doi.org/10.1186/1471-2164-12-245
51. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: characterising errors in ion torrent PGM data. PLoS Comput Biol 9:e1003031. https://doi.org/10.1371/journal.pcbi.1003031
52. Potapov V, Ong JL (2017) Examining sources of error in PCR by single-molecule sequencing. PLoS One 12:e0169774. https://doi.org/10.1371/journal.pone.0169774
53. Broad Institiute (2014) Picard tools: Java-based command-line utilities for manipulating SAM files

Chapter 19

HLA Haplotype Frequency Estimation from Real-Life Data with the Hapl-o-Mat Software

Jürgen Sauter, Christian Schäfer, and Alexander H. Schmidt

Abstract

HLA haplotype frequencies are of use in a variety of settings. Such data is typically derived either from family pedigree data by targeted typing or statistical analysis of large population-specific genotype samples. As established tools for the latter approach lacked ability to treat the amount, ambiguity, and inhomogeneity found in genotype data in hematopoietic stem cell donor registries, we developed Hapl-o-Mat to alleviate these specific shortcomings.

Key words HLA, Immunogenetics, Population genetics, Bioinformatics, Haplotype, Expectation-maximization algorithm, Open-source software

1 Introduction

Knowledge of human leukocyte antigen (HLA) haplotypes and their respective frequencies is helpful, for example, in disease association studies [1] and population genetics [2]. In the context of unrelated hematopoietic stem cell transplantation (HSCT), population-specific HLA haplotype frequencies are of particular interest in strategic donor registry planning [3, 4] and for individual donor searches using advanced HLA matching algorithms [5–7]. However, the estimation of haplotype frequencies from HLA genotyping data is typically challenged by the large amount of genotype data, the complex HLA nomenclature [8], and the heterogeneous and ambiguous nature of typing records. To meet these challenges, we have developed the open-source software Hapl-o-Mat [9]. It estimates haplotype frequencies from population data including an arbitrary number of loci using an expectation-maximization algorithm [10–13]. Its key features are the processing of different HLA typing resolutions within a given population sample and the handling of ambiguities recorded via multiple allele codes [14] or genotype list strings (GLS) [15]. The software was successfully validated using former

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3_19,

implementations as well as artificial populations with known haplotype frequency distributions.

Hapl-o-Mat is freely available on GitHub (https://github.com/DKMS/Hapl-o-Mat). In this Chapter, we provide a detailed step-by-step guide on the usage of Hapl-o-Mat in a Linux environment. Starting by downloading the source code of the software, we elucidate installation and preparation of the tool and conclude with a demonstration sample for genotypes containing multiple allele codes. The software itself is not limited to a single operation system: It can be compiled in a Microsoft Windows environment as well. Instructions for operation systems different than Linux are included in the GitHub repository.

2 Materials

2.1 Installation

2.1.1 Prerequisites

To follow this tutorial, a Linux system, a C++ compiler supporting C++11, Python, and a working internet connection are required. The tutorial is based on Ubuntu 14.04.4 LTS and GNU compiler collection (GCC) version 4.8.4. This Ubuntu version comes with Python. Every step is processed in a terminal window. Basic knowledge of Linux terminal operation is recommended.

2.1.2 Download

Download Hapl-o-Mat into your Linux home directory (or any other directory of your choice) by typing

```
> git clone https://github.com/DKMS/Hapl-o-Mat.git
```

Change to the new directory containing Hapl-o-Mat by entering

```
> cd Hapl-o-Mat
```

Check what is inside by using the command

```
Hapl-o-Mat> ls
```

The directory should contain the files as listed in Table 1 where important files for using Hapl-o-Mat are marked in boldface.

To estimate haplotype frequencies, only the folder *prepareData* and the files *Makefile*, *parametersGLS*, *parametersGLSC*, *parametersMAC*, and *parametersREAD* need to be considered.

2.1.3 Compilation

Hapl-o-Mat is compiled using GCC and a Makefile. To compile, type

```
Hapl-o-Mat> make
```

to create the executable *haplomat*.

