DISCOVERING GENETIC DRIVERS IN ACUTE GRAFT-VERSUS-HOST DISEASE AFTER ALLOGENEIC HEMATOPOIETIC STEM CELL TRANSPLANTATION

A DISSERTATION
SUBMITTED TO THE FACULTY OF
THE UNIVERSITY OF MINNESOTA
BY

HU HUANG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ADVISORS:

CALEB KENNEDY, PH.D.
CLAUDIA NEUHAUSER, PH.D.

MAY, 2019

Acknowledgements

Completion of a Ph.D. requires one's motivation, dedication, and determination. More importantly, it needs a network of support and guidance from academics, mentors, family, and friends. During the past six years, I was lucky enough to have such a fantastic and generous group of people.

First and foremost, I am deeply indebted to my academic advisor Dr. Claudia Neuhauser for all the intellectual and emotional support, and encouragement she gave me since the beginning of my graduate study, especially during the tough times in the Ph.D. pursuit. Without her guidance and motivation, it would not have been possible to achieve what I accomplished today.

I would also like to extend my deepest gratitude to my research advisor, Dr. Caleb Kennedy. The last four years at the National Marrow Donor Program (NMDP) has been a fantastic experience, and I thank Dr. Kennedy wholeheartedly, not only for his tremendous academic support but also for his continuous encouragement, profound belief in my abilities and giving me the freedom to explore many creative ideas.

Special thanks go to Dr. Hongbo Xie at the Queensland University of Technology (formerly at Jiangsu University), who encouraged me to embark on the path of machine learning applications in biomedical science and engineering research, and provided me

with a fantastic training in scientific research and writing. My intellectual pursuits and American dream were largely inspired by him.

I would like to express my most profound appreciation to my preliminary and final exam committee, Dr. Abeery Madbouly, and Dr. Chad Myers, for their time, interest, helpful comments, and productive discussions. I gratefully acknowledge the assistance of Dr. Madbouly from the Center for International Blood and Marrow Transplant Research (CIBMTR) throughout my research at NMDP.

The members of the NMDP Bioinformatics Research Group and CIBMTR have contributed immensely to my personal and professional growth during the past four years. Special thanks to Wei Wang, Michael Halagan, Jason Brelsford, Pradeep Bashyal, Stephanie Fingerson, Julia Udell, Stephen Spellman, Cynthia Vierra-Green, Colleen Brady, Jane Pollack, Debra Turner, and Jayesh Iyer, for their technical support, challenging discussions, valuable advice and collaboration, and great friendships.

I gratefully acknowledge the funding sources that made my Ph.D. work possible. The University of Minnesota, Bioinformatics and Computational Biology (BICB) Program provided Graduate Fellowship for my first year and the NMDP/CIBMTR and BICB program funded the rest of my Ph.D. research.

My time at the University of Minnesota, Twin Cities was made colorful and enjoyable thanks to the many friends who came into my life. I am grateful to Praveen Kumar and Madhu Kumari for making stressful times fun, comforting and memorable. To Dr. Pajau

Vangay for academic and emotional support whenever I needed the most. To my old roommate Dr. Youssef Roman, for the inspiring stories and motivational conversations. To Jeremy Dowd for introducing me to the local communities through volunteering opportunities, especially connecting with the GLBTQ community. Those were the precious memories and experiences that made my graduate life in Twin Cities extremely special.

Lastly, I would like to thank my parents who raised me with grit and supported me in all my academic pursuits. Without their relentless support, I would not have been able to fulfill my dream. And most of all, for my loving, supportive, encouraging, and patient fiancé, Tom, who made the stressful graduate life fun and exciting with amazing National Parks adventures. His unwavering support during the final stages of this Ph.D. is so much appreciated.

It has been a marvelous ride!

Hu Huang

University of Minnesota, Twin Cities

May 2019

To my love, Tom,

who made this journey extra special and exciting!

Abstract

Acute graft-versus-host disease (GVHD) is one of the major complications after allogeneic hematopoietic stem cell transplantation (allo-HCT) that cause non-relapse morbidity and mortality. Although the increasing matching rate of the human leukocyte antigen (HLA) genes between donor and recipient (DR) has significantly reduced the risk of GVHD, clinically significant GVHD remains as a transplantation challenge, even in HLA-identical transplants. Candidate gene studies and genome-wide association studies have revealed susceptible individual genes and gene pairs from DR pairs that are associated with acute GVHD; however, the roles of genetic disparities between donor and recipient remain to be understood.

To identify genetic factors linked to acute GVHD, we investigated the classical HLA and non-HLA genes and conducted a genome-wide clinical outcome association study. Assessment of 4,646 antigen recognition domain (ARD)-matched unrelated donor allo-HCT cases showed that the frequency of mismatches outside the ARD in HLA genes is very low when the DR pairs are matched at ARD. Due to the low frequency of amino acid mismatches in the non-ARD region and their reportedly weak alloimmune reactions, we suggest that the non-ARD sequence mismatches within the ARD-matched DR pairs have limited influence on the development of acute GVHD, and may not be a primary factor. The genome-wide clinical outcome association study between DR pairs observed multiple autosomal minor histocompatibility antigens (MiHAs) restricted by HLA typing, though their association with acute GVHD outcome was not statistically significant. This

result suggests that HLA mismatching outweighs other genetic mismatches as contributors to acute GVHD risk. In the cases of female donors to male recipients, we identified the significant association of the Y chromosome-specific peptides encoded by *PCDH11Y*, *USP9Y*, *UTY*, and *NLGN4Y* with the acute GVHD outcome.

Additionally, we developed a machine learning-based genetic variant selection algorithm for ultra-high dimensional transplant genomic studies. The algorithm successfully selected a set of genes from over 1 M genetic variants, all of which have evidence to be linked to the transplant-related complications.

This work offers evidence and guidance for further research in acute GVHD and allo-HCT and provides useful bioinformatics and data mining tools for transplant genomic studies.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

Allogeneic hematopoietic stem cell transplantation (allo-HCT) has been widely used as a curative treatment for a series of hematologic malignancies and inherited genetic diseases. Bone marrow is an organ that is rich in hematopoietic stem cells (HSCs) and was used as the primary source of donor stem cells in allo-HCT until the early 2000s (Griffioen, van Bergen, and Falkenburg 2016). Peripheral blood-derived stem cells (PBSCs) have become a popular alternative of HSCs since its first published report (Goldman et al. 1978). Especially in the autologous setting, PBSCs are preferred over bone marrow due to ease of collection and quick engraftments (Cutler and Antin 2001). Although PBSCs were more cautiously monitored in allo-HCTs, the number of allogeneic PBSC transplant cases has increased dramatically since the report of mobilized PBSCs through Granulocyte colony-stimulating factor (G-CSF) or Granulocyte-macrophage colony-stimulating factor (GM-CSF) (Schmitz et al. 1995; Welniak, Blazar, and Murphy 2007) and the reports of the similar transplant outcomes using PBSCs to that of bone marrow stem cells. The use of PBSCs in allo-HCT exceeded that of the allo-HCT cases with bone marrow stem cells since 2014 (D'Souza and Zhu 2017). Another rich source of HSCs is umbilical cord blood (UCB), which has gained popularity as an attractive graft source (Benito et al. 2004; Sullivan 2008). UCB derived HSCs are immunologically naive, resulting in attenuated donor-derived immune response compared to bone marrow stem cells (Sullivan 2008). Multiple retrospective studies have suggested that transplant cases that use UCB-derived stem cells yield similar transplant outcomes to that of the bone marrow transplant cases, supporting a broader applications of UCB

stem cells and the promotion of cord blood banking (Eapen et al. 2007; Barker and Wagner 2003; Hwang et al. 2007; Ballen, Gluckman, and Broxmeyer 2013; Lou, Zhao, and Chen 2018).

The most crucial criterion in allo-HCT is to identify the tissue compatibilities through the major histocompatibility complex (MHC) between a potential donor and the recipient to avoid graft rejection. More specifically, an ideal graft source would have identical human leukocyte antigen (HLA) genes as the recipient. The HLA gene system is one of the most extensively studied human gene regions due to its high polymorphism (Robinson et al. 2015; Gourraud et al. 2014; Horton et al. 2004; Xie et al. 2003) and its critical role in allogeneic solid organ transplant (liver, kidney etc.) and allo-HCT (Petersdorf 2013). The HLA region has also been used extensively to study population diversity (Sanchez-Mazas and Meyer 2014), the evolutionary history of human ancestry (Gourraud et al. 2014; Uinuk-Ool, Takezaki, and Klein 2003), the association with multiple immunodeficiency diseases, such as HIV/AIDS (Goulder and Walker 2012), and other genetic diseases and cancer (Horton et al. 2004; de Bakker et al. 2006; Shiina et al. 2009).

One of the main transplant-related complications after HLA matching allo-HCT is called graft-versus-host disease (GVHD), which is caused by the graft immune cells (mainly T cells) that recognize peptides from the host as non-self antigens. This recognition initiates a chain of immune reactions and the cells in the recipient are attacked. These peptides are called the minor histocompatibility antigens (MiHAs) in distinction to the major histocompatibility complex (MHC) that leads to graft rejection. When the graft T

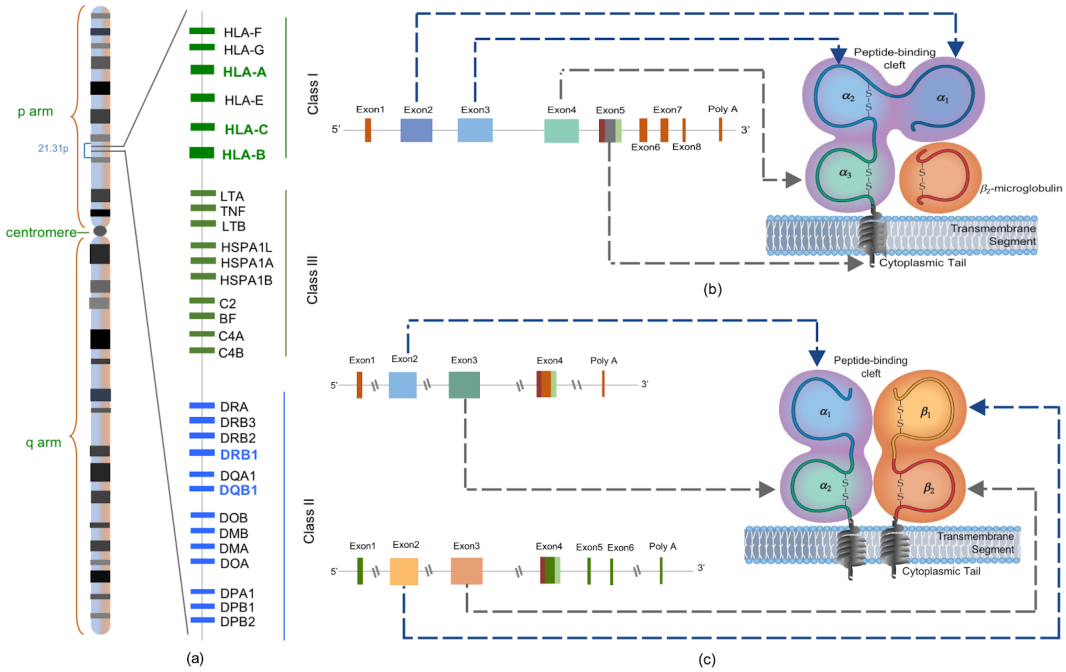


Figure 1.1 Illustration of genomic position of classical HLA genes and classical MHC molecules. The MHC molecule artwork is adapted from https://microbeonline.com/difference-mhc-class-mhc-class-ii-proteins/. (a) Genomic position of HLA genes on chromosome 6, (b) HLA class I genes functional composition and the corresponding MHC class I molecule, and (c) HLA class II genes functional composition and the corresponding MHC class II molecule. Blue dotted arrows indicate the corresponding exon regions and the encoded antigen recognition domains (ARDs). Exons 2 and 3 in HLA class I genes and Exon 2 in HLA class II genes encode the ARD, respectively.

cells recognize MiHAs in the tumor cells, it leads to graft-versus-leukemia (GVL) effect, which is beneficial to disease remission. Graft selection in allo-HCT is to minimize the GVHD while maximizing GVL effect.

The human leukocyte antigen (HLA) gene system

The HLA genes are located on chromosome 6 at p21.31 and consist of three different classes of genes, as shown in Figure 1.1(a). Class I and II genes are structurally and functionally related and play an essential role in allo-HCT. Class III genes are not directly

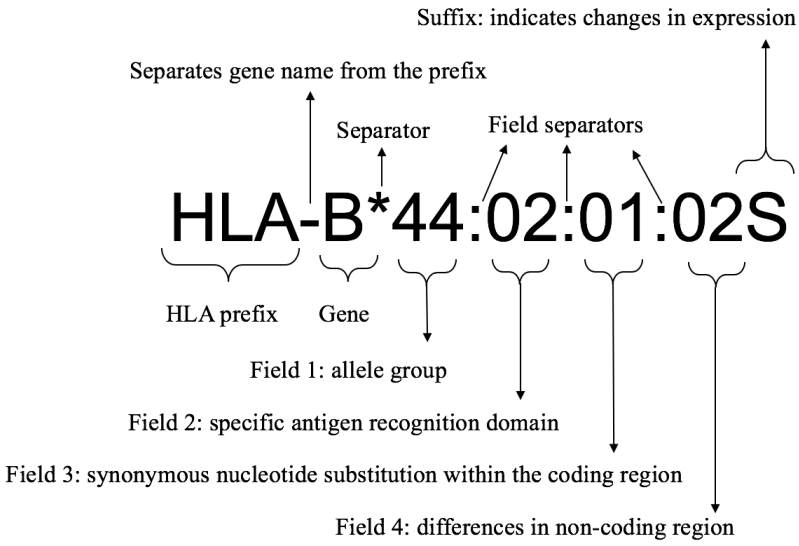


Figure 1.2 Illustration of the HLA gene nomenclature.

involved in donor-recipient histocompatibility testing, however, they function through inflammatory reactions (Petersdorf et al. 2018). The HLA genes are further categorized into classical and nonclassical genes based on the gene polymorphism and the direct interactions with T cells. Classical class I genes include HLA-A, -B and -C. These genes have high polymorphisms in exons 2, 3 and 4 that define different alleles and encode an **α**1, an **α**2, and an **α**3 domain of the MHC molecules. As shown in Figure 1.1(b), the **α**1 and the **α**2 domain form a peptide-binding groove which presents a peptide of 8 to 10 amino acids in length to CD8+ T cells (Halenius, Gerke, and Hengel 2015). This region is called the antigen recognition domain (ARD). Similarly, HLA-DRB1 and -DQB1 belongs to the classical class II genes, which encode MHC molecules that present peptides with more variable lengths, ranging from 10 to 20 amino acids to CD4+ T cells.

Graft histocompatibility testing largely depends on the determination of the classical HLA gene allele types of graft stem cells and the patients, which is referred to as HLA typing. For effective communications across HLA typing laboratories, clinicians and

researchers, the Harmonization of Histocompatibility Typing Terms Working Group defined a consensual Histocompatibility Typing Terms (Nunes et al. 2011). The HLA typing results are reported based on the World Health Organization (WHO) HLA Nomenclature Report (Marsh and WHO Nomenclature Committee for Factors of the HLA System 2018) and the IMGT/HLA database (Robinson et al. 2015). Figure 1.2 shows an example of HLA allele in a four-field name. Low-resolution typing refers to the alleles with only one field (e.g., HLA-B*44:XX) which corresponds to a serologic equivalent (B44), and a high-resolution HLA typing usually includes alleles with two-field names (e.g., HLA-B*44:02), which indicates alleles that encode the same protein sequence on the ARD of the MHC molecule. When an allele cannot be determined at a high resolution, the allele is reported at between high- and low-resolution and is assigned the name mostly based on associated population and the common and well-documented alleles (Mack et al. 2013).

In practice, the allele types at five classical HLA gene loci, i.e., HLA-A, -B, -C, -DRB1 and -DQB1, are determined and matched for each DR pair. All five gene loci matched (10 of 10 allele-matched) allo-HCT cases have shown significantly improved survival rate and transplant outcomes (Fürst et al. 2013). However, approximately 40% of patients after 10/10 HLA-matched allo-HCT still suffer from relapse and non-relapse causes of death, such as GVHD (D'Souza and Zhu 2017). In allo-HCT, GVHD is the major cause of non-relapse morbidity and mortality, affecting up to 40~60% (M. Jagasia et al. 2012), accounting for 20% of deaths after allogeneic HSCT (Pasquini and Zhu 2015). Maximizing the graft-versus-leukemia (GVL) effect and limiting GVHD has been the main research topic to improve the effectiveness of allo-HCT. Specifically, these

effects are caused by the polymorphic peptides presented on the surface of HLA Class I/II molecules, which activate the donor-derived cytotoxic T cells and subsequent immune responses. The influence of mismatches of classical HLA genes between donor and recipient on GVHD have mainly been investigated and well established at the ARD level (Fürst et al. 2013); however, there has not yet been an established study on the influence of mismatches of non-ARD exons and/or introns on the GVHD outcomes after allo-HCT.

Pathogenesis of acute graft-versus-host disease

GVHD is an immunologically mediated complex disease resulting from donor-derived T cell activation and attack on recipient normal cells due to the genetic disparities between donor and recipient. Clinically, GVHD has an acute form and a chronic form. Acute GVHD shows damaging symptoms on the skin, liver, and the gastrointestinal tract, while chronic GVHD has more diverse manifestations (for instance, nails, mouth, eyes) and sometimes resembles autoimmune syndromes. For epidemiological studies, these two forms were defined based on the symptoms that occur before or after Day 100 after transplantation. However, this definition does not correctly distinguish the two forms of GVHD, and hence a recent updated National Institutes of Health classification clarified the categorization and scoring form for acute and chronic GVHD by including late-onset acute GVHD (after Day 100) and the “common” features that indicate signs and symptoms found in both acute and chronic GVHD (M. H. Jagasia et al. 2015). Table 1.1 and Table 1.2 show the severity of each organ in acute GVHD, and the overall clinical grading of acute GVHD adapted from 1994 Consensus Conference on acute GVHD grading (Przepiorka et al. 1995; Ball, Egeler, and EBMT Paediatric Working Party 2008).

The development of acute GVHD conceptually includes three main sequential phases (Ferrara et al. 2009), or more specifically five basic steps (Socié and Blazar 2009). The first phase includes the activation of antigen presenting cells (APCs) due to host tissue damage caused by the underlying disease or the allo-HCT conditioning regimen. The second phase is activation and expansion of donor-derived T cells in response to the host and graft APCs as well as inflammatory cytokines (Bader et al. 2004). The last phase is an effector phase with a complex cascade of both cellular mediators (cytotoxic T lymphocytes, NK cells) and inflammatory mediators (TNF-α, INF-γ, IL-1). The dysregulation of cytokines resulted from these mediators eventually leads to the clinical manifestations of acute GVHD (Antin and Ferrara 1992; Couriel et al. 2004; Zeiser et al. 2004; Ball, Egeler, and EBMT Paediatric Working Party 2008).

The genetic disparities between donor and patient lead to the immunogenic polymorphic peptides, or the minor histocompatibility antigens (MiHAs). To the best of our knowledge, a very limited number of HLA restricted MiHAs have been identified and characterized in the literature, and the global roles of these MiHAs on the development of acute GVHD remain to be understood. Griffioen *et al.* reviewed forty-eight HLA Class I-restricted and eight HLA Class II-restricted autosomal MiHA genes that have been discovered and characterized through in vivo immune responses (Griffioen, van Bergen, and Falkenburg 2016). They were individually and specifically searched on candidate genes such as HLA-ligandomes, hematopoiesis-restricted genes and single nucleotide polymorphism (SNP) association with the clinical outcome after transplantation. However, the interactions among these SNPs is still unclear, as well as their collective influences on acute GVHD.

To date, genome-wide association studies (GWAS) and candidate gene studies have identified SNPs associated with acute GVHD, including SNPs that cause the genetic disparities between the donor and the patient, i.e. MiHA SNPs (Griffioen, van Bergen, and Falkenburg 2016), and SNPs that modify gene functions (Petersdorf et al. 2015). However, the genetic risks for acute GVHD outcome have not been well defined yet (Hansen et al. 2010). Most such studies have focused on single locus variations individually and tested them for association with acute GVHD. Unlike the underlying assumptions of these studies, however, genes tend to work interactively within specific functional pathways, contributing to the disease phenotypes.

It is reported that non-HLA genetic factors also play an essential role in HLA-identical allo-HCT (Yang and Sarwal 2017), including immune response and regulatory pathway gene polymorphisms (Lin et al. 2005b; Mullally and Ritz 2007), killer-cell immunoglobulin-like receptors (KIRs)(Littera et al. 2012), MHC Class I polypeptide-related sequence A (MICA) (Park et al. 2016; Chojecki 2017; Fuerst et al. 2016), and minor histocompatibility antigens (MiHAs) (Griffioen, van Bergen, and Falkenburg 2016; Martin et al. 2017). These factors may trigger allo-immune responses in the recipient who has received HLA-identical stem cell transplant, either through incompatible receptor-ligand interactions or encoding non-self peptides that trigger the donor's immune cells. Non-HLA gene studies are mainly targeted on the immune response related genes (candidate gene approach); however, the complex interactions among different genes in acute GVHD after HLA-identical allo-HCT remains to be understood. Gene-gene interaction, or epistasis, in the context of allogeneic transplantation, exhibits its unique features and challenges, compared to epistasis in traditional population

Table 1.1 Organ staging of acute GVHD, adapted from Przepiorka et al. (Przepiorka et al. 1995)

Stage	Skin	Liver	GI Tract
0	No rash due to GVHD	Bilirubin (< 2.0 mg/100 ml)	None (< 280 ml/m^2)
I	Maculopapular rash <25% of body surface area without associated symptoms	Bilirubin (2.0–3.0 mg/100 ml)	Diarrhea >500–1000 ml/day (280–555 ml/m^2); nausea and emesis
II	Maculopapular rash or erythema with puritis or other associated symptoms ≥25% of body	Bilirubin (3.0–5.9 mg/100 ml)	Diarrhea 1000–1500 ml/day (556–833 ml/m^2); nausea and emesis
III	Generalized erythroderma; symptomatic macular, papular or vesicular eruption with bullous formation or desquamation covering ≥50% of body surface area	Bilirubin (6.0–14.9 mg/100 ml)	Diarrhea >1500 ml/day (>833 ml/m^2); nausea and emesis
IV	Generalized exfoliative dermatitis or bullous eruption	Bilirubin (>15 mg/100 ml)	Diarrhea >1500 ml/day (>833 ml/m^2); nausea and emesis. Abdominal pain or ileus

GI: Gastrointestinal tract

Table 1.2 Clinical grading of the severity of acute GVHD, adapted from Przepiorka et al. (Przepiorka et al. 1995)

Grade	Skin	Liver	GI tract	Functional impairment
0	0	0	0	0
I	1–2	0	0	0
II	1–3	1	1	1
III	2–3	2–3	2–3	2
IV	2–4	2–4	2–4	2–4

GI: Gastrointestinal tract

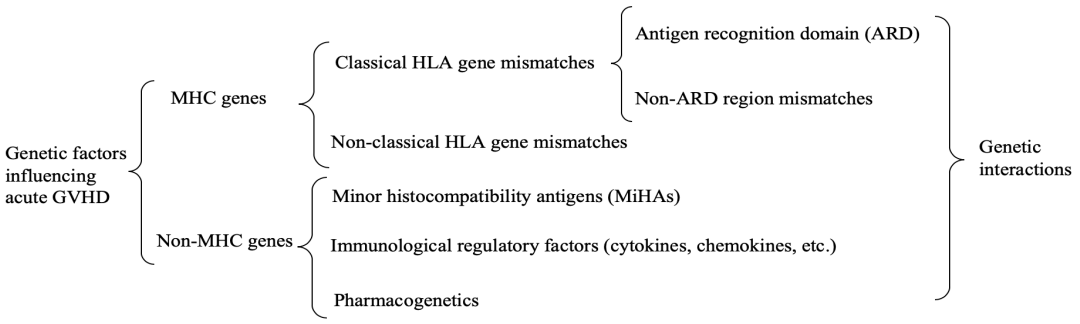


Figure 1.3 Genetic factors that influence the acute GVHD outcome in allo-HCT

genetic models where the interactions are modeled under the assumptions of different inheritable genetic models (e.g., additive, dominant, recessive models). In the case of transplantation, it is likely that the synergistic interactions of several genes in a biochemical pathway from both the donor and the recipient play more critical roles in the development of acute GVHD. For example, the synergism between IL10 gene of the recipient and IL10 receptor β gene of the donor in the IL10 metabolic pathway reportedly modulates the severity of acute GVHD (Lin et al. 2005a; Tseng et al. 2009). Figure 1.3 shows the genetic factors that are involved in the development of acute GVHD.

Main contributions of the dissertation

Chapter 2 explores the HLA classical gene sequence diversity at scale and evaluates the mismatches outside the ARD region between ARD matched donor and recipient and their potential impact on the transplant outcomes. This research is the first systematic evaluation of the role of mismatches outside the ARD of the classical HLA genes on the allo-HCT outcomes. Additionally, the bioinformatics pipeline incorporates multiple sequence analysis with functional annotation for donor-recipient pairs. Note that the HLA Class II genes are highly prone to misalignment due to their extremely long and

polymorphic intron regions if aligned directly to the whole gene reference sequences. Targeted region alignment for Class II genes in the proposed pipeline significantly improves the alignment accuracy. All these features will benefit the HLA research community and provide new insights on HLA mismatch types in allo-HCT outcomes. This work has been presented in the 42nd and 44th ASHI annual symposiums, respectively (H. Huang et al. 2016, 2018).

In Chapter 3, we expand our focus on the mismatches between donor and recipient to the whole genome level. We developed a bioinformatics analysis pipeline to compare the genomic sequences between donor and recipient, especially the non-synonymous mismatches restricted by HLA typing. We successfully identified Y-chromosome encoded minor histocompatibility antigens in sex-mismatched transplant cases that may be directly linked the acute GVHD symptoms. The bioinformatics tools provide a systematic analysis of whole genome sequences of the HLA-matched donor-recipient pairs and identify potential MiHAs efficiently *in silico*. This work has been published in *Blood Advances* (W. Wang et al. 2018).

The complex transplant-related outcome goes beyond the genetic mismatches between donor and recipient. In Chapter 4, we therefore propose a data mining technique to investigate the complex genetic relationships among donor and recipient genomes and the transplant outcomes. Specifically, we develop a feature selection method, called iRBA-RF, to identify the most informative SNPs that are linked to the transplant outcome (e.g., acute GVHD) and provide possible biological functional explanations. The proposed model does not require prior biological knowledge, such as the functional

pathways and immunological signatures or any assumptions. The framework examines gene sets and their association with the transplant outcomes, as opposed to investigating individual genes associated with the outcomes, which is the current state in the literature. The proposed algorithm provides novel bioinformatics tools for transplant genomic studies where exploring the ultra-high dimensional genotype data is one of the biggest challenges. It will bring new insights into understanding genetic factors that drive GVHD/GVL effects after HLA-matched allogeneic HSCT. This work has been submitted to a peer-reviewed journal and undergoing the manuscript review process.

Chapter 2 HLA gene sequence diversity and the impact of sequence mismatches in allogeneic hematopoietic stem cell transplantation outcomes

Abstract

In allogeneic hematopoietic stem cell transplantation (allo-HCT), HLA allele matching between donor and recipient (DR) has traditionally focused on the polymorphic antigen recognition domain (ARD). While mismatching at the ARD is known to influence transplant outcomes, it is unclear about the role of the mismatches outside the ARD. An estimate of up to 70% of the allo-HCT recipients still develops post-transplant complications, such as acute graft-versus-host disease (GVHD), even after the use of 10 out of 10 (HLA-A, -B, -C, DRB1, -DQB1) HLA allele-matched unrelated donors. In order to determine whether the sequence mismatches outside the ARD region are linked to the transplant-related complications, we assessed the genetic sequence variations of the classical HLA genes at a general population and characterized the frequency of mismatches in non-ARD regions when DR pairs are matched at the ARD level. The analysis of 15,865 healthy donors' classical HLA gene sequences revealed that in addition to the expected high sequence variations in the ARD region, the noncoding regions (5'-, 3'- untranslated regions and introns) and non-ARD exons also exhibit high sequence diversities. Despite the high sequence variation across different regions, a

subsequent analysis of 4,646 10/10 HLA matched DR pairs' HLA gene sequences showed limited mismatches outside the ARD regions. For HLA Class I alleles, 95.19% of the ARD matched alleles have identical sequences across the whole gene region. 0.67% of the mismatches were synonymous variants from the ARD region, while 0.17% and 0.10% of mismatches observed in the non-ARD exons were synonymous and nonsynonymous variants, respectively. The intronic variation accounted for 4.39% of the mismatches. Similarly, for HLA Class II alleles, 0.28% of mismatches were synonymous ARD variants, and the mismatches in the non-ARD exons were also very rare (synonymous: 0.28%; nonsynonymous: 0.16%). A high degree of variation was observed in the intronic regions of the HLA class II genes, with only 77.3% of the allele pairs shared having identical sequences. Overall, 0.22% and 4.56% of HLA class I and class II allele pairs, respectively, showed both exon and intron mismatches. In conclusion, due to the low frequency of amino acid mismatches in the non-ARD region and their reportedly weak allo-immune reactions, we suggest that the non-ARD sequence mismatches within 10/10 HLA ARD matched DR pairs have limited influence on the development of post-transplant complications, such as acute GVHD, and may not be a primary factor.

Keywords: human leukocyte antigen (HLA), antigen recognition domain, allogeneic hematopoietic stem cell transplantation (allo-HCT), graft-versus-host disease (GVHD)

Introduction

Allogeneic hematopoietic stem cell transplantation (allo-HCT) is the main curative treatment option for a series of blood and bone marrow related disorders, including

leukemia and lymphoma. One of the most crucial criteria in allo-HCT is to identify the histocompatibility (tissue compatibility) between the potential donor and the patient. The human leukocyte antigen (HLA) complex genes, also known as the major histocompatibility complex (MHC) genes, play a critical role in determining the histocompatibility of a potential donor with a recipient in a stem cell transplant. HLA complexes recognize and bind to antigens produced from non-self proteins and initiate corresponding immune responses, causing graft rejection, graft-versus-host disease (GVHD), and other post-transplant disorders.

For allo-HCT purposes, the ideal donors are HLA genotypically identical siblings of the patients. In theory, the likelihood of having such a donor for a patient depends on the number of siblings, for instance, 25% for those with one sibling, 44% for those with two siblings, and 58% for those with three. Besse et al. showed that the actual probability of having an HLA-identical sibling depends on the patient's age and their ancestry information and that in the U.S., only 13% to 51% of patients can find such a sibling donor (Besse et al. 2016). In other words, 49% to 87% of allo-HCT patients are expected to rely on an unrelated donor source. The current gold standard for donor selection is to test the HLA compatibility of 10 alleles at five classical HLA loci, i.e., HLA-A, -B, -C, -DRB1, and -DQB1, as mismatches at either of these locus are significantly associated with deteriorated survival rates and post-transplant complications (S. J. Lee et al. 2007; Fürst et al. 2013; Petersdorf 2015; Morishima et al. 2015; Kekre et al. 2016; Petersdorf 2017). Specifically, several multicenter retrospective allo-HCT outcome studies have shown that single-allele mismatched transplant cases have an estimated risk of acute GVHD ranges from 13% to 69% (Ciurea et al. 2011; Nakamae et al. 2010; Mehta et al.

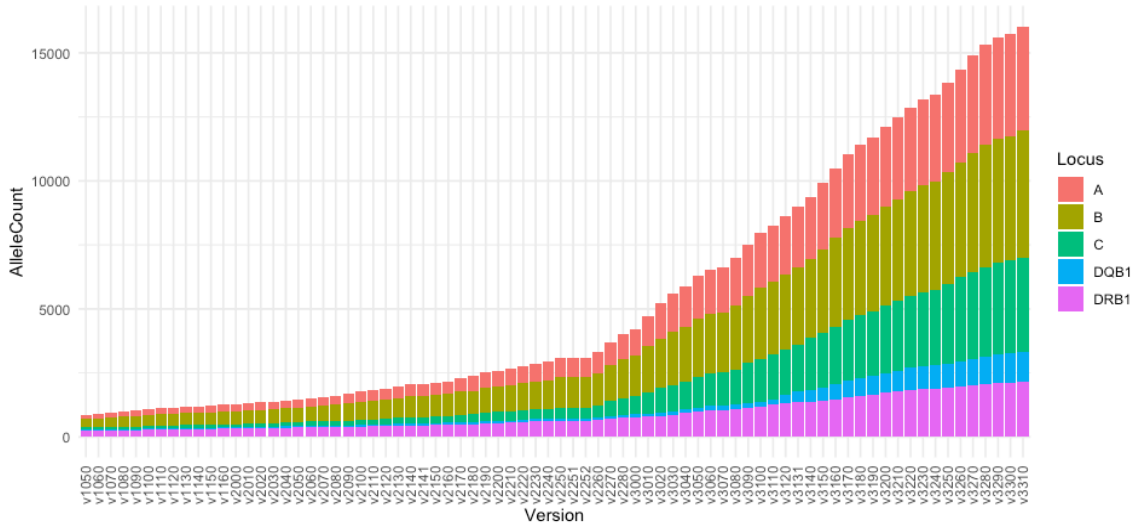


Figure 2.1 The dramatic growth of curated HLA classical gene alleles since 1998 (v1.0) from 964 alleles to 16,040 alleles (v3.31.0).

2004; Kekre et al. 2016). Thus, HLA-A, -B, -C, -DRB1, -DQB1 (10/10) allele-matched unrelated donors are considered as the optimal alternatives when HLA identical siblings are not available.

In practice, only the genetic regions in HLA genes that encode the antigen binding groove are characterized and compared between the potential donor and the recipient. The influence of genetic mismatches within the antigen-recognition domain (ARD) on the transplant outcome has been primarily investigated and well characterized based on the assumption that the ARD mismatches are the major player in the post-transplant complications. However, little is known about the functional implication of mismatches outside the ARD on allo-HCT outcomes.

The high-throughput next-generation sequencing (NGS) platform has enabled the accurate and high-resolution typing of HLA genes. As a result, novel alleles have been

Table 2.1 Summary of curated HLA classical gene alleles
in IMGT/HLA database (v3.31.0) (January 2018)

HLA Gene Class	Locus	#Alleles	#Full-length alleles	%Full-length alleles
Class I	HLA-A	4,080	849	20.8 %
	HLA-B	4,948	969	19.6 %
	HLA-C	3,684	873	23.7 %
*Class II	HLA-DRB1	2,146	87	4.1 %
	HLA-DQB1	1,176	173	14.7 %
Total		16,034	2,951	18.4 %

* HLA Class II genes have extremely long intronic sequence (over 2000 bp). It is difficult accurately sequence the full gene region. For HLA class II genes, the sequences that cover from intron 1 through intron 3 are counted towards the “full-length alleles” in this summary.

discovered at a tremendous rate in recent years, as shown in Figure 2.1. However, due to the particular interest on the ARD regions as well as challenges in characterizing full gene region sequences, as of January 2018, only less than 19% of the recognized classical five loci HLA alleles (HLA-A, -B, -C, -DRB1, -DQB1) have characterized full-length gene sequences or sequences outside the ARD region in the IMGT/HLA database Release 3.31.0, as shown in Table 2.1 (Robinson et al. 2015; Marsh and WHO Nomenclature Committee for Factors of the HLA System 2018; European Bioinformatics Institute 2018). For HLA Class II alleles (HLA-DRB1 and -DQB1), the extremely long and repetitive intron regions (over 2000 bp) pose challenges on sequence alignment, assembly, and haplotype assignment, so that less than 8% of recognized Class II genes have characterized sequences outside ARD. Sequence variations outside the ARD

region are not well characterized, and the impact of donor-recipient (DR) pair HLA mismatches in non-ARD regions are poorly understood.

Based on the HLA nomenclature system, alleles that share the same nucleotide sequence in the ARD regions are designated as the G allele groups, while alleles that encode the same polypeptide sequences in this region are designated as the P allele groups(Marsh and WHO Nomenclature Committee for Factors of the HLA System 2018). For instance, HLA-A*01:01:01G includes 72 alleles that have the same ARD exon sequences but different sequences in non-ARD exons and non-coding regions [5'-, 3'-untranslated region (UTR) and introns]. Similarly, HLA-A*01:01P includes 157 alleles that encode the same polypeptides in ARD region but different amino acid sequences in the rest of the HLA molecule. The impact of amino acid substitutes outside the ARD region in the G group alleles on T cell recognition has not been well characterized, except the case where it results in the loss of HLA expression that affects the allorecognition process.

Studies have shown that up to 70% of the patients who undergo 10/10 HLA matched unrelated allo-HCT still suffer from post-transplant complications, such as acute GVHD (Shaw et al. 2017; D'Souza and Zhu 2017). The purpose of this study is to determine whether the sequence mismatches outside the ARD region are linked to these transplant-related diseases after 10/10 HLA matched allo-HCT. We first assess the genetic sequence variations of the classical HLA genes at a general population, especially the non-ARD exons and non-coding regions, and then characterize the

frequency of mismatches in non-ARD regions when DR pairs are fully matched at the ARD level.

Materials and methods

Healthy donors' HLA gene sequences for gene sequence diversity analysis

15,865 healthy donors including four broad races (African American: 15.4%; Asian and Pacific Islander: 11.6%; Caucasian: 22.0%; Hispanic Origin: 28.9%; Native American: 22.1%) were selected from the Be The Match Registry® operated by the National Marrow Donor Program. The HLA Class I (HLA-A, -B, -C) alleles were sequenced through the Illumina MiSeq platform (Illumina, San Diego, CA). Several thousand pair-ended short reads (150 bp) were then assembled into consensus sequences that include from 5'-untranslated region (UTR) to 3'-UTR. HLA Class II (HLA-DRB1, -DQB1) alleles have extremely long intronic regions with repetitive nucleotide sequence blocks. For instance, introns 2 and 3 of the HLA-DRB1 gene are 2,229 bp and 701 bp long, respectively, whereas exons 2 and 3 are only 270 bp and 282 bp long, respectively. These intron regions lead to a challenge in characterizing the full-length allele sequences. Typically, a targeted sequencing scheme is employed to extract the sequences of exons 2 and 3 for HLA Class II alleles. Here we employed the Pacific Biosciences RS-II platform to obtain high-quality elongated sequences with partial intronic regions.