2.2 Data Preparation and First Run

Hapl-o-Mat relies on information on the HLA nomenclature. As the HLA nomenclature evolves over time it is important to update data from time to time. Hapl-o-Mat relies for this information on

Table 1
Content of Hapl-o-Mat root directory

File name	Description
COPYING	The GNU general public license
detailedGettingStartedLinux	Guide for using Hapl-o-mat under Linux
detailedGettingStartedWindows	Guide for using Hapl-o-mat under windows
examplePopulations	Some genotype population data we are going to work with in the section tutorials
gettingStarted	A shorter form of this tutorial
Include	A part of Hapl-o-Mat's source code; if you do not want to change code, do not touch it
Makefile	Instructions for building Hapl-o-mat; you might need to adapt it, if you use another compiler than GCC
parametersGLS, parametersGLSC, parametersMAC, parametersREAD	Parameter files for Hapl-O-mat; we are going to discuss these in the section parameters
prepareData	Here is everything to create the data required by Hapl-o-mat
README.md	Read me
Src	Another part of Hapl-o-Mat's source code; if you do not want to change code, do not touch it
systemTest	Run the system test after changing code to for code validity. Refer to its README
textsForGettingStarted	Raw files for the guides including this guide

a number of files (*see* Table 2), which must be placed in the folder *Hapl-o-Mat/data* for Hapl-o-Mat to work.

Hapl-o-Mat comes with an automated script building these nomenclature files. To do so, type

```
Hapl-o-Mat> cd prepareData
```

and start the script with

```
Hapl-o-Mat/prepareData> python BuildData.py
```

This command downloads all relevant data, processes them, and moves the created files to the folder *Hapl-o-Mat/data*.

For manual data compilation, follow the short instructions in *Hapl-o-mat/prepareData/README* or the detailed version in *Hapl-o-mat/prepareData/detailedExplanationManuallyPrepareData.pdf*.

After the completion of the script, return to the main Hapl-o-Mat folder by typing

Table 2
Files defining the HLA nomenclature for Hapl-o-Mat

File name	Description
AllAllelesExpanded.Txt	A list of relevant existing HLA alleles with their enclosed more-digit typing resolutions
AlleleList.Txt	If your input data in GLS format includes a missing single-locus genotype, it can be replaced by combining all alleles of the same locus from this file. You only must create it in this case
Ambiguity.Txt	Data for the ambiguity filter
LargeG.Txt	A list of G-groups [8] with their enclosed alleles in 8-digit resolution
MultipleAlleleCodes.Txt	A list of multiple allele codes and their translation to alleles in 4-digit resolution
P.Txt	A list of P-groups [8] with their enclosed alleles in 8-digit resolution
Smallg.Txt	A list of g-groups [16] with their enclosed alleles in 8-digit resolution

```
Hapl-o-Mat/prepareData> cd ..
```

To check whether the installation process has been completed successfully, Hapl-o-Mat can be started for the first time by entering

```
Hapl-o-Mat> ./haplomat MAC
```

3 Methods

3.1 Haplotype Frequency Estimation

After successful setup, Hapl-o-Mat is ready for haplotype frequency estimation. This section illustrates the generic workflow and explains the detailed steps for an individual analysis.

3.1.1 Workflow Overview

1. Build or update the data comprising information on the HLA nomenclature using the python script *Hapl-o-Mat/prepareData/BuildData.py*.
2. Prepare the genotype population data to be analyzed. Identify how genotyping ambiguities are recorded (MAC or GLS, see section Genotype Data Formats and Table 3) and choose the input format accordingly. Adapt the format of input data, e.g., include the header line or separate alleles by tabulator.
3. Set parameters in the parameter file corresponding to your input format (*see* Table 3).
4. Copy the executable *haplomat*, the folder *data*, the parameter file, and the input population data into one specific folder. Create any folders you specified in the parameter file. All the other files are not needed to run Hapl-o-Mat but for setup only.

5. Run Hapl-o-Mat in the specific folder.

Building and updating of HLA nomenclature data were explained in the previous chapter. To illustrate usage, this tutorial makes use of example data included in the initial Hapl-o-Mat download. The example data is contained in the folder *Hapl-o-Mat/examplePopulations*. For all formats, three-locus (HLA-A, -B, -DRB1) haplotypes are going to be estimated from this data.

3.1.2 Genotype Data Formats

Hapl-o-Mat estimates haplotypes from population genotype data. It supports different formats of recording genotype data. To use Hapl-o-Mat, data needs to be in one of the formats listed in Table 3.