Multiple sequence alignment was performed on the five loci allele sequences using

Clustal Omega (Sievers et al. 2011), and low-quality sequences were removed from the downstream analysis. This was followed by a gene annotation pipeline to annotate exon and intron regions. We employed the Shannon entropy of each aligned nucleotide site at each of the five loci to characterize the allele diversity. The Shannon entropy is defined as

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i$$

where p_i is the frequency of each nucleotide type in the cohort. If all four types of nucleotides (A, T, C, and G) have an equal frequency (25%) at a position, then this position has the maximum entropy of 2 (or 2.32 in the case of considering alignment gaps as there are five possible values at each position, i.e., A, T, C, G and a gap), and hence it is considered the most variable position. At a most conserved position, there is only one type of nucleotides, and the entropy has the minimum value of 0.

Donor-recipient pair HLA gene sequences comparison

A cohort of 10/10 (HLA-A, -B, -C, -DRB1, -DQB1) high-resolution matched retrospective transplant cases (*n*=4,646) performed between 2000 and 2017 was selected from the Center for International Blood and Marrow Transplant Research (CIBMTR) Repository, and the donor-recipient pairs' HLA gene alleles were sequenced. HLA Class I alleles were sequenced through the Pacific Biosciences RS-II platform to characterize the full gene regions, whereas HLA Class II alleles were sequenced through the Illumina MiSeq platform for targeted exon regions. Figure 2.2 shows the developed comprehensive HLA allele sequence comparison pipeline that identifies and annotates the mismatched positions between two alleles by their functional regions and their protein sequence

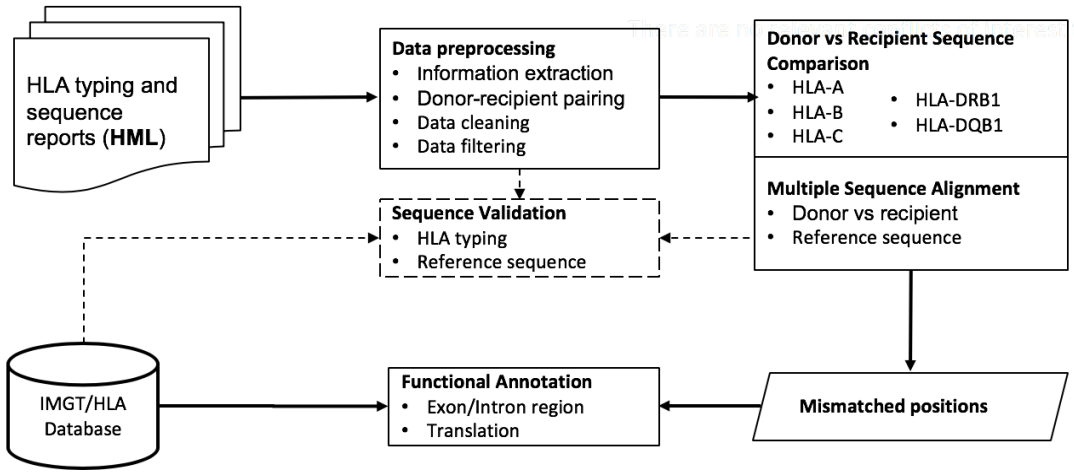


Figure 2.2 Diagram of the donor-recipient pair HLA gene sequence comparison pipeline

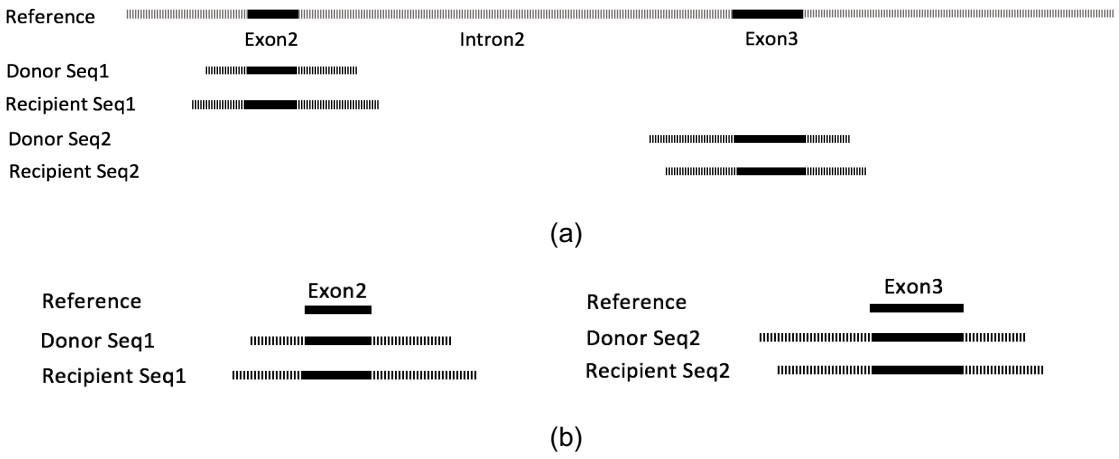


Figure 2.3 Targeted alignment for HLA class II gene sequences

differences using IMGT/HLA Database (v3.31.0). The sequence alignment for comparison was performed through the MUSCLE algorithm (Edgar 2004).

The HLA Class II gene sequences often reported with partial intronic regions, as shown in Figure 2.3(a), which can cause misalignment when the query sequences are aligned against the full-length reference sequence. For instance, the exon region of the query

sequence may be stretched to the intron region after alignment with gaps in between. In order to avoid this issue, we adopted a target region alignment, i.e., instead of aligning against the full-length reference sequence, the query sequences were aligned against only exons 2 and 3 from the reference, respectively, as shown in Figure 2.3(b). This strategy significantly improved the alignment accuracy and efficiency for HLA Class II genes.

Results

Information theory-based measurement shows the genetic sequence variations in non-ARD exons and non-coding regions are as diverse as in the ARD regions.

Figure 2.4 shows the nucleotide variation across the alignment positions. After alignment, the untranslated regions (URTs) generally show a high sequence variation for HLA Class I alleles. This may be due to the high polymorphism in the UTR regions; on the other hand, it also may be caused by the varying lengths of UTR regions for different alleles which introduced gaps in the alignment. In order to investigate the genetic variation without the alignment gaps, the positions with at least one gap were artificially given the value of -0.5 to filter out from the scatter plot, as shown in the right column of Figure 2.4.

Figure 2.5 illustrates the entropy distribution in each functional region. As expected, the ARD regions show high sequence variation in general (entropy values between 0.01 and

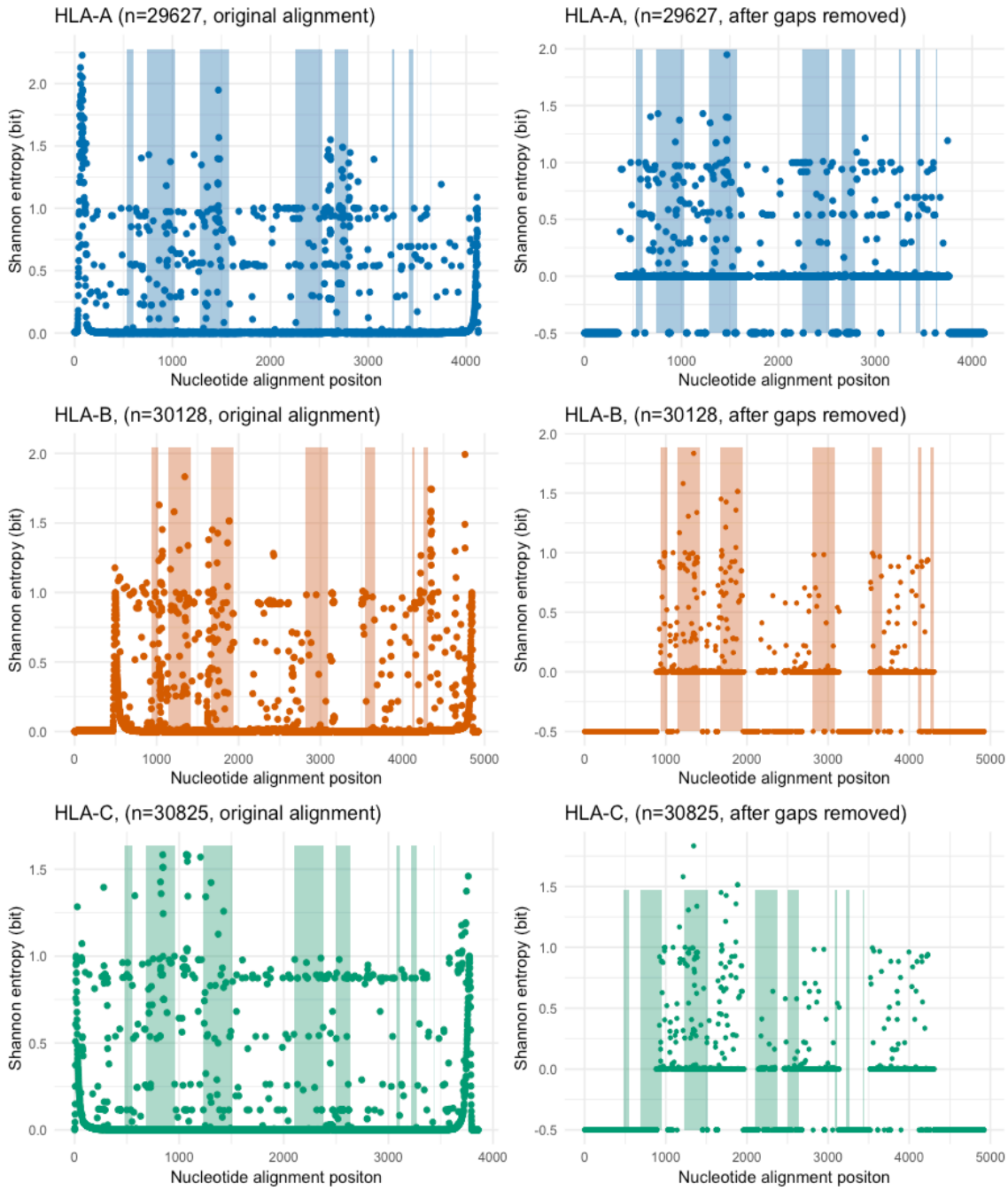


Figure 2.4 HLA class I gene alignment and their sequence diversity. Left column shows the entropy changes on the original alignment, and the right column shows the entropy after removing gaps from alignment. The positions which have at least one gap is given the value of -0.5. The shaded areas indicate exon regions.

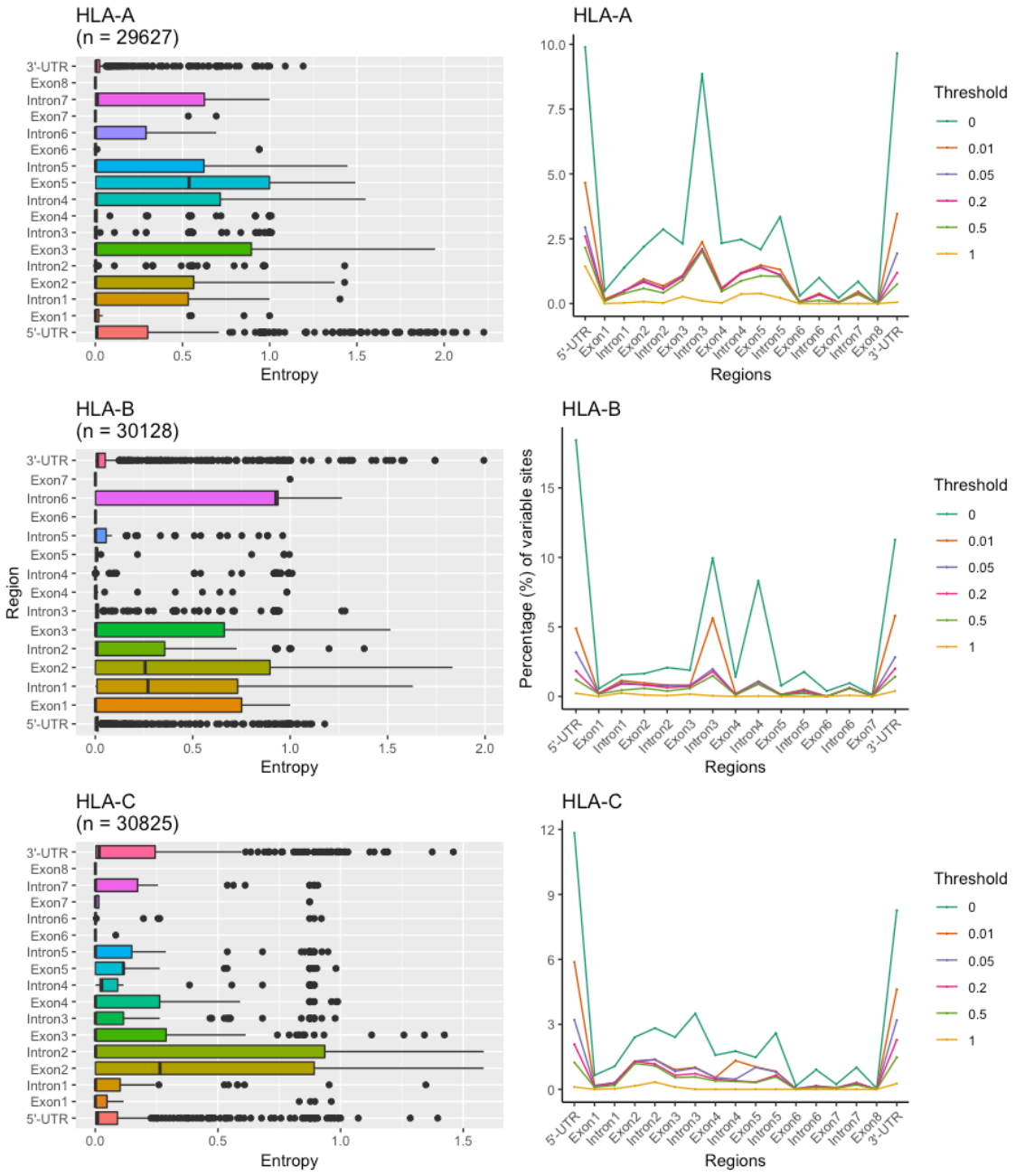


Figure 2.5 HLA class I gene sequence variable sites by functional regions after sequence alignment. Left column: distribution of entropy values in each region; Right column: the proportion of diverse position within each region above different entropy thresholds.

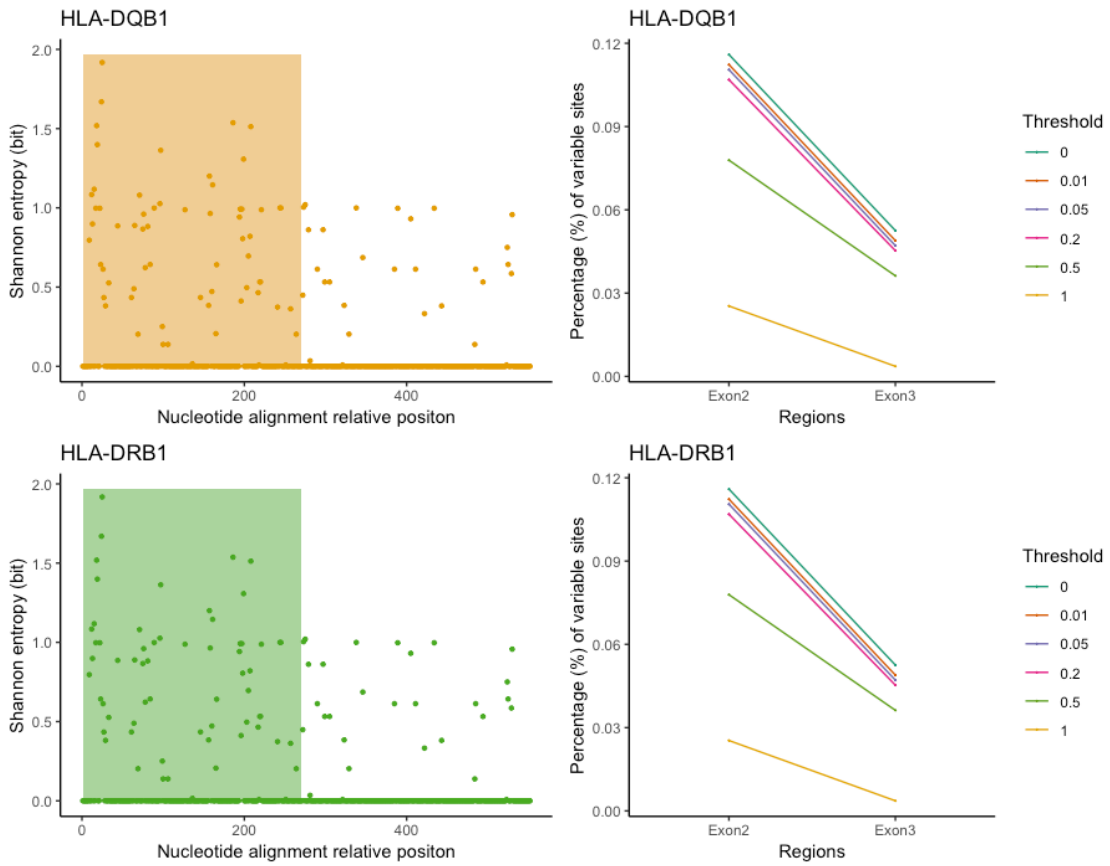


Figure 2.6 HLA class II gene alignment and their sequence.
Left column: shaded area shows exon 2 while the rest shows exon 3.
Right column: proportion (%) of variation sites within exons 2 and 3.

0.5). Highly variable sites (entropy>0.01) appear mostly in the noncoding regions (intron 3 in HLA-A; introns 3 and 4 in HLA-B; introns 2, 3 and 5 in HLA-C). Among non-ARD exon regions, exons 5 and 6 show high variation. For HLA Class II alleles, due to the incomplete intron sequences and alignment challenges, we only assessed the exons 2 and 3. Figure 2.6 shows similar results as HLA Class I genes, where non-ARD exon (exon 3) also shows a high variation.

10/10 high-resolution ARD-matched donor-recipient pairs have limited mismatches outside the ARD region

When a donor-recipient pair is matched at ARD regions, we observed very limited mismatches outside the ARD region. For HLA Class I alleles, 95.19% of the ARD matched alleles have identical nucleotide sequences between donor and recipient pairs. There are 0.3% of nucleotide sequence mismatches within the ARD region; however, all of them are synonymous variants, which are translated into the same protein sequences. In the non-ARD exon region, we observed 0.3% and 0.13% of synonymous and nonsynonymous nucleotide mismatches, respectively. The intronic variation accounted for 4.17% of the mismatches.

Similarly, for HLA Class II alleles, 0.28% of mismatches were synonymous ARD variants, and the mismatches in the non-ARD exons were also very rare, including 0.28% synonymous and 0.16% nonsynonymous variants. However, due to the high polymorphism in the intronic regions of the Class II genes, 26.48% of mismatches were intronic, and only 77.3% of allele pairs shared identical nucleotide sequences. For some allele pairs, nucleotide mismatches were observed in more than one region simultaneously. Specifically, 0.22 % and 4.56% of Class I and Class II allele pairs, respectively, showed mismatches in more than one functional region (exon and intron). Table 2.2 shows the detailed breakdown of mismatched allele pairs by HLA gene locus. A total of 25 different allele pairs showed nonsynonymous variants in non-ARD exons leading to different amino acid sequences of the MHC molecules, as shown in Table 2.3. Seven of these allele pairs showed different protein structures in the transmembrane

Table 2.2 Percentage (%) of mismatches between alleles from 10/10 ARD matched donor-recipient pairs HLA full gene sequences by locus.

Syn: synonymous; non-syn: nonsynonymous;

non-coding: intron and/or the untranslated regions (UTRs).

Locus	Identical (%)	Syn ARD (%)	Non-syn ARD (%)	Syn non-ARD exon (%)	Non-syn non-ARD exon (%)	Non-coding (%)	Total # allele pairs
A	95.19	0.67	0	0.17	0.10	4.39	9286
B	96.64	0.17	0	0.17	0.13	2.96	9290
C	94.50	0.05	0	0.18	0.16	5.17	9266
DRB1*	77.58	0.03	0	0.17	0.23	22.20	9090
DQB1*	77.08	0.53	0	0.45	0.10	30.67	9292
Class I	95.44	0.30	0	0.30	0.13	4.17	27842
Class II*	77.33	0.28	0	0.28	0.16	26.48	18382

* Class II genes were sequenced at the following targeted regions: full-length exon 2 (270 bp) with partial intron 1 and partial intron 2, full-length exon 3 (282 bp) with partial intron 2 and partial intron 3.

Table 2.3 Observed non-ARD nonsynonymous mismatched alleles between donor and recipient

Allele 1	Allele 2	Exon/Intron location; amino acid change from Allele 1 to Allele 2; protein location	Count of observed DR pairs	Allele 2: Common and well documented
HLA-A*23:01:01	HLA-A*23:17	Exon 5; His283Pro; TM	3	WD
HLA-A*02:01:01:01/02L	HLA-A*02:66	Exon 4; Thr225Ile; **α**3	1	Not CWD
HLA-A*02:01:01:01/02L	HLA-A*02:559	Exon 4; His188Arg; **α**3	1	Not CWD
HLA-A*02:01:01:01/02L	HLA-A*02:09	Exon 4; Ala236Glu; **α**3	1	Common
HLA-A*11:01:01:01	HLA-A*11:86	Exon 4; Gly221Arg; **α**3	1	Not CWD
HLA-A*24:02:01:06	HLA-A*24:79	Exon 4; Gly265Asp; **α**3	1	Not CWD
HLA-A*01:01:01	HLA-A*01:37	Exon 4; Thr228Met ; **α**3	1	Not CWD
HLA-B*44:02:01:03	HLA-B*44:27:01	Exons 4, 5, 7; Val199Ala, Val282Ile, A305T,	6	Common

		Cys325Ser; **α**3, TM, Cyt		
HLA-B*35:01:01:01	HLA-B*35:57	Exon 4; Val194Ile; **α**3	3	Not CWD
HLA-B*51:01:01:01	HLA-B*51:193	Exon 5; Ile282Val; TM	2	Not CWD
HLA-B*07:05:01	HLA-B*07:06	Exon 5; Ile282Val; TM	1	Common
HLA-C*07:01:01:01	HLA-C*07:18	Exon 6, Intron 4; Ala324Val; Cyt	5	Common
HLA-C*04:01:01:01	HLA-C*04:82	Exon 5; Leu300_Gly301insAVL; TM	1	WD
HLA-C*12:03:01:01	HLA-C*12:143	Exon 6; Ala325Val; Cyt	1	Not CWD
HLA-C*05:01:01:02	HLA-C*05:03	Exon 4; Glu183Asp, His184Pro, Val194Ile, Ala199Val; **α**3	1	Not CWD
HLA-C*15:02:01:01	HLA-C*15:13	Exon 4; Glu229Gln; **α**3	1	WD
HLA-DRB1*14:01:01	HLA-DRB1*14:54:01	Exon 3; Tyr112His; **β**2	20	Common
HLA-DRB1*13:01:01	HLA-DRB1*13:117	Exon 3; Arg133Trp; **β**2	1	Not CWD
HLA-DQB1*03:01:01/04	HLA-DQB1*03:19:01	Exon 3; Thr185Ile; **β**2	2	Common
HLA-DQB1*06:02:01	HLA-DQB1*06:111	Exon 3; Gly141Ser; **β**2	2	Not CWD
HLA-DQB1*03:01:01:02/03	HLA-DQB1*03:19	Exon 3; Thr185Ile; **β**2	1	Common
HLA-DQB1*03:01:01:01/02/03	HLA-DQB1*03:09	Exon 3; Gly168Ala, Asp169del; **β**2	1	Common
HLA-DQB1*03:03:02:01	HLA-DQB1*03:31	Exon 3; Val116Ile; **β**2	1	Not CWD
HLA-DQB1*02:01:01	HLA-DQB1*02:02:01	Exon 3; Asp135Gly; **β**2	1	Common
HLA-DQB1*02:02:01	HLA-DQB1*02:10	Exon 3; Val142Ile; **β**2	1	Not CWD

CWD: common and well-documented allele; WD: well-documented allele; TM: transmembrane; Cyt: cytoplasmic tail; CWD alleles are based on CWD release 2.0.

segment or the cytoplasmic tail, while the rest differed in **α**3 (Class I) or **β**2 (Class II) chain below the antigen binding cleft. Especially, HLA-B*44:02:01:03 and HLA-B*44:27:01 differ in all three regions outside the ARD, i.e., **α**3-chain, the transmembrane segment, and the cytoplasmic tail. Twelve of the alternative alleles (shown as Allele 2 in Table 2.3) were categorized as common or well-documented alleles according to the CWD catalogue 2.0.0 (Mack et al. 2013).

Discussion and conclusion

MHC genes are the most polymorphic gene regions among known the human gene system (Janeway et al. 2001). In this study, we showed the sequence diversity of the classical HLA genes in a general population, especially the high sequence variations in the non-ARD exons as well as the noncoding regions (5'-, 3'-UTR, or introns). Although the ARD region has been the focus in the practice of HLA matching, the high sequence variation outside the ARD region may also exert influence on the transplant outcomes.

The investigation of the 10/10 HLA matched transplant cases showed that the ARD-matched DR pairs have limited mismatches in the non-ARD exons (0.43%) and we observed the nonsynonymous mismatches at an even lower rate (0.13%). Alleles that share the same ARD sequences likely have the same non-ARD exon sequences. The low mismatch rate in non-ARD exons may be due to the high degree of linkage disequilibrium between the exons of the HLA alleles (Smith et al. 2005). The intron sequence mismatches were observed at a higher rate (4.17% for Class I and 26.48% for Class II), which may be explained by the different evolutionary forces in coding and non-

coding regions of HLA alleles (Cereb, Hughes, and Yang 1997; Meyer and Blasczyk 2000; Bergström et al. 2000). Intron variation is caused by the combined effects of mutation, recombination, and selection and it is reported that the recombination events concentrated in the introns near ARD exons (Cereb, Hughes, and Yang 1997; Kotsch and Blasczyk 2000; Meyer and Blasczyk 2000). Our observation of high intronic sequence variations is consistent with these reports. For HLA Class II alleles, we observed a much higher rate (26.48%) of intronic mismatches than a prior report with a smaller sample size (less than 2% of intronic mismatches, *n*=360) (Hou et al. 2017). This discrepancy may be due to the differences in typing methodologies and sequencing of highly repetitive intronic regions. As noted by Hou et al., in order to assess the impact of non-ARD mismatches on clinical outcomes of allo-HCT, an outcome association study may need at least 5,916 transplant pairs according to a log-rank test for 80% power at a significant level of $\alpha = 0.05$ (Hou et al. 2017).

However, the analysis of 4,646 transplant cases in this study suggests that sequence mismatching rate outside the ARD region between ARD-matched donor and recipient is relatively low and the mismatches of the corresponding amino acid sequences may be uncommon. Moreover, a recent study based on *in vitro* T cell assays by Roelen *et al.* suggested that non-ARD polypeptide mismatches may result in weak immunogenicity as they observed minimal T cell reactivity (Roelen et al. 2018). Even when there are amino acid mismatches outside the ARD region, they are unlikely the leading player to trigger the allo-immune reaction. Thus, we suggest that the non-ARD sequence mismatches between donor and recipient have limited influence on the development of post-transplant complications, such as acute GVHD, and may not be a primary factor.

Several retrospective analysis of allo-HCT cases confirmed that donor-recipient HLA matching at the ARD level is the most critical factor that influences overall survival and the development of acute GVHD (S. J. Lee et al. 2007; Fürst et al. 2013; Kollman et al. 2016). Furthermore, when more than one HLA matching donor is available, non-HLA characteristics, such as donor age and sex, have shown to be linked to the post-transplant complications (Kollman et al. 2016; Petersdorf 2017). In order to uncover the genetic mechanisms of the development of acute GVHD and facilitate donor selection for allo-HCT with better transplant outcomes, it will be beneficial for future research to investigate the non-MHC genes, genetic and immunologic regulatory pathways and donor-recipient synergistic interactions. In the next two chapters, we discuss the non-MHC gene factors and a method to detect potential genetic interactions from the DR pair whole genome sequences.

Chapter 3 Chromosome Y-encoded antigens associate with acute graft-versus-host disease in sex-mismatched stem cell transplant [1]

Abstract

Allogeneic hematopoietic stem cell transplantation (allo-HCT) is a curative option for blood cancers, but the coupled effects of graft-versus-tumor and graft-versus-host disease (GVHD) limit its broader application. Outcomes improve with matching at HLAs, but other factors are required to explain residual risk of GVHD. In an effort to identify genetic associations outside the major histocompatibility complex, we conducted a genome-wide clinical outcomes study on 205 acute myeloid leukemia patients and their fully HLA-A, -B, -C, -DRB1, and -DQB1–matched (10/10) unrelated donors. HLA-DPB1 T-cell epitope permissibility mismatches were observed in less than half (45%) of acute GVHD cases, motivating a broader search for genetic factors affecting clinical outcomes. A novel bioinformatics workflow adapted from neoantigen discovery found no

[1] This research was originally published in *Blood Advances*. "Wei Wang*, Hu Huang*, Michael Halagan, Cynthia Vierra-Green, Michael Heuer, Jason E. Brelsford, Michael Haagenson, Richard H. Scheuermann, Amalio Telenti, William Biggs, Nathaniel M. Pearson, Julia Udell, Stephen Spellman, Martin Maiers, and Caleb Kennedy. **Chromosome Y–encoded antigens associate with acute graft-versus-host disease in sex-mismatched stem cell transplant**. Blood Advances 2, no. 19 (2018): 2419-2429." © the American Society of Hematology.

W.W. and H.H. made equally substantial contributions to this research. The main contributions of H.H. include the development of bioinformatics workflows to analyze the whole-genome sequence data, identity-by-descent (IBD) calculation, result visualization and interpretation, and manuscript review.

associations between acute GVHD and known, HLA-restricted minor histocompatibility antigens (MiHAs). These results were confirmed with microarray data from an additional 988 samples. On the other hand, Y-chromosome–encoded single-nucleotide polymorphisms in 4 genes (*PCDH11Y*, *USP9Y*, *UTY*, and *NLGN4Y*) did associate with acute GVHD in male patients with female donors. Males in this category with acute GVHD had more Y-encoded variant peptides per patient with higher predicted HLA-binding affinity than males without GVHD who matched X-paralogous alleles in their female donors. Methods and results described here have an immediate impact for allo-HCT, warranting further development and larger genomic studies where MiHAs are clinically relevant, including cancer immunotherapy, solid organ transplant, and pregnancy.

Introduction

Allogeneic hematopoietic cell transplantation (allo-HCT) can cure certain inherited diseases and acquired malignancies of the blood, yet biological mechanisms that provide beneficial effects, such as graft-versus-tumor (GVT),(Copelan 2006; Horowitz et al. 1990; Miller et al. 2010) also contribute to life-threatening graft-versus-host disease (GVHD)(Ferrara et al. 2009). Outcomes improve dramatically with donor-recipient matching of HLAs, but GVHD still occurs at frequencies of up to 70% in fully matched unrelated pairs, and to a lesser degree in related, HLA-identical transplant recipients (Gooley et al. 2010), suggesting unaccounted-for genetic factors impact clinical responses.

Minor histocompatibility antigens (MiHAs) are germline-encoded immunogenic peptides

presented by specific HLA molecules on the surface of cancer cells or normal tissues. Although donor and recipient mismatching at the major histocompatibility complex (MHC) confers the highest proportional risk of GVHD, many clinically relevant MiHAs with defined HLA restriction have been identified (Oostvogels, Lokhorst, and Mutis 2016), including Y-chromosome–encoded antigens that affect outcomes in sex-mismatched HCT (Popli et al. 2014). In other nonmalignant conditions, such as solid organ transplant or pregnancy, MiHAs carry risk of rejection (Kim and Gill 2009; Wagner 2012; Pfeffer and Thorsby, n.d.) or miscarriage (Christiansen, Steffensen, and Nielsen 2011; H. S. Nielsen et al. 2010), respectively.

In leukemia, there is evidence for tumor-specific antigenicity by exogenous activation of gene expression (J. Molldrem et al. 1996; J. J. Molldrem et al. 1999; Gao et al. 2000), gene fusion (Cai et al. 2012), and alternative splicing (Pont et al. 2016). In all cancers, driver and passenger mutations mark tumor progression (Lindsley et al. 2017; Lazarian, Guièze, and Wu 2017; Alexandrov et al. 2013), which may guide biomarker discovery (Falchook et al. 2016; A. C. Huang et al. 2017) and individualized treatment (Budczies et al. 2017; Strønen et al. 2016; Rajasagi et al. 2014). A subset of cancer variants give rise to immunoreactive neoantigens encoded by somatic changes in tumor DNA, and these changes are presented exclusively by tumors and targeted by patients' normal immune systems (Tran et al. 2015). In a clinical setting, this effect may theoretically be exploited for GVT in allo-HCT (Burkhardt and Wu 2013) or precision medicine approaches to cancer immunotherapy.

Despite the physiological connection between MiHAs and neo-antigens, there are important differences that should guide genomic analysis. Neoantigen discovery from

DNA or RNA sequences requires high sensitivity to detect rare or private variants from heterogenous tumor tissue (Y.-C. Lu and Robbins 2016), which is often chemically preserved (Srinivasan, Sedmak, and Jewell 2002). Sequencing patient and adjacent normal samples adds cost but also reduces false positives (Jones et al. 2015; Garofalo et al. 2016). MiHAs, on the other hand, arise from heritable germline polymorphisms that may be common in populations and accessible with less expensive microarrays or lower-coverage sequencing panels (Sampson et al. 2014). In both cases, antigens are only immunoreactive if they are displayed by patient HLA in affected tissues. Therefore, it is important to annotate variants with predicted MHC restriction, binding affinity, and tissue-specific expression.

We sought a controlled, clinical-outcomes–based study in HLA-matched donor–recipient pairs to discover genetic variation outside the MHC that may contribute to the risk of acute GVHD following allo-HCT.

Methods

Study design

The study population consisted of high-resolution HLA-A–, HLA-B–, HLA-C–, HLA-DRB1–, and HLA-DQB1–matched (10/10) unrelated donor and recipient allo-HCT pairs. Patients were selected to obtain equal numbers with and without clinical evidence of grade II-IV acute GVHD, which was assessed as described based on severity or degree of organ involvement before day 100 after transplant (Deeg and Antin 2006). All patients received myeloablative conditioning for acute myeloid leukemia (AML) or other blood

cancers in complete remission (CR1 or CR2). After quality control and other filtering (see Methods), acute GVHD positive and negative cohorts were balanced for age, disease status, self-reported race or ethnicity, GVHD prophylaxis, and other factors (Table 3.1).

Clinical data collection

Clinical data were collected by the Center for International Blood and Bone Marrow Transplant Research (CIBMTR), a collaboration between the National Marrow Donor Program and the Medical College of Wisconsin representing a worldwide network of transplant centers that contribute detailed data on HCT. The CIBMTR conducts research in compliance with all applicable federal regulations pertaining to the protection of human research participants. All participants provided informed consent for participation in the CIBMTR research program, including submission of biological samples to the Research Repository, and this study was approved by the National Marrow Donor Program Institutional Review Board.

HLA typing and histocompatibility matching

HLA matching was determined at high resolution for HLA-A, HLA-B, HLA-C, HLA-DRB1, and HLA-DQB1 through retrospective typing of stored pre-transplant samples and/or reported by the transplant center and match assessment performed per CIBMTR criteria as previously described (Spellman et al. 2008). 5-locus haplotype matching was performed with the HapLogic algorithm (Dehn et al. 2016).

Whole-genome sequencing

250 donor and 250 HCT recipient samples (500 samples total) were sequenced at

Human Longevity, Inc. (San Diego, CA) to a mean coverage depth of 303 with 2 3 150 bp paired reads using Illumina HiSeq X instruments. 125 pairs came from transplants with clinical evidence of acute GVHD; 125 pairs came from transplants without evidence of GVHD. Ten recipient samples did not produce adequate sequencing data. A further 2 recipient samples and 1 donor sample failed the heterozygosity test that was applied to remove contaminated samples. An additional 32 samples were missing data for their paired donor or recipient and were removed from analysis. The final set included 205 pairs of donor-recipient samples (102 acute GVHD and 103 non-GVHD). Secondary analysis with Isaac alignment and variant calling pipeline (Raczy et al. 2013) resulted in 1 binary alignment map (Li et al. 2009) and 1 variant call format (Danecek et al. 2011) file per sample using the human genome reference assembly hg38. Variants with below average read depth (30X) were excluded from analysis.

Microarray data and analysis

The microarray data and primary analysis for Supplementary Table 3.1 have been described previously (Madbouly et al. 2017).

Bioinformatics

Genomic similarity was measured using identity-by-descent (IBD) sequencing with default parameters (Browning and Browning 2013). This technique determines phase for donor and patient genotypes to form haplotype segments of varying lengths, which indicate common ancestry. Normalizing the lengths of these segments to those of specific genomic features (including the whole genome itself) gives a relative measure of genetic similarity for each feature (Figure 3.1). For comparison, the null distribution of

Table 3.1 Counts of patients and donors in the acute graft-versus-host (GVHD) and non-GVHD groups. P-values were calculated using the Pearson chi-square test for comparing discrete variables or the Kruskal-Wallis test for comparing continuous variables. Abbreviations: CR – complete remission, PBSC – peripheral blood stem-cells, TCE – T cell epitope, CMV – cytomegalovirus.