3.1.3 Parameters

Each input format for genotype data requires a different set of parameters. The parameters are saved in the corresponding files *parametersMAC*, *parametersGLSC*, *parametersGLS*, and

Table 3
Input data format options

Data format	Description
MAC	**M**ultiple **a**llele **c**odes: Ambiguities are encoded by multiple allele codes (MAC). Except for the first line, input files hold an individual's identification number and genotype per line. Genotypes are saved allele by allele without locus name. Identification number and alleles are TAB-separated. The first line of the file is a header line indicating the name of the first column and the loci of the other columns. Alleles from identical loci must be placed next to each other. For an example refer to "Hapl-o-mat/examplePopulations/populationData_a.Dat."
GLSC	**G**enotype **l**ist **s**trings **c**olumn-wise: Genotypes with or without ambiguities are saved as genotype list strings [15]. Input files hold an individual's identification number and genotype per line. Identification number and single-locus genotypes are TAB-separated. For an example refer to "Hapl-o-mat/examplePopulations/populationData_b.Dat"
GLS	**G**enotype **l**ist **s**trings: Genotypes with or without ambiguities are saved as genotype list strings (GLS). Population data is saved in two files. The pull file contains an individual's identification number and a list of integer numbers, so-called GLS-ids, referring to her or his single-locus genotype. The GLS-ids are separated from the identification number via ";" and from each other via ":". The second file, the glid file, contains a translation from GLS-ids starting with "1" to actual single-locus genotypes. GLS-id and genotype are separated via ";". A GLS-id of "0" is interpreted as a missing typing at the corresponding locus and does not require a translation in the glid file. For an example refer to "Hapl-o-mat/examplePopulations/populationData_c.Pull" and "Hapl-o-mat/examplePopulations/populationData_c.Glid"
READ	**READ**: Ambiguities are completely resolved and alleles are already translated to the desired typing resolutions. The input data is formatted in the same way as Hapl-o-mat records processed genotype data. This allows for easily repeating a run without the need to resolve genotype data again

Table 4
Default input parameters for Hapl-o-Mat

Parameter	Description
FILENAME_HAPLOTYPES	Name of the file which temporarily saves haplotype names
FILENAME_GENOTYPES	Name of the file which saves resolved genotypes
FILENAME_ HAPLOTYPEFREQUENCIES	Name of the file which saves haplotypes and estimated haplotype frequencies
FILENAME_EPSILON_LOGL	Name of the file which saves stopping criterion and log-likelihood per iteration
INITIALIZATION_ HAPLOTYPEFREQUENCIES	Initialization routine for haplotype frequencies. It takes the following values: • "Equal": All haplotype frequencies are initialized with the same frequency, the inverse number of haplotypes • "numberOccurrence": Haplotype frequencies are initialized according to the number of occurrence of haplotypes in the initial genotype analysis • "Random": Haplotype frequencies are initialized randomly • "Perturbation": Haplotype frequencies are initialized as in numberOccurrence and then randomly modified by a small (<10%) positive or negative offset
EPSILON	Value for the stopping criterion. The algorithm continues as long as the maximal change between consecutive haplotype frequency estimations is larger than the assigned value
CUT_HAPLOTYPEFREQUENCIES	Estimated haplotype frequencies smaller than this value are removed from the output
RENORMALIZE_ HAPLOTYPEFREQUENCIES	Takes values "true" and "false". If "true", estimated haplotype frequencies are normalized to sum up to 1. Within machine precision, this becomes necessary if estimated haplotypes are removed, e.g., via the option CUT_HAPLOTYPEFREQUENCIES
SEED	Sets the seed of the used pseudo random number generator. If set to "0", the seed is initialized by the system time

parametersREAD. All input formats have the parameters in common. These are listed in Table 4.

Depending on the input format, additional parameters are required. These are noted in Table 5.

Whenever specifying a file name including folders, these folders have to be created before running Hapl-o-Mat.

3.2 Example: Input Format MAC

This example is based on the population data in the folder *Hapl-o-Mat/examplePopulations/populationData_a.dat*. As ambiguities are recorded as multiple allele codes, the input format is MAC.

Table 5
Format-dependent input parameters for Hapl-o-Mat

Parameter	Input format	Description
FILENAME_INPUT	MAC, GLSC, READ	The file name of the input population data
FILENAME_PULL	GLS	The file name of the pull file
FILENAME_GLID	GLS	The file name of the glid file
LOCI_AND_RESOLUTIONS	MAC, GLS, GLSC	Loci included into analysis and desired typing resolution per locus; the list is separated by "," and contains the locus names followed by ":" and the desired typing resolution, e.g., A:g,B:4d,C:g. Supported typing resolutions and their abbreviations are g-groups (g), P-groups (P), G-groups (G), 2-digit fields (2d), 4-digit fields (4d), 6-digit fields (6d), and 8-digit fields (8d)
LOCIORDER	GLS	Specify the order of loci the individual's GL-ids correspond to. Loci are separated via ","
RESOLVE_MISSING_GENOTYPES	GLS	Takes values "true" and "false." If set to true, a missing typing is replaced by a combination of all alleles from AlleleList.Txt at the locus. Else, individuals with a missing typing are discarded from analysis
MINIMAL_FREQUENCY_GENOTYPES	MAC, GLS, GLSC	Genotypes which split into more genotypes than the inverse of this number are discarded from analysis
DO_AMBIGUITYFILTER	MAC, GLS, GLSC	Takes values "true" and "false." The option "true" activates the ambiguity filter
EXPAND_LINES_AMBIGUITYFILTER	MAC, GLS, GLSC	Takes values "true" and "false." If set to "true," matching lines with additional genotype pairs in the ambiguity filter are considered