	AGVHD	No AGVHD	
Variable	N (%)	N (%)	p-value
Number of Recipients	102	103	
Number of centers	47	50	
Patient-related			
Recipient age at transplant			0.09
0-9 years	3 (3)	1 (1)	
10-19 years	10 (10)	6 (6)	
20-29 years	16 (16)	17 (17)	
30-39 years	21 (21)	15 (15)	
40-49 years	23 (23)	42 (41)	
50-59 years	29 (28)	22 (21)	
Median (Range)	41 (1-66)	44 (9-66)	0.58
Recipient race/ethnicity			0.79
Caucasian, non-Hispanic	87 (92)	94 (92)	
African-American, non-Hispanic	1 (1)	2 (2)	
Asian, non-Hispanic	2 (2)	3 (3)	
Hispanic, Caucasian	5 (5)	3 (3)	
Unknown	7 (N/A)	1 (N/A)	
Recipient sex			0.009
Male	64 (63)	46 (45)	
Female	38 (37)	57 (55)	
Karnofsky score			0.47
10-80	24 (24)	32 (31)	
90-100	72 (71)	66 (64)	
Missing	6 (6)	5 (5)	
Disease-related			
Disease status at transplant			0.61
Early (CR1)	71 (70)	75 (73)	

Intermediate (CR2+)	31 (30)	28 (27)	
Transplant-related			
Stem cell source			0.24
Marrow	27 (26)	35 (34)	
PBSC	75 (74)	68 (66)	
TCE nonpermissiveness			0.22
Ambiguous DPB1 allele	0	1 (1)	
Permissive DPB1	54 (54)	62 (63)	
Non-permissive DPB1	46 (46)	35 (36)	
Data Missing	2 (N/A)	5 (N/A)	
GVHD Prophylaxis			0.80
Tacrolimus + MMF +- others	22 (22)	20 (19)	
Tacrolimus + MTX +- others (except MMF)	56 (55)	57 (55)	
Tacrolimus + others (except MTX, MMF)	5 (5)	2 (2)	
Tacrolimus alone	3 (3)	1 (1)	
CSA + MMF +- others (except Tacrolimus)	2 (2)	3 (3)	
CSA + MTX +- others (except Tacrolimus, MMF)	12 (12)	18 (17)	
CSA + others (except Tacrolimus, MTX, MMF)	1 (1)	1 (1)	
CSA alone	1 (1)	1 (1)	
Donor/Recipient sex matching			0.007
Male/Male	44 (42)	40 (39)	
Male/Female	24 (25)	38 (37)	
Female/Male	21 (21)	6 (6)	
Female/Female	13 (13)	19 (18)	
Donor/Recipient CMV serostatus			0.68
Negative/Negative	33 (32)	29 (28)	
Negative/Positive	37 (36)	36 (35)	
Positive/Negative	10 (10)	14 (14)	
Positive/Positive	19 (19)	23 (22)	
Unknown	3 (3)	1 (1)	
Donor age at donation			0.18
18-19 years	4 (4)	1 (1)	
20-29 years	45 (44)	54 (52)	
30-39 years	29 (28)	31 (30)	

40-49 years	16 (16)	15 (15)	
50 and older	8 (8)	2 (2)	
Median (Range)	30 (19-56)	28 (20-52)	0.11
Donor race/ethnicity			0.46
Caucasian, non-Hispanic	92 (94)	92 (91)	
African-American, non-Hispanic	0	2 (2)	
Asian, non-Hispanic	2 (2)	4 (4)	
Hispanic, Caucasian	1 (1)	0	
Hispanic, race unknown	3 (3)	3 (3)	
Unknown	4 (N/A)	2 (N/A)	
Year of transplant			0.61
2000	0	3 (3)	
2001	2 (2)	4 (4)	
2002	1 (1)	0	
2003	1 (1)	3 (3)	
2004	5 (5)	8 (8)	
2005	15 (15)	11 (11)	
2006	16 (16)	15 (15)	
2007	17 (17)	17 (17)	
2008	11 (11)	8 (8)	
2009	14 (14)	9 (9)	
2010	18 (18)	22 (21)	
2011	2 (2)	3 (3)	
Follow-up among survivors, Months			
N Eval	44	50	
Median (Range)	60 (33-99)	61 (30-123)	0.93

normalized IBD in each region is simulated from an all-by-all pairing of donors and recipients (excluding actual HLA-matched pairs). X and Y chromosomes were excluded from analysis. Removal of low-quality variants due to read misalignment resulted in small broken intervals in the ARD and MHC, explaining lower than expected genetic similarity for HLA-matched donor–recipient pairs within these regions.

Comparisons of donor-recipient variant call format files (Figure 3.2, Figure 3.3, Figure 3.4; Supplementary Figure 3.2) was performed with RTG tools (Cleary et al. 2014) to generate patient-specific variants, which were functionally annotated with snpEff (Cingolani et al. 2012). Sex-mismatched pairs were considered as special cases with Y-chromosome–specific variants in male recipients or genomic locations aligned to paralogous sites on the X chromosome. In all samples, missense and nonsense variants were mapped to their corresponding primary transcript and translated into amino acid sequences for proteasomal cleavage site prediction with netChop 3.1 (Keşmir et al. 2002). MHC binding prediction was performed with netMHCpan 3.0 (M. Nielsen and Andreatta 2016) using patient HLA typing to determine MHC restriction. Ranked peptides were further annotated with minor allele frequencies from dbSNP (NCBI Resource Coordinators 2016) build 147. Acute GVHD usually affects the skin, liver, and gastrointestinal tract (Jacobsohn and Vogelsang 2007). While patient-specific MiHA expression is most informative, collecting these data requires invasive tissue biopsy specimens. Therefore, we opted to corroborate our results with public data from the Genotype-Tissue Expression Project (The GTEx Consortium 2015; GTEx Consortium 2013; Keen and Moore 2015) using previously described methods (Kryuchkova-Mostacci and Robinson-Rechavi 2017) to associate MiHAs with a measure of broad tissue-specific gene expression. The entire workflow is freely available at https://github.com/wwang-nmdp/MiHAIP.

Visualization of X-Y paralogous regions (Figure 3.3A-C) were performed with manual curation using the BLAST-like alignment tool (Kent 2002).

Statistics

P values were calculated using the χ^2 test for Table 3.1 and the Wilcoxon rank sum test with continuity correction for Figure 3.1, Figure 3.2A-C, Figure 3.3(A,D,E), and Figure 3.4B. All other *P* values were calculated using hypergeometric tests with sample and population counts limited to patients with specified MHC restriction. Benjamini-Hochberg false discovery was applied to correct for multiple hypothesis testing in Supplementary Table 3.1. All tests were performed in R with default parameters.

Results

Donor-recipient matching extends beyond five HLA loci

There is strong evidence that HLA-DPB1 T-cell epitope (TCE) matching correlates with allo-HCT outcome (Zino et al. 2004, 2007; Crocchiolo et al. 2009; Katharina Fleischhauer et al. 2012; Shaw et al. 2013; K. Fleischhauer et al. 2014; Pidala et al. 2014). Generally speaking, mismatched alleles between donor and recipient may be benign (permissible) or alloreactive (nonpermissible) in either direction (graft versus host or host versus graft), with clinical consequences that include GVHD or rejection, respectively. Several methods are available to determine the direction and permissibility of HLA-DPB1 mismatching. Although pairs in this cohort were not explicitly matched at this locus at the time of transplant, a retrospective analysis revealed 16%, 60%, 68%, and 76% of donor–recipient pairs were matched by HLA-DPB1 allele, TCE permissibility (Zino et al. 2004), expression (Goyal et al. 2017), or functional distance (Crivello et al. 2015), respectively (Supplementary Figure 3.1). This is consistent with baseline

likelihoods of finding HLA-DPB1 matches with productive 10/10 searches (Tram et al. 2017). We found that HLA-DPB1 allele mismatching did not associate with acute GVHD (P = .92), whereas TCE mismatching did associate as expected (P = .038; 1-sided Fisher's exact test), leaving 56 out of 102 acute GVHD cases (55%) unaccounted for by mismatching at 6 HLA loci.

We hypothesized that HLA-matched unrelated donor-recipient pairs share genetic material outside the MHC. We used IBD inference (Browning and Browning 2013) to measure broad genomic similarity (see Methods), which revealed matching at the MHC regions as expected (Figure 3.1A). Overall, high rates of IBD were observed at the MHC, indicated by many outliers in randomized pairs, which can be attributed to very strong and recent natural selection acting upon these loci in the human population (Albrechtsen, Moltke, and Nielsen 2010). Genetic similarities extended further, albeit to a lesser degree, across chromosome 6 (Figure 3.1B) and genome-wide (P < 2.2e-16; Figure 3.1C). Unexpectedly, there was a single outlier in control experiments where donors and patients were randomly paired. This simulated pair shared 50% of their DNA, likely representing a parent-child or full siblings. To protect confidentiality, we did not analyze the relationship further.

Autosomal MiHAs do not associate with acute GVHD

To investigate patient-specific variation further, we developed an integrative bioinformatics workflow adapted from neoantigen discovery to perform comparative analysis of all HLA-matched donor-recipient pairs regardless of TCE permissibility (Supplementary Figure 3.2).

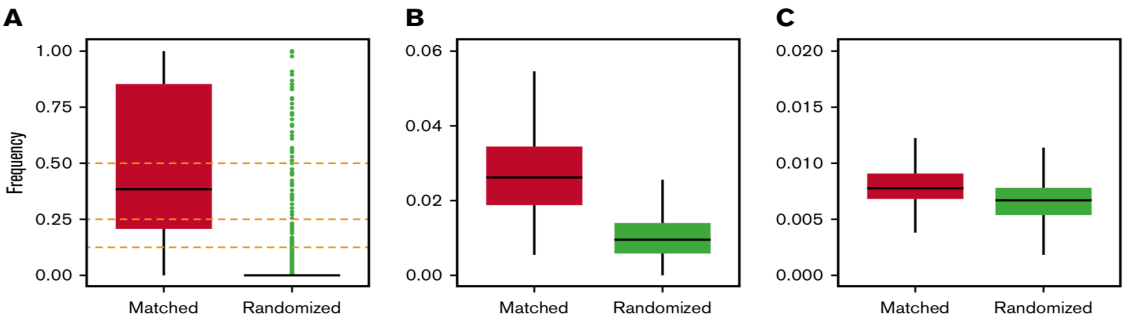


Figure 3.1 Frequency distributions of IBD segments normalized by the total lengths of regions of interest for the following: (A) MHC, including HLA-A, HLA-B, HLA-C, HLA-DR, and HLA-DQ; (B) chromosome 6; (C) and the whole genome. Horizontal black bars represent median values. Outliers (gray) are shown only for panel A (see text for details).

The acute GVHD and non-GVHD groups displayed comparable numbers of missense variants (P = .32; Figure 3.2A) and known MiHAs (P = .80; Figure 3.2B) restricted with patient HLA (P = .76; Figure 3.2C). Ordering MHC-restricted MiHAs by log-ratio (Figure 3.2D) revealed *DPH1* (rs35394823) and *LB-NISCH-1A* (rs887515) as the lowest and highest ranking; however, no associations achieved statistical significance. Thus, we expanded our study to include pre-existing single-nuclear polymorphism microarray data from non-overlapping patient samples. With the addition of 988 HLA-matched donor–recipient pairs (456 acute GVHD, 532 non-GVHD), no statistically significant associations were identified for 17 known MiHAs represented in both data sets (Supplementary Table 3.1).

Y-chromosome–encoded variants associate with acute GVHD

There were 89 sex-mismatched cases in our cohort (Figure 3.3A). Male recipients with female donors (F>M) were more likely to develop acute GVHD (78%) than male

recipients with male donors (52%, $P < .02$) Sequence analysis of the entire Y chromosome of F>M pairs identified only 6 missense variants (relative to the reference genome hg38) encoding a total of 9 variant peptides in 10 out of 21 recipients (48%) with acute GVHD. By contrast, the Y chromosomes of all 6 non-GVHD males matched the reference (Figure 3.3B). The variant peptides were confined to 4 genes: *PCDH11Y*, *USP9Y*, and *UTY*, which have reactive minor histocompatibility epitopes determined *in vitro* (Ofran et al. 2010; Vogt et al. 2000), and *NLGN4Y*, a neuroligin with unknown HCT significance. Except *PCDH11Y*, which is specific to the brain and heart, all genes have broad tissue expression (Supplementary Figure 3.3) and thus make qualified candidates for MiHA presentation in GVHD-affected tissues. Filtering by class I MHC restriction (Figure 3.3C) revealed several variant and reference peptides with strong affinity for their respective HLA allele in both the acute GVHD and non-GVHD groups (Figure 3.3D); however, there were significantly more predicted binders per GVHD male ($P < .015$; Figure 3.3E), suggesting a possible compound effect of multiple Y-linked MiHAs. HLA-DPB1 alone did not explain the association, as 12 out of 21 F>M patients with acute GVHD (57%) were permissibly matched compared with 3 out of 5 without GVHD (60%) ($P < .26$; one patient was not typed at HLA-DBP1).

Paralogous X-Y mismatching explains acute GVHD risk in male recipients with female donors

Risk of chronic GVHD from allo-HCT is higher in male patients with female donors because of B-cell alloreactivity (Popli et al. 2014; Nakasone, Sahaf, and Miklos 2015; Miklos et al. 2005; Sahaf, Yang, and Arai 2013), which is detectable by antibody response that occurs after (Nakasone et al. 2015), but not before, transplant

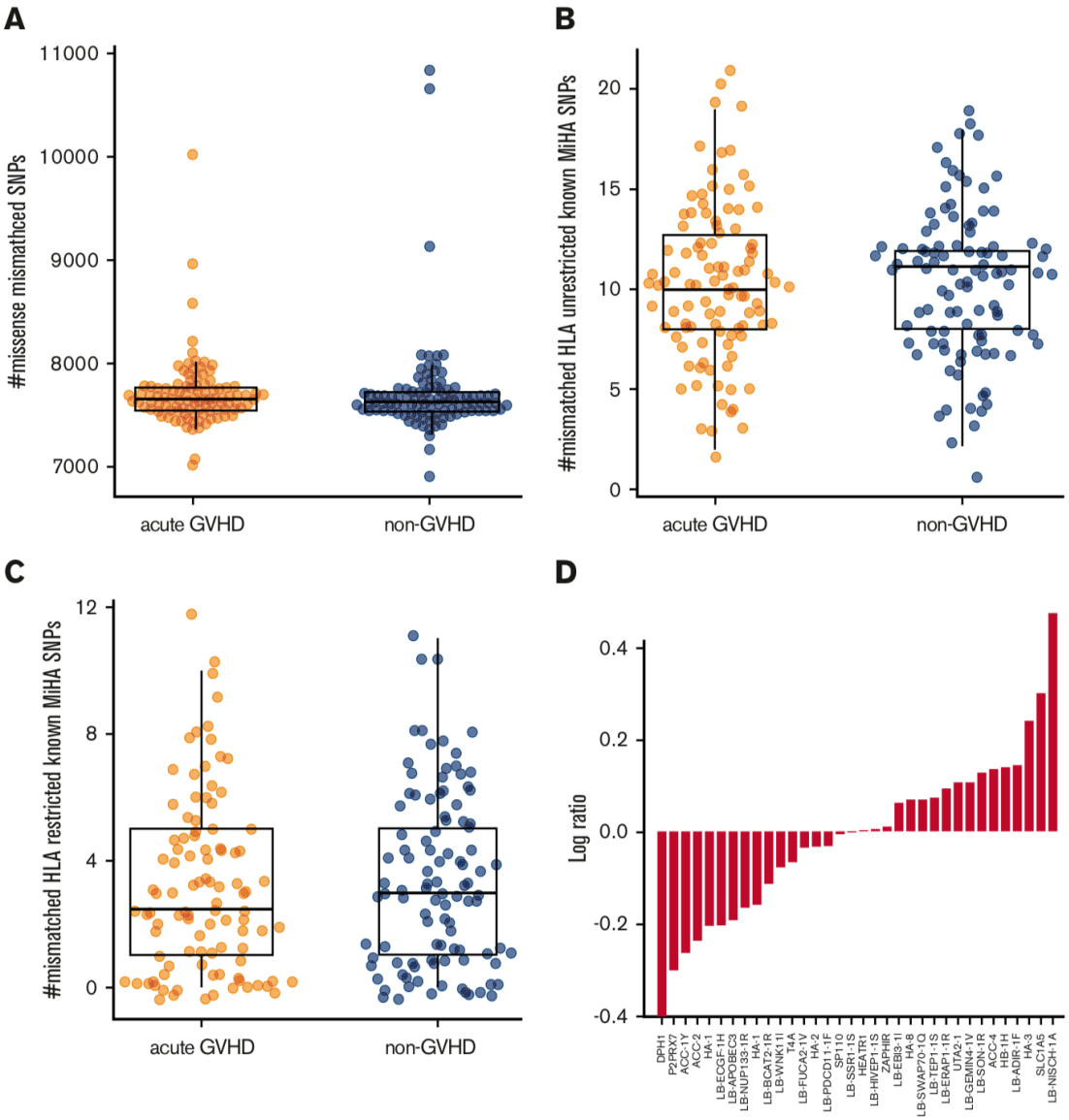


Figure 3.2 Autosomal variants do not associate with GVHD. The number of patient-specific missense variants (A) as well as known unrestricted (B) and HLA-restricted (C) MiHAs is comparable in the acute GVHD and non-GVHD groups. (D) Known, HLA-restricted MiHAs ordered by log-odds ratio (acute GVHD to non-GVHD). SNP, single-nucleotide polymorphism.

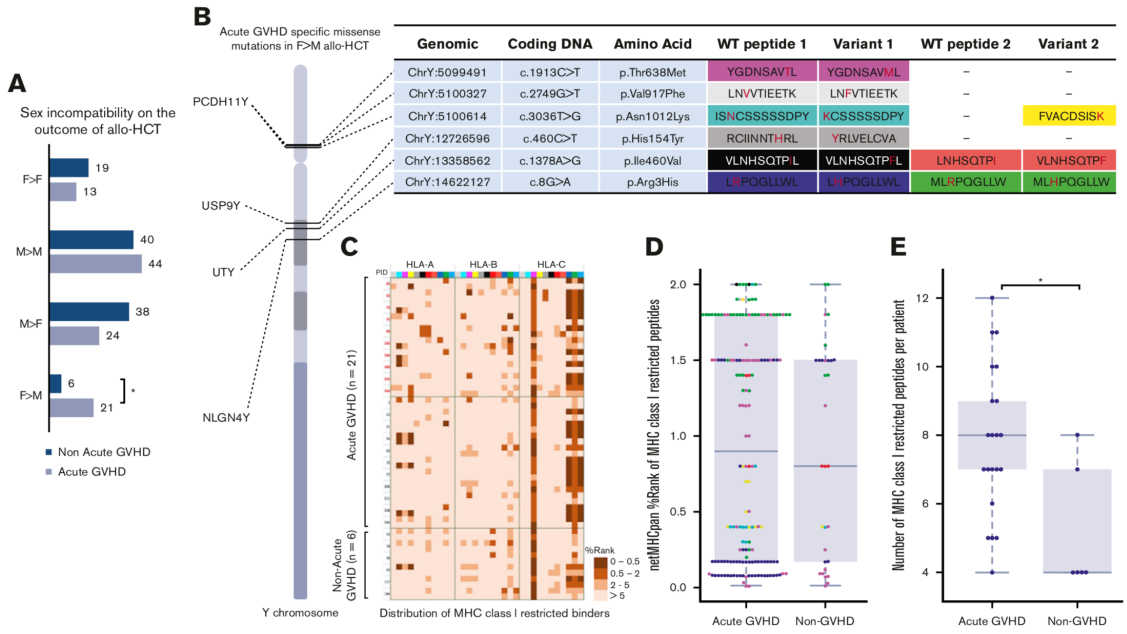


Figure 3.3 Y-chromosome variants associate with acute GVHD. (A) Acute GVHD in sex-matched and sex-mismatched donor–recipient pairs, including a statistically significant association (star) in female-to-male (F>M) allogeneic stem cell transplant. Variants were identified in 4 genes (*PCDH11Y*, *USP9Y*, *UTY*, and *NLGN4Y*), which are displayed with approximate locations on the Y chromosome (B). Precise genomic coordinates and nucleotide and amino acid positions are tabulated, with variant residues shown in red. In some cases, alternative proteasomal cleavage prediction resulted in multiple peptides. (C-E) HLA-restricted affinity prediction for each color-coded peptide is shown for acute GVHD and non-GVHD patients (C) and summarized (D), with application of the recommended threshold for strong binders per male recipient (E). WT, wild-type.

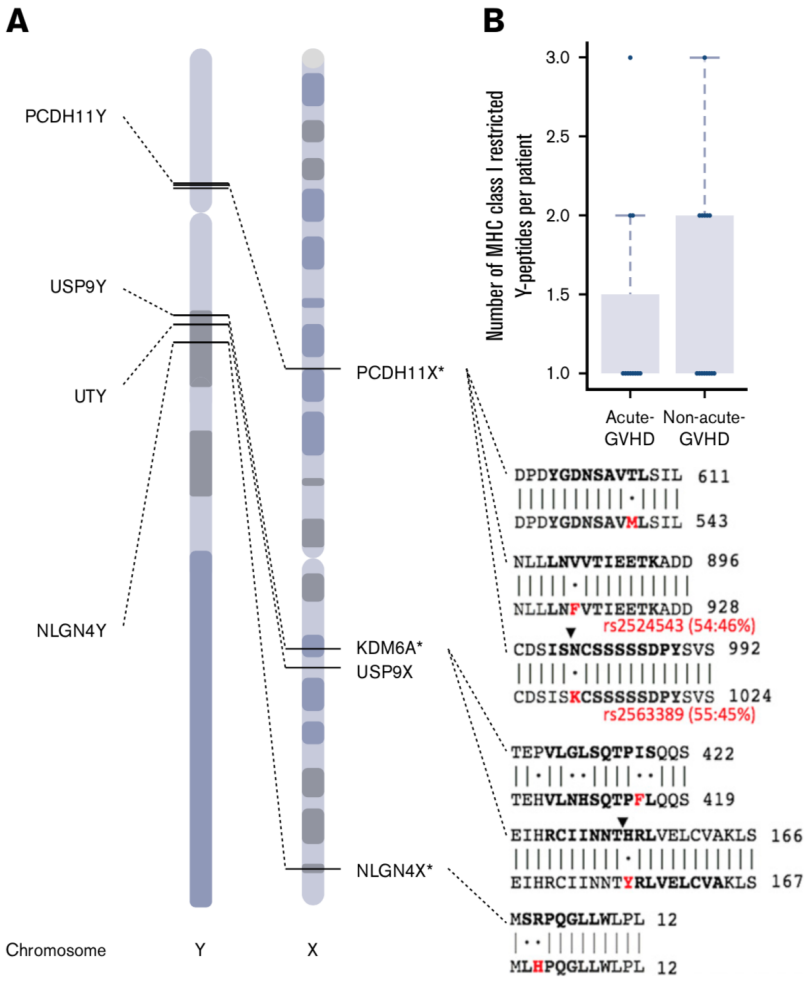


Figure 3.4 Paralogous X-Y mismatching explains acute GVHD risk in male recipients with female donors. (A) Six missense variants on the Y chromosome are exclusive to sex-mismatched male patients with acute GVHD. These variants correspond to 9 variant peptides, which are restricted to 4 genes. Genomic positions for *PCDH11Y*, *USP9Y*, *UTY*, and *NLGN4Y* are shown with dotted lines to paralogous genes on the X chromosome. Protein coding sequence alignments indicating identity matches (bars) and mismatches (dots) are shown for regions that contain individual variant residues (red) and peptides with high-affinity prediction (bold). Ending amino acid coordinates are given to the right of each sequence alignment. Two male-specific variants are named biallelic polymorphisms rs2524543 and rs2563389 with minor allele frequencies 46% and 45%, respectively (red). Note the predicted cleavage sites (black triangles) created by coding variants in *PCDH11Y* and *USP9Y*. *For clarity, other alternative cleavage sites are not shown (see Figure 4 for details), and chromosome X is shown in reverse (39) orientation. (B) The number of predicted high-affinity binding peptides per patient in male-to-male allogeneic HCT recipients with and without acute GVHD.

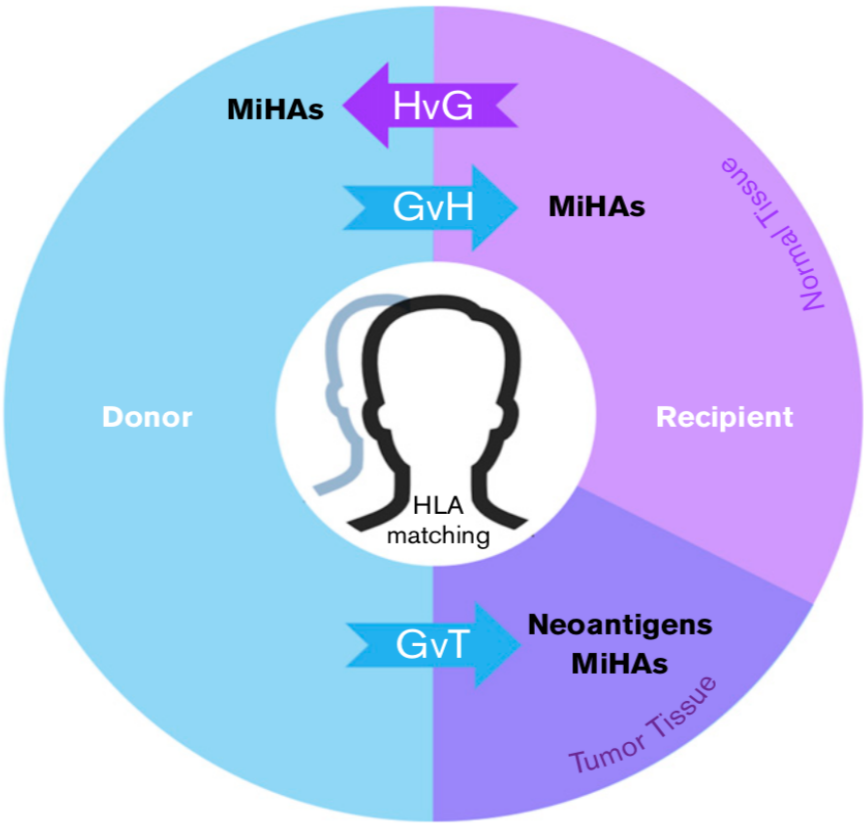


Figure 3.5 MiHAs contribute to the therapeutic benefits and adverse effects of allo-HCT. In the donor-to-recipient direction, germline-encoded variant peptides (some of which may be presented by recipient HLA molecules) are expressed on both normal and tumor tissue and thus may contribute to GVT or graft-versus-host (GvH) effects. Tumor-specific somatic mutations that encode immunoreactive neoantigens contribute to GVT. In the recipient-to-donor direction, MiHAs may have host-versus-graft (HvG) effects in various clinical contexts leading to rejection in HCT and solid organ transplant (SOT) or miscarriage in pregnancy. Matching of HLAs reduces alloreactive responses from donor or host immune systems in transplantation settings. With cord blood HCT, HLA matching is usually performed at fewer (6) loci. This model does not fully illustrate the genomic and immunological complexities of graft predominance with multiple unit infusion. With haploidentical pairs, GvH effects in the recipient are controlled nongenetically with prophylaxis. Predictable patterns of germline inheritance determine match rates at HLA (4/8) and MiHAs (50%) with consequent effects on GVT. The presence and therapeutic benefit of neoantigens (because they are not heritable) are predicted to be independent of graft source.

(Nakasone et al. 2016). Results presented here extend risk in this patient segment to the acute form of GVHD by implicating male-specific variants in four genes. Although definitive clinical recommendations require confirmatory analysis, it is possible to investigate the genetic basis for risk in this cohort.

PCDH11Y, *USP9Y*, *UTY*, and *NLGN4Y* have paralogs on the X chromosome (Lahn, Pearson, and Jegalian 2001) with 72%, 91%, 86%, and 24% amino acid identity, respectively. We mapped Y-encoded variant peptides from each male patient to paralogous sites on their female donor's X chromosomes. All the variant peptides observed in males with acute GVHD mismatched corresponding sites (Figure 3.4A). By contrast, the 6 males without GVHD were X-Y matched at these sites, suggesting their donor-female immune systems were educated, and consequently non-alloreactive, to same-as-self peptides encoded at these positions.

The only other category with increased (albeit statistically insignificant) risk of acute GVHD were male recipients with male donors (Figure 3.3A). Y-Y mismatching was explored as a possible explanation; however, the number of predicted high-affinity binders per patient (maximum 3) was comparable between recipients with and without acute GVHD in the M>M direction (P = .52; Figure 3.4B) and considerably lower than acute GVHD recipients in the F>M direction (maximum 12; Figure 3.3E). Furthermore, all female donors, regardless of recipient sex, lacked variants representing high-affinity binding peptides. These findings associate acute GVHD risk, with explanatory genetic factors, specifically in male patients with female donors, at least in this cohort.

Discussion

Allo-HCT is a curative option for many disorders, yet side effects limit its widespread application. GVHD remains a principal barrier to more effective treatment and improved quality of life, but immune responses that contribute to therapeutic benefit and adverse events are physiologically coupled (Figure 3.5). In malignant conditions, tumor and normal cells are genetically distinct and analytically separable. In the context of allo-HCT, immunoreactive peptides resulting from tumor-specific somatic mutations (neoantigens) may contribute specifically to GVL. On the other hand, the tissue-specific expression and immunogenicity of germline polymorphisms (MiHAs) determine their relative contributions to GVL or GVHD. As treatment options advance, it is important to precisely define genetic factors that affect (or do not affect) clinical outcomes.

Despite research associating several autosomal MiHAs with clinical outcomes, none are routinely matched in allo-HCT. Target tissue expression partially determines the predominance of GVT or GVHD. For example, HA-2 (rs61739531) is expressed in cells of hematopoietic origin (Sellami et al. 2010) where there is evidence for GVL in AML with low risk of GVHD (Mutis et al. 1999). However, expression patterns are not wholly determinant. For example, ZAPHIR (rs2074071) associates with GVT, but not GVHD, in renal cell carcinoma patients receiving non-myleoablative allo-HCT (Broen et al. 2011). Similarly, other ubiquitously expressed MiHAs are associated with GVL in chronic myelogenous leukemia without evidence of GVHD, suggesting complex allo-reactivities from antigen processing, presentation, and costimulation (Griffioen et al. 2012). This is consistent with studies of cancer vaccines where therapeutic benefit results from the

synergistic effects of multiple cancer-specific neoepitopes combined with immune checkpoint blockade (Ott et al. 2017; Sahin et al. 2017). In all cases, MHC restriction is an important qualifier, but filtering patients by HLA reduces the number of samples available for retrospective analysis.

Here, we analyzed common autosomal MiHAs with characterized HLA restriction in separate cohorts of 205 and 988 matched samples. We extended the capabilities of commonly used bioinformatics tools to aid comparative genomic analysis of donor–recipient pairs, incorporating MHC matching and antigen restriction as well as HLA-predicted binding affinity and tissue expression into a common workflow. This study was designed specifically to interrogate acute GVHD in AML patients who were in remission at the time of transplant. Consequently, leukemic cell counts were relatively low, and whole-genome sequences represented primarily germline polymorphism. Thus, bioinformatics analysis focused on MiHAs with broad tissue expression patterns. Future studies will apply these methods to patients with active disease, analyzing somatic variants (possible neoantigens) expressed in cells of hematopoietic origin within larger cohorts that are balanced for GVL-related outcomes including relapse.

Our analysis of autosomes revealed no statistically significant associations with acute GVHD among individual MiHAs. These results confirm a recent genome-wide association study of unrelated allo-HCT where MHC mismatching outweighed other genetic factors as contributors to GVHD risk (Martin et al. 2017). As with neoantigens, it seems plausible that multiple recipient-specific variants contribute to GVHD; however, unlike clonal expansion of somatic mutations in cancer, population-genetic mechanisms

account for the co-occurrence of germline-encoded MiHAs. Indeed, there is evidence that arbitrary HLA-matched donor-recipient pairs may present thousands of MiHAs (Jameson-Lee et al. 2014), which have a cumulative effect on T-cell responses (Razzaq et al. 2016). Although MiHAs are individually common (with minor allele frequencies in our cohort ranging from 19% to 61%; Supplementary Table 3.1), alloreactive combinations may be rare, making it difficult to power case–control studies. Indeed, segmenting patients into subsets sharing ≥2 MiHAs lacked statistical power even when HLA restriction was limited to common alleles. Larger unrelated cohorts are necessary. Additionally, we plan comparable studies in related and haploidentical HCT pairs where shared donor-recipient haplotypes should reduce the number of MiHA combinations under consideration. These studies will also assess whether results reported here are relevant for patients receiving non-calcineurin-based GVHD prophylaxis.

Our comprehensive analysis of sex-linked variation revealed multiple MiHAs encoded on the Y chromosome that associate with acute GVHD specifically in F>M allo-HCT patients. Relative to other chromosomes, the Y is better suited to case–control MiHA association studies, because it lacks population-scale genetic variability due to extremely low rates of diversifying recombination (Wilson Sayres, Lohmueller, and Nielsen 2014). Furthermore, since genetic and phenotypic sex are tightly coupled, it is easy to presegment genomic analysis into clinically weighted categories such as sex match or mismatch. Our limited cohort of primarily white patients suggests the majority of Y haplotypes in this population increase risk of acute GVHD for males with female donors. This is consistent with previous observations of increased chronic GVHD and lower relapse in F>M allo-HCT (Stern et al. 2008; Alois Gratwohl et al. 2009; A. Gratwohl

2012), adding a genetic basis for choosing HLA-matched male donors over nonparous females (Loren et al. 2006; Kollman et al. 2016). However, in cases where a female donor is otherwise the best option for male patients, results reported here may help select a more suitable match.

Chapter 4 Iterative feature selection method to discover predictive variables and interactions for high-dimensional transplant genomic data [2]

Abstract

After allogeneic hematopoietic stem cell transplantation (allo-HCT), donor-derived immune cells can trigger devastating graft-versus-host disease (GVHD). The clinical effects of GVHD are well established; however, genetic mechanisms that contribute to the condition remain unclear. Candidate gene studies and genome-wide association studies have shown promising results, but they are limited to a few functionally derived genes and those with strong main effects. Transplant-related genomic studies examine two individuals simultaneously as a single case, which adds additional analytical challenges. In this study, we propose a hybrid feature selection algorithm, iterative Relief-based algorithm followed by a random forest (iRBA-RF), to reduce the SNPs from the original donor-recipient paired genotype data and select the most predictive SNP sets in association with the phenotypic outcome in question. The proposed method does not assume any main effect of the SNPs; instead, it takes into account the SNP interactions. We applied the iRBA-RF to a cohort (*n*=331) of acute myeloid leukemia

[2] This research has been submitted to a journal and undergoing a peer-review process. "Hu Huang, Cynthia Vierra-Green, Stephen Spellman, Caleb, Kennedy. **Iterative feature selection method to discover predictive variables and interactions for high-dimensional transplant genomic data**. (under review)." BioRxiv preprint DOI: https://doi.org/10.1101/605428.
H.H. conceived the idea of machine learning application, designed and conducted the computational experiments, visualized and interpreted the results, and wrote the manuscript.

(AML) patients and their fully 10 of 10 (HLA-A, -B, -C, -DRB1, and -DQB1) HLA-matched healthy unrelated donors and assessed two case-control scenarios: AML patients vs healthy donor as case vs control and acute GVHD group vs non-GVHD group as case vs control, respectively. The results show that iRBA-RF can efficiently reduce the size of SNPs set down to less than 0.05%. Moreover, the literature review showed that the selected SNPs appear functionally involved in the pathologic pathways of the phenotypic diseases in question, which may potentially explain the underlying mechanisms. This proposed method can effectively and efficiently analyze ultra-high dimensional genomic data and could help provide new insights into the development of transplant-related complications from a genomic perspective.

Keywords: allogeneic stem cell transplantation; whole genome array genotype; acute graft-versus-host disease; acute myeloid leukemia; machine learning; feature selection; Relief-based algorithm (RBA); random forest

Introduction

Acute graft-versus-host disease (GVHD) is one of the major complications after HLA-matching allogeneic hematopoietic stem cell transplantation (allo-HCT) that cause non-relapse morbidity and mortality, affecting up to 40~60% of transplant patients and accounting for 20% of deaths after allogeneic HCT. It is an immunologically mediated complex disease. To date, genome-wide association studies (GWAS) and candidate gene studies have identified SNPs associated with acute GVHD, including SNPs that cause the genetic disparities between the donor and the patient, i.e., the minor histocompatibility antigen (MiHA) single nucleotide polymorphisms (SNPs)(Griffioen, van

Bergen, and Falkenburg 2016), and SNPs that modify gene functions (Petersdorf et al. 2015). However, the genetic risks for acute GVHD outcome have not been well defined yet(Hansen et al. 2010). Most such studies have focused on single locus variants individually or a few candidate gene locations and tested them for association with acute GVHD. Unlike the assumptions of these studies, however, genes tend to interact within specific regulatory and functional pathways, contributing to the disease development.

Next-generation sequencing technologies have enabled affordable high-throughput whole genome microarray genotyping and sequencing. These technologies pose multiple unique challenges in transplant-related genomic studies that need to be addressed and taken into consideration. First, each allo-HCT case involves in two individuals, the donor and the patient, both of whose genomes directly influence the transplant outcomes. Thus, the genomic association models should consider two genomes simultaneously as a single 'sample,' whereas, in common disease association studies, only either the donor or recipient genome is considered as a single sample. Second, the transplant-related outcomes are caused by the genomic disparities between donor and recipient with their synergistic interactions, and hence there is no inheritability of the diseases. Third, the allele frequencies may not play as much of an important role as in the common disease association studies; instead, the combinations and mismatches of donor-recipient (DR) pair genotypes may be more influential. Fourth, the cohorts in the transplant genomic studies are more heterogeneous and harder to control than in common disease studies. Each year, there are limited transplant cases due to the challenges of finding HLA-matching unrelated donors and hence it is harder to recruit groups that share most of the conditions. Furthermore, the cohort size usually is very

small compared to the common disease studies, and this also leads to the lack of publicly available transplant-related genomic databases.