3.2.1 Preparations

Enter the folder *Hapl-o-Mat/examplePopulations* by typing

```
Hapl-o-Mat> cd examplePopulations
```

Create a folder named "a" with

```
Hapl-o-Mat> mkdir a
```

Enter the folder by using the command

```
Hapl-o-Mat> cd a
```

Then provide the data required by Hapl-o-Mat by copying the folder "Hapl-o-Mat/data" to "a" with

```
Hapl-o-Mat/a> cp –r ../../data .
```

Additionally, copy the executable *haplomat* and the file *parametersMAC* to folder *a* by typing

```
Hapl-o-Mat/a> cp ../../haplomat ../../parametersMAC .
```

Check that everything is in the folder *a* by typing

```
Hapl-o-Mat/a> ls
```

The output should contain a directory *data*, an executable *haplomat*, and the file *parametersMAC*.

3.2.2 Parameters

According to the format of the input genotype data, the parameter file is *parametersMAC*. Open it in the text editor *gedit* (however, any text editor of your choice and convenience suffices) with

```
Hapl-o-Mat/a> gedit parametersMAC
```

and set the following values:

```
#file names
FILENAME_INPUT=../populationData_a.dat
FILENAME_HAPLOTYPES=run/haplotypes.dat
FILENAME_GENOTYPES=run/genotypes.dat
FILENAME_HAPLOTYPEFREQUENCIES=run/hfs.dat
FILENAME_EPSILON_LOGL=run/epsilon.dat
#reports
LOCI_AND_RESOLUTIONS=A:g,B:g,DRB1:g
MINIMAL_FREQUENCY_GENOTYPES=1e-5
DO_AMBIGUITYFILTER=false
EXPAND_LINES_AMBIGUITYFILTER=false
#EM-algorithm
INITIALIZATION_HAPLOTYPEFREQUENCIES=perturbation
EPSILON=1e-6
CUT_HAPLOTYPEFREQUENCIES=1e-6
RENORMALIZE_HAPLOTYPEFREQUENCIES=false
SEED=1000
```

Save the file by using [CTRL] + [C] and close the editor with [CTRL] + [Q].

Finally, create the directory *run* by typing

```
Hapl-o-Mat/a> mkdir run
```

3.2.3 Run Hapl-o-Mat

Compute haplotype frequencies from the genotype input data by running Hapl-o-Mat.via

```
Hapl-o-Mat/a> ./haplomat MAC
```

Hapl-o-Mat reports parameters, statistics on the resolved genotype data and the expectation-maximization algorithm, and the run time. This output can be saved in an extra file *Log.txt* by running Hapl-o-Mat via

```
Hapl-o-Mat/a> ./haplomat MAC > Log.txt
```

3.2.4 Results

To examine results, change into the directory *run* by

```
Hapl-o-Mat/a> cd run
```

Check for directory content by typing

```
Hapl-o-Mat/a/run> ls
```

The folder should contain three files named *hfs.dat*, *genotypes.dat*, and *epsilon.dat*. The main result file reporting haplotype frequencies is "hfs.dat" and can be examined using the text editor *gedit* by typing

```
Hapl-o-Mat/a/run> gedit hfs.dat
```

or any other text editor. The file contains two columns. The first column reports haplotypes, the second the associated frequencies as determined by Hapl-o-Mat. Haplotypes are saved in the GLS format. The file is sorted by descending haplotype frequencies.

The file *genotypes.dat* contains resolved genotypes. It can be examined using the text editor *gedit* by typing

```
Hapl-o-Mat/a/run> gedit genotypes.dat
```

or any other text editor. The first column corresponds to the individual's identification number as used in the input file. The second column indicates how ambiguities per single-locus genotypes have been resolved. If no ambiguities occurred or no additional genotypes were formed, an "N" is stated in the column. If an ambiguity occurred and was resolved by building all possible allele combinations, the type is I. Activating the ambiguity filter leads to additional types: A, if one matching line in the ambiguity file was found; and M, if multiple matching lines were found. The third column shows the genotype frequencies and the fourth column the genotypes themselves. The genotypes are saved in the GLS format. If an individual's genotype splits into a set of genotypes, each genotype is written to one line starting with the same identification number. The corresponding frequencies become non-integer and sum up to 1.