Alloimmune complications after stem cell transplantation, such as acute GVHD, not only involve immune responses to conventional exogenous antigens but also responses to alloantigens. The latter is unique to transplant cases. The major player in GVHD is the activated T cells that recognize and eliminate alloantigens. These T cell functions are influenced by the complex interactions between regulatory networks, pathways, extracellular environment and the unique conditions induced by transplantation procedures (Perkey and Maillard 2018). Thus, it is reasonable to assume that both donor's and recipient's genomes matter in the development of acute GVHD. However, most transplant-related outcome studies often focus on patients' genomes, and very few studies have examined both HLA-matching donor and recipient genomes together (Martin et al. 2017). Here, we assume the donor's genome as equal weight as the recipients and form a paired genotype encoding matrix from each transplant case. With a sufficiently large sample size and appropriate models, we can capture the interacting signals from the paired genome.

Similar to the general whole-genome research in common disease studies as Moore and Ritchie outlined (Moore and Ritchie 2004), transplant-related genomic research also faces three major challenges. The first challenge is to identify meaningful genetic variants along with clinical characteristics that are susceptible to transplant-related complications. The genetic variants include SNPs, genes, or specific gene regions. As described above, transplant-related complications are mostly caused by the genetic

disparities between donor and recipient and the combination of their clinical and demographic characteristics (e.g., sex, age, race, and ethnicity), rather than the disease heritability. The second challenge is to build robust and powerful predictive models that take both genetic and demographic variables into account and output the probability of developing adverse transplant outcomes given a candidate graft characteristic. The predictive models will help facilitate effective and optimized donor search strategies with the best transplant outcomes. While the first two challenges are from statistical and machine learning aspects, the last challenge is to interpret the genetic variants and the predictive models from a biological perspective and further advance our understanding of the transplant-related complications. Biological functional interpretation will help optimize the donor selection process, improve the transplant outcomes and prevent transplant-related complications. It is the most important and difficult challenge and requires a deep understanding of human immunology as well as genetic regulatory mechanisms. Wet lab bench experiments would be the most effective way to validate the hypotheses but it would be too time-consuming and could become implausible if there are too many factors to control. It is one of the current leading translational bioinformatics research focus areas.

Traditional logistic regression models, χ^2-test, odds-ratio are efficient and intuitive when finding simple linear relationships from a large-scale data set; however, they have limited power in modeling high-order non-linear relationships among variables, especially for ultra-high dimensional data. Whole genome microarray genotype data usually cover over 500,000 base pairs of genetic variables and a majority of them may be considered as noise since they do not show any susceptibility to the diseases in question. Data mining

or machine learning techniques build models without any linearity assumptions on the data and can identify the high-order interactive relationship among variables. This is especially attractive to genomic data mining tasks. From a machine learning point of view, there are two main tasks in this context: 1) select the most informative variables from the over 1 million SNPs; 2) predict the disease risk from the selected variables using classifiers. From a clinical point of view, these selected variables should be interpretable. Unlike Mendelian diseases, transplant-related outcomes are influenced by non-linear interactions of multiple genes between donor and recipient. Transplant-related outcomes are more likely a joint effect of multi-factors rather than one single main effect factor. The attribute or feature interaction methods in machine learning seem more appropriate in this case. The data-mining methods can detect nonlinear relationships that traditional regression-based models cannot represent, and this is especially true for dealing with high-dimensional data. In addition, the data-mining algorithms may also uncover the interactions between variables other than their main effects. Applications of machine learning in detecting gene-gene interactions in genetic epidemiology are reviewed in (McKinney et al. 2006; Cordell 2009; Koo et al. 2013).

The purpose of this study is to investigate the application of machine learning techniques in transplant genomics. More specifically, we propose a hybrid feature selection model (iRBA-RF) by incorporating the iterative Relief-based algorithms (iRBA) and a random Forest (RF).

The rest of the paper is organized as follows. First, we define the transplant genomics and outcome association study in the machine learning context. Second, we briefly

review feature selection and classification models. Then we apply the proposed iRBA-RF model to transplant cases to identify critical genetic factors. Lastly, we show the predictive results and provide a possible biological interpretation, as well as the applicability, limitations and future work.

Methods

Problem Definition

In allo-HCT, histocompatibility of stem cells is the primary concern of graft selection, and there are many factors involved in the donor screening process. In this study, we retrospectively investigated 10 of 10 (HLA-A, -B, -C, -DRB1, -DQB1) HLA allele fully-matched unrelated donor transplant cases, and explored the potential genetic variants that may influence the transplant outcomes. In addition to minor histocompatibility antigens (MiHAs), there are other genes involving in regulatory immunological pathways that are critical to the development of GVHD. In complex diseases, there is overwhelming evidence that non-additive synergistic effects of multiple genetic factors play an essential role in the development of the diseases. As described before, we consider the donor genome the same weight as the recipients.

In order to investigate the applicability of the proposed model in the transplant related genomic studies, we assess the following two case-control scenarios: 1) Scenario 1 (AML case-control): acute myeloid leukemia (AML) patients as case and their HLA-matched healthy donors as the healthy control; 2) Scenario 2 (acute GVHD case-control): the donor-recipient (DR) pairs where the patients developed the acute GVHD

symptoms as the case and the DR pairs where the patients did not show any adverse symptoms as the controls.

The main difference between these two scenarios in the context of machine learning is how the genotypes are represented as a feature matrix. Scenario 1 is a common case-control situation where each individual's genotype vector is a single observation, and the AML disease condition is the phenotypic outcome to be predicted. In Scenario 2, an observation is defined as the combined genotype vectors of the recipient and the donor, where the length of the vector is doubled compared to Scenario 1. In addition to the DR genotypes, other clinical characteristics are also included in the model, such as the HLA typing and the donor-recipient sex-mismatch status.

iRBA-RF: a hybrid feature selection model for detecting attribute interactions

In bioinformatics, the "large *p* small *n*" problem is a common challenge, including when it comes to genomic association analysis. The most common problems in genomics data are 1) noisy data 2) heterogeneous data types, and 3) ultra-high dimensional feature space. In machine learning, the feature selection procedure is employed to avoid the "curse of dimensionality" for small samples with high dimensions (Friedman 1997; Domingos 1998, 1999). The objective of feature selection is to select the most relevant feature subset to achieve the best classification/prediction performance without losing the generalization power (accuracy, speed, and generalization). A strong feature relevance indicates the feature is necessary for the predictive model, while an irrelevant feature does not contribute to the predictability. In some cases, the presence of certain

features would decrease the predictability of the model, in which case they are considered as noise. For a formal theoretic derivation of feature relevance, interested readers may refer to (Bell and Wang 2000).

Depending on the feature search strategy and the level of predictive classifier integration, there are three different categories of feature selection methods: filter, wrapper and embedded. Filter approaches are independent of classifiers; instead, they examine the intrinsic properties and relationship between the phenotype in question. Specifically, the information theoretic metrics, such as mutual information (Ding and Peng 2005; Peng, Long, and Ding 2005) and entropy/information gain (Xing et al. 2001; Eom and Zhang 2004), are popular options to measure the intrinsic properties. Since these approaches do not involve training a classifier, they are computationally fast and applicable to a large dataset. Detailed reviews of feature selection techniques in bioinformatics can be found in (Saeys, Inza, and Larrañaga 2007; Bolón-Canedo et al. 2014).

Since we are interested in interpretable variables that are linked to the phenotypes within a reasonable computation time, we adopt the filter-based approaches. More specifically, we propose a hybrid feature selection model that combines an iterative Relief-based algorithm and a random forest (iRBA-RF), to iteratively eliminate the irrelevant features and select the top-ranked features, respectively. In the next subsections, we describe the details of each algorithm.

Iterative RBA for variable elimination

The Relief-based algorithms (RBAs) was inspired by instance-based learning (Aha, Kibler, and Albert 1991; Callan, Fawcett, and Rissland 1991), where it draws instances at random and iteratively compute and updates the weights of features based on their nearest neighbors and their phenotypes. The features that distinguish the selected instance from its neighbors of a different class get more weight. The original Relief algorithm only compares one nearest neighbor of each class, which is sensitive to noisy data and restricted to a binary classification problem. There have been many studies to address the limitations and improve the performance of the original Relief algorithm. The most widely used RBA is ReliefF (Kononenko 1994), which relies on the nearest *k* neighbors, instead of one. By comparing the entire vector of values across all attributes among neighbors, ReliefF can capture the attribute (feature) interactions and has gained popularity in data mining applications. Figure 4.1 shows an example of ReliefF on acute GVHD outcome data set with k =3 nearest neighbors in each class, respectively.

However, ReliefF is not robust to noisy features where it cannot capture the correct signal. An improved ReliefF called Tuned ReliefF (TuRF)(Moore and White 2007) was proposed to iteratively remove features that have low-quality scores, in most cases are noisy features. More extended RBAs were later developed and applied in genomic data analysis, including Spatially Uniform ReliefF (SURF) (Greene et al. 2009), SURF* (Greene et al. 2010), SWRF* (Stokes and Visweswaran 2012), Multiple-Threshold* (MultiSURF*)(Granizo-Mackenzie and Moore 2013), and MultiSURF (Urbanowicz, Olson, et al. 2018). They use different strategies to select neighboring hits and misses

SNP genotypes (features)

Transplant outcome

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	
Observation (O_i)	0	0	2	2	1	2	0	1	1	0	Acute GVHD
Nearest Hit (H_1)	0	0	2	2	1	1	0	1	1	0	Acute GVHD
Nearest Hit (H_2)	0	0	2	2	1	2	0	1	1	0	Acute GVHD
Nearest Hit (H_3)	0	0	2	2	1	1	0	1	1	0	Acute GVHD
Nearest Miss (M_1)	0	1	2	2	2	0	1	0	1	0	Non-GVHD
Nearest Miss (M_2)	0	1	0	1	2	0	1	0	1	0	Non-GVHD
Nearest Miss (M_3)	0	1	1	0	2	0	1	0	1	0	Non-GVHD

$$W^i_{SNP4} = W^{i-1}_{SNP4} - \frac{1}{3n}\sum_{j=1}^{3} diff(SNP4, O_i, H_j) + \frac{1}{3n}\sum_{j=1}^{3} diff(SNP4, O_i, M_j) = W^{i-1}_{SNP4} + \frac{2}{3n}$$

$$W^i_{SNP6} = W^{i-1}_{SNP6} - \frac{1}{3n}\sum_{j=1}^{3} diff(SNP6, O_i, H_j) + \frac{1}{3n}\sum_{j=1}^{3} diff(SNP6, O_i, M_j) = W^{i-1}_{SNP6} - \frac{2}{3n}$$

Figure 4.1 Illustration of ReliefF algorithm with $k = 3$ nearest hits and misses, respectively, on transplant outcome data.

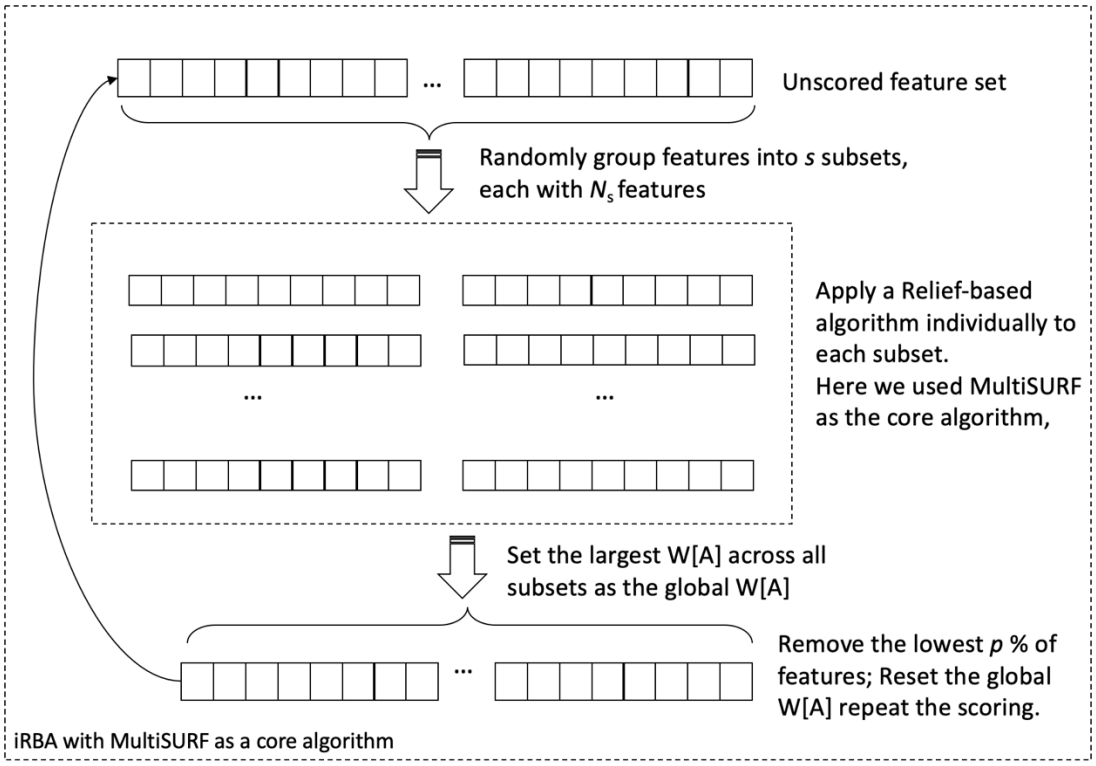


Figure 4.2 Illustration of iRBA, adapted from (Urbanowicz, Meeker, et al. 2018).

and calculate their weights to improve sensitivities and computational efficiency. Furthermore, unlike the original Relief algorithm, these improved versions can handle incomplete data and extend to multi-class problems. For an in-depth review of RBA-based feature selection methods, readers may refer to Urbanowicz et al. (Urbanowicz, Meeker, et al. 2018).

In typical genomic association studies, there are over 500,000 SNPs to be examined. Especially in the context of transplantation, donor-recipient pair genotypes may include over 1 million SNPs. This poses a challenge in computational efficiency. For such ultra-high dimensional genomic data, iterative and efficient approaches that are wrapped around and integrated into the above core RBAs are recommended. VLSReliefF (Eppstein and Haake 2008) algorithm is reported to be able to detect feature interactions in a very large feature space both efficiently and accurately. The main idea is to randomly select s subsets of the feature set with N_s features and individually apply ReliefF to each group to calculate local feature weights. The global weights of each feature are the maximum value of the local feature weights among the subsets. In this study, we follow the framework of VSLReliefF and repeat the process multiple times to remove low-quality features iteratively, as shown in Figure 4.2. Instead of ReliefF, here we choose MultiSURF as the core RBA since it has shown to outperform in multi-way interaction detection as well as various associations, compared to the other RBA algorithms (Urbanowicz, Olson, et al. 2018).

Random forests for feature importance ranking and variance selection

Random forests (RF) are ensembles of tree-structured classifiers that are constructed in the following random fashion: each tree is grown using a bootstrap sample, i.e., aggregated sampling with replacement, of original training set and a randomly chosen subset of features and a majority voting scheme to ensemble individual trees, as illustrated in Figure 4.3 (Breiman 2001). Instead of using the whole set of a training set, each tree is trained on the bootstrapped sample set, and the rest samples are used as a validation set to estimate the tree's classification error. This validation set is called the out-of-bag (OOB) samples. The OOB scheme is used to monitor the generalization error, strength, and correlation of trees in the forests, as well as the variable importance. As more trees added to the RF, it is guaranteed to converge with a limited generalization error and does not suffer from overfitting problem due to *the Law of Large Numbers* (Breiman 2001).

In addition to its effective predictive ability, RFs also measure the importance of the variables in terms of their relevance to the phenotypic outcome. This function has shown great potential in genome-wide association studies and bioinformatic applications due to its effectiveness and potential interpretability. The original RF measures the feature importance using two different metrics.

The first variance importance metric is called Gini impurity index-based feature importance (GIFI). At a node in a tree, the objective is to reduce the class ambiguity as

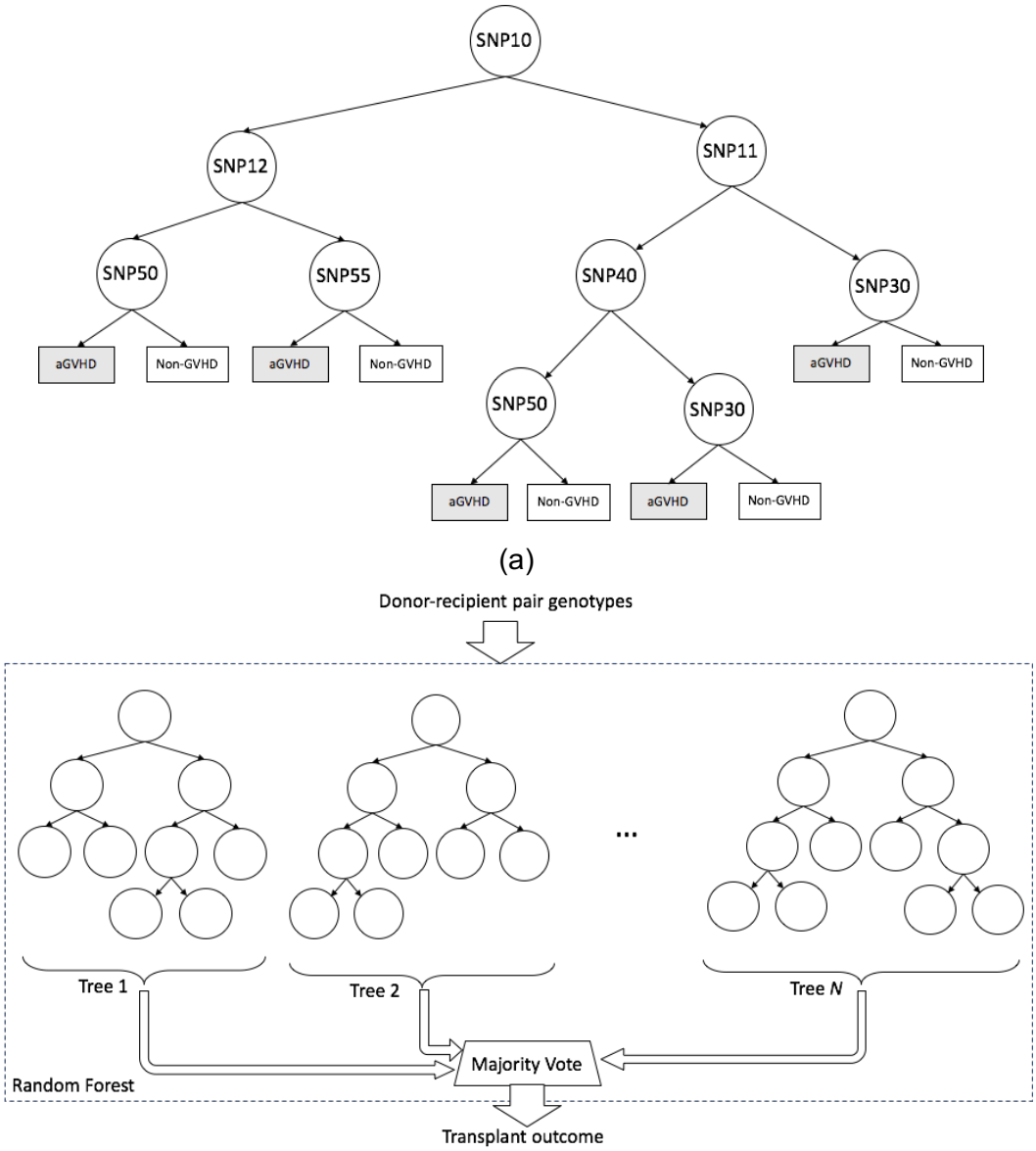


Figure 4.3 Diagram of single decision tree and the random forests. (a) a single decision tree in the forest; (b) Random Forest classifying transplant outcome from the donor-recipient pair genotypes.

the tree grows and the split at a node is determined by the feature that reduces the class ambiguity the most when the sample passes down the split. In RF, the impurity of splits is measured by the Gini impurity index (Breiman et al. 1984), defined as follows: suppose at a node m, N_m observations are trained using feature set $R_m =$

$\{f_{m1}, f_{m2}, \ldots, f_{md}\}$. Write (x_i, y_i) to denote each observation, where x_i has d-dimensional features and y_i is the corresponding outcome label of K possible classes, $y_i \in \{1, 2, \ldots, K\}$. The frequency of class k at node m is defined as

$$p_{mk} = \frac{1}{N_m} \sum_{i=1}^{N_m} I(y_i = k),$$

where $\sum_{k=1}^{K} p_{mk} = 1$. The final class of the observation at the node is determined as $\underset{k}{\operatorname{argmax}}\, p_{mk}$, i.e., the majority class in the node m. For binary classification ($K = 2$), the Gini impurity index is defined as

$$GI(m) = \sum_{k=1}^{2} p_{mk}(1 - p_{mk}) = \sum_{k=1}^{2} p_{mk} - \sum_{k=1}^{2} p_{mk}^2 = 1 - \sum_{k=1}^{2} p_{mk}^2$$

In our case-control cases, there are two classes: AML patient as 1 and healthy donor as 0 for scenario 1; or acute GVHD group as 1 and non-acute GVHD group as 0 for scenario 2. In both cases, the Gini impurity index is

$$GI(m) = p_1(1 - p_1) + p_2(1 - p_2) = 2p_1p_2 = 2p_1(1 - p_1)$$

where p_1 and p_2 are the probabilities of the two classes mentioned above, respectively, and $p_1 + p_2 = 1$.