The file *epsilon.dat* reports on the evolution of the stopping criterion and log-likelihood while iterating expectation and maximization steps. It can be examined using the text editor *gedit* by typing

```
Hapl-o-Mat/a/run> gedit epsilon.dat
```

or any other text editor. The first column states the maximal change between consecutive haplotype frequency estimations. This change is compared to the stopping criterion set in the respective parameter file. The second column is the non-normalized log-likelihood.

References

1. Crawford DC, Nickerson DA (2005) Definition and clinical importance of haplotypes. Annu Rev Med 56:303–320. https://doi.org/10.1146/annurev.med.56.082103.104540
2. Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. Am J Hum Genet 60(4):772–789
3. Hurley CK, Fernandez Vina M, Setterholm M (2003) Maximizing optimal hematopoietic stem cell donor selection from registries of unrelated adult volunteers. Tissue Antigens 61(6):415–424
4. Schmidt AH, Sauter J, Pingel J, Ehninger G (2014) Toward an optimal global stem cell donor recruitment strategy. PLoS One 9(1):e86605. https://doi.org/10.1371/journal.pone.0086605
5. Eberhard HP, Feldmann U, Bochtler W, Baier D, Rutt C, Schmidt AH, Muller CR (2010) Estimating unbiased haplotype frequencies from stem cell donor samples typed at heterogeneous resolutions: a practical study based on over 1 million German donors. Tissue Antigens 76(5):352–361. https://doi.org/10.1111/j.1399-0039.2010.01518.x
6. Steiner D (2012) Computer algorithms in the search for unrelated stem cell donors. Bone Marrow Res 2012:175419. https://doi.org/10.1155/2012/175419
7. Bochtler W, Gragert L, Patel ZI, Robinson J, Steiner D, Hofmann JA, Pingel J, Baouz A, Melis A, Schneider J, Eberhard HP, Oudshoorn M, Marsh SG, Maiers M, Muller CR (2016) A comparative reference study for the validation of HLA-matching algorithms in the search for allogeneic hematopoietic stem cell donors and cord blood units. Hla 87(6):439–448. https://doi.org/10.1111/tan.12817
8. Marsh SGE, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Fernandez-Vina M, Geraghty DE, Holdsworth R, Hurley CK, Lau M, Lee KW, Mach B, Maiers M, Mayr WR, Muller CR, Parham P, Petersdorf EW, Sasazuki T, Strominger JL, Svejgaard A, Terasaki PI, Tiercy JM, Trowsdale J (2010) Nomenclature for factors of the HLA system, 2010. Tissue Antigens 75(4):291–455. https://doi.org/10.1111/j.1399-0039.2010.01466.x
9. Schäfer C, Schmidt AH, Sauter J (2017) Hapl-o-mat: open-source software for HLA haplotype frequency estimation from ambiguous and heterogeneous data. BMC Bioinformatics 18(1):284. https://doi.org/10.1186/s12859-017-1692-y
10. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Series B (Methodological) 39(1):1–38
11. Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12(5):921–927
12. Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56(3):799–810
13. Polańska J (2003) The EM algorithm and its implementation for the estimation of frequencies of SNP-haplotypes. Int J Appl Marth Comp Sci 13(3):419–429
14. NMDP (2016) Allele code lists. National marrow donor program. https://bioinformatics.bethematchclinical.org/HLA-Resources/Allele-Codes/Allele-Code-Lists/. 2016
15. Milius RP, Mack SJ, Hollenbach JA, Pollack J, Heuer ML, Gragert L, Spellman S, Guethlein LA, Trachtenberg EA, Cooley S, Bochtler W, Mueller CR, Robinson J, Marsh SGE, Maiers M (2013) Genotype list string: a grammar for describing HLA and KIR genotyping results in a text string. Tissue Antigens 82(2):106–112. https://doi.org/10.1111/tan.12150
16. Schmidt AH, Baier D, Solloch UV, Stahr A, Cereb N, Wassmuth R, Ehninger G, Rutt C (2009) Estimation of high-resolution HLA-A, -B, -C, -DRB1 allele and haplotype frequencies based on 8862 German stem cell donors and implications for strategic donor registry planning. Hum Immunol 70(11):895–902. https://doi.org/10.1016/j.humimm.2009.08.006

INDEX

Sebastian Boegel (ed.), *HLA Typing: Methods and Protocols*, Methods in Molecular Biology, vol. 1802,
https://doi.org/10.1007/978-1-4939-8546-3,

G

H

I

K

L

M

N