The GIFI score of a feature in a tree is calculated as the sum of the Gini impurity decrease from a parent node to its children nodes over all nodes in the tree. The GIFI score of a feature in the RF is then defined as the sum (or average) of the Gini importance values among all trees in the forest.

The second feature importance is based on the feature's predictability. After estimating the OOB prediction error during the training phase, the feature values in the OOB data set are randomly permuted and fed into the trained RF. The difference between the OOB prediction error and the permuted prediction error is defined as the prediction-based feature importance. If this value is a large positive value, the corresponding feature has high predictability and is favored high in the ranking; whereas negative or zero values indicate the features are not predictive and thus are discarded in ranking.

It has been shown that both of these metrics suffer a certain degree of selection bias when ranking features. The GIFI favors the features with many possible split points, i.e., categorical variables with many categories or continuous variable (Strobl et al. 2007). In genomic variance selection, it tends to be in favor of SNPs with high minor allele frequencies (MAF) (Nicodemus 2011; Boulesteix et al. 2012). Many studies have proposed correction methods to eliminate bias. Altmann et al. (Altmann et al. 2010) proposed to permute the response (phenotypic outcome) to calculate the null importance distribution while preserving the relationships between features. The algorithm is shown to reduce the feature selection bias induced by the GIFI but also provides the significance level *P*-values for each feature. Later, Janitza et al. (Janitza, Celik, and Boulesteix 2016) proposed an alternative approach to improve the computational speed while correcting the feature selection bias and providing the *P*-values for each feature. Nembrini et al. (Nembrini, König, and Wright 2018) provided a unified framework with a corrected impurity importance measure (AIR) to calculate the GIFI fast and they claimed that AIR outperforms the previous approaches in terms of computational performance and statistical power. All these bias correction methods have been incorporated and

implemented in the R package `ranger` (Wright and Ziegler 2017), and the Altmann-corrected GIFI is adopted in this study.

The prediction error-based feature importance (PEFI) does not have these issues; however, it tends to favor the features that locate closer to the root node since they tend to affect the prediction accuracy of a larger set of observations and the permutation-based importance favors these variables (Strobl et al. 2007). A modified PEFI was proposed by Ishwaran (Ishwaran 2007), where it follows the same procedure as in the original RF, except instead of permuting the features in the out-of-bag data and test on the trained trees from the in-bag data, here the trees are randomized by using left-right random daughter assignment at each of the features. When a case is dropped down to the node with the feature in question, the left and the right daughter nodes of the following lower trees are chosen randomly with the same probability to till it reaches the leaf node. This procedure promotes the poor leaf node values for cases that pass through the nodes that split on the feature.

The predictability of the selected feature set is assessed by using OOB samples with the overall classification error, area under the receiver operating characteristic curve (AUC), and the normalized Brier score defined by Ishwarn and Lu(Ishwaran and Lu 2018). Brier score is more stable than AUC when assessing the classifier performance. A value of 100 normalized Brier score indicates random guessing and 0 being a perfect classifier.

Figure 4.4 shows the proposed iRBA-RF feature selection model. During the first stage, noise and phenotypically irrelevant features are removed through the iRBA using

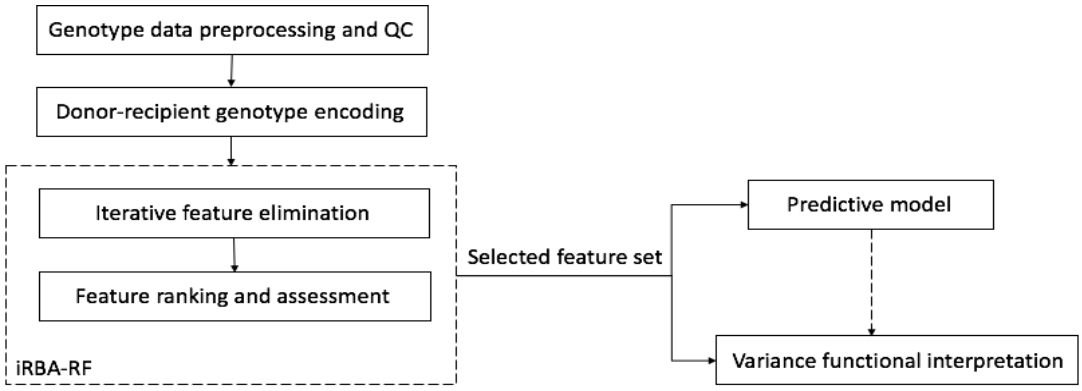


Figure 4.4 Illustration of iRBA-RF feature selection model

MultiSURF as its core RBA. By removing the lowest ranked features, it retains the multi-way interaction relationships between features from MultiSURF. The refined feature set is then fed into the RF model in the second stage. The RF then train models and rank the features through GIFI and/or PEFI metrics. In this study, we implemented the model by incorporating the `scikit-rebate` library written in Python (Urbanowicz, Olson, et al. 2018) (available at https://github.com/EpistasisLab/scikit-rebate) and two random forest R packages, `ranger` (Wright and Ziegler 2017) and `randomForestSRC` (Ishwaran et al. 2008; Ishwaran and Lu 2018).

Data Collection and Preprocessing

A retrospective cohort of blood cancer patients and their HLA matching donors have been selected in this study. The microarray genotype data collection and primary analysis have been described in (Madbouly et al. 2017). In order to reduce the bias induced by disease types and the reference population, we chose AML patients and their transplant cases and used the original genotypes without imputation. After data quality control [Supplementary Material 4.7], 331 transplant cases (662 individuals in

total) of AML patients and HLA matching donors with 630,793 genotyped autosomal SNPs were included in this study. SNPs from the sex chromosome were excluded from this study; however, sex-mismatch conditions were considered as clinical characteristics in Scenario 2 acute GVHD case-control context.

As described in the Methods section, we investigated the iRBA-RF model in two scenarios. In Scenario 1, the formatted genotype matrix has a size of 662×630,793 and the AML disease status as its target label; in Scenario 2, the formatted genotype matrix has a size of 331×1,261,586 and the acute GVHD status as the target label.

Results

Scenario 1: AML case-control experiment

The original 630,793 SNPs were reduced to 200 SNPs through the iRBA-RF and they were further reduced to 176 SNPs and 164 SNPs using GIFI and PEFI, respectively. Table 4.1 shows the top 30 SNPs ranked by the GIFI scores with their significance *P*-values. Of the 176 GIFI-based SNPs, 103 SNPs showed statistically significant scores at the confidence level $\alpha = 0.05$. The full list for the 200 SNPs can be found in the Supplementary Table 4.1.

The PEFI scores are further assessed through the delete-*d* Jackknife subsampling scheme as proposed in (Ishwaran and Lu 2018). Figure 4.5 illustrates the 95% asymptotic normal confidence intervals for the top 50 SNPs ranked by the median PEFI scores. For a full list of features by PEFI, please refer to Supplementary Table 4.2.

Compared to the SNPs listed in Table 4.1, 9 SNPs (rs2694642 (*USP34*), rs928770 (*KCNJ15*), rs10936248, rs2293836 (*NRXN3*), rs6915644 (*EYS*), rs1173099, rs788871, rs17329514, rs675992) are ranked in the top 30 in both cases, whereas 3 SNPs (rs10002187, rs6106323, rs1365342) from Table 4.1 ranked between 31 and 50 in Figure 4.5.

Table 4.1 Top 30 SNPs linked to AML, which are ranked by the Gini impurity importance using the bias-corrected Altmann-GIFI. For illustration purpose, here lists the top 30 SNPs out 200 SNPs from the proposed feature selection model.

Rank	Marker	CHR:POS	Gene(s)	Major	Minor	MAF	Importance score	*p*-values
1	rs2694642	chr2:61369045	*USP34*	A	G	0.315	1.669	0.010
2	rs928770	chr21:38265545	*KCNJ15*	C	T	0.287	0.932	0.010
3	rs10936248	chr3:161818648		C	T	0.383	0.854	0.020
4	rs4698732	chr4:14336718		C	T	0.442	0.832	0.010
5	rs4692262	chr4:27923209	LOC105374552*	C	A	0.372	0.770	0.040
6	rs11869908	chr17:72674911	*SLC39A11*	G	T	0.210	0.768	0.010
7	rs7718156	chr5:172669363	*NEURL1B*	C	A	0.321	0.719	0.010
8	rs2293836	chr14:79840947	*NRXN3*	T	C	0.112	0.694	0.010
9	rs6915644	chr6:65106919	*EYS*	G	A	0.397	0.692	0.010
10	rs10002187	chr4:149994262	*DCLK2***	G	A	0.190	0.689	0.010

11	rs749773	chr2:165877228	*TTC21B*	T	C	0.199	0.644	0.020
12	rs285206	chr20:43667902	*MYBL2*	T	C	0.202	0.605	0.010
13	rs10926025	chr1:239949314	*LOC105373224*	C	T	0.298	0.597	0.010
14	rs12675334	chr8:83887031		A	G	0.396	0.587	0.020
15	rs9819506	chr3:172452314	*GHSR*; BZW1P1**; TNFSF10***; FNDC38***	C	T	0.427	0.578	0.030
16	rs6106323	chr20:2169032	STK35**; LOC105372502**	G	A	0.257	0.574	0.010
17	rs2914290	chr5:7629643	*ADCY2*	C	T	0.181	0.565	0.010
18	rs1365342	chr4:37097911	*LOC101928721*	G	A	0.329	0.564	0.040
19	rs1173099	chr9:90679392	DIRAS2**; OR7E109P***	T	G	0.260	0.556	0.010
20	rs2222514	chr7:123206473	SLC13A*; LYPLA1P1**	A	G	0.432	0.548	0.010
21	rs12714359	chr2:2603252	LOC105373389***; LOC107985839***	A	G	0.367	0.548	0.020
22	rs2185591	chr20:43133279	*PTPRT*	C	A	0.423	0.542	0.020
23	rs788871	chr1:30591356	*MATN1***	T	C	0.474	0.528	0.030

24	rs17329514	chr18:71998994	LOC102725148***; LOC105372189***	A	G	0.101	0.522	0.010
25	rs41135	chr5:96830323	*ERAP1*	G	A	0.369	0.519	0.020
26	rs428148	chr2:70583509	*TGFA*; ADD2****	C	T	0.298	0.518	0.030
27	rs10794031	chr10:125876751	*DHX32*	A	G	0.439	0.517	0.020
28	rs725856	chr4:39746658	*UBE2K*	A	G	0.124	0.510	0.010
29	rs675992	chr1:17888266	ACTL8***	A	G	0.255	0.492	0.040
30	rs2025009	chr14:68376888	*RAD51B*	C	G	0.476	0.491	0.020

*: genes that are within 10 kb range of upstream or downstream from the marker

**: genes that are outside 10 kb but within 50 kb range from the marker

***: genes that are outside 50 kb but within 100 kb range from the marker

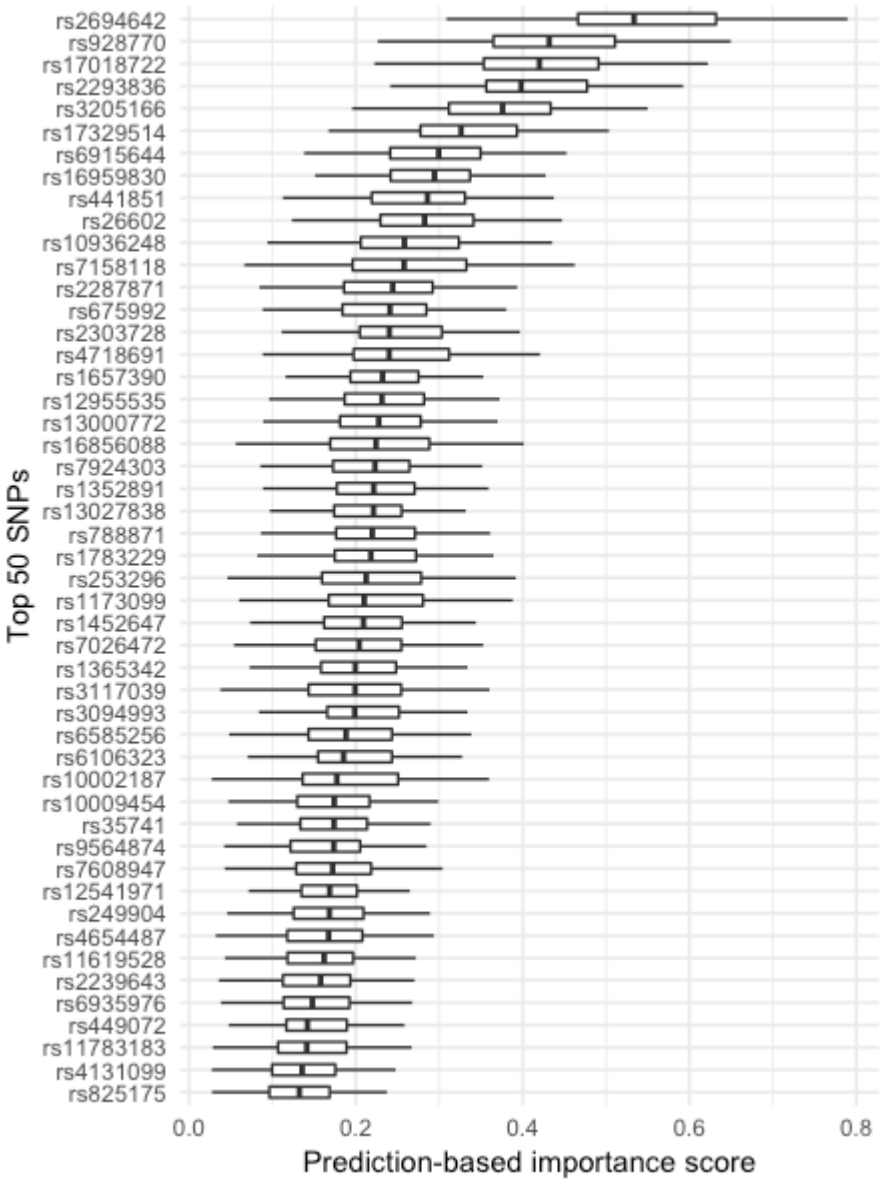


Figure 4.5 Delete-*d* Jackknife 95% asymptotic normal confidence intervals for the top 50 SNPs in the AML case-control scenario. The large positive variance importance values indicate the high predictability of the features, whereas zero and negative values suggest noise variables.

Scenario 2: aGVHD case-control

In the case of acute GVHD, the genotype matrix has twice as many dimensions as Scenario 1, since the donor and recipient genotypes were concatenated in the same

vector for each case. The original genotype matrix has a total of 1,261,586 SNPs, and after the iRBA-RF, the number was reduced to 400 SNPs. The classical HLA typing (HLA-A, -B, -C, -DQB1, -DRB1) and DR pair sex mismatch status are major factors that influence the transplant outcome, and hence these two types of variables were added to the reduced genotype matrix before RF feature ranking. A total of 411 variables (400 SNPs, 10 HLA gene typing, 1 sex mismatch status) were ranked through the RF using GIFI and PEFI metrics, respectively.

342 out of 411 variables were selected through GIFI scores, only 124 of which showed statistically significant scores at the confidence level $\alpha = 0.05$. Top 30 variables by GIFI is listed in Table 4.2, and the full list can be found in Supplementary Table 4.3. Similar to Scenario 1, PEFI scores are assessed through delete-*d* jackknife subsampling procedure and estimated the 95% asymptotic normal confidence interval. 297 variables were selected through PEFI scores, top 50 of which are shown in Figure 4.6. The full list of PEFI features can be found in Supplementary Table 4.4.

Compared to GIFI features in Table 4.2, 6 SNPs [rs10936748 (*LOC105374224*), rs3818283 (*TEK*), rs17161332 (*SGCZ*), rs17236893 (*LOC101928583*), rs10974006, rs2868956] are ranked in the top 30 in both cases, whereas 4 SNPs [rs1341852 (*LOC105370228*), rs11160228, rs4863533, rs504371 (*C6orf118*)] from Table 4.2 ranked between 31 and 50 in Figure 4.6.

Table 4.2 Top 30 variables linked to acute GVHD, which are ranked by the bias-corrected Altmann-GIFI. For illustration purposes, here lists the top 30 variables out 411 SNPs from the iterative feature selection model.

Rank	Marker	CHR: POS	Gene(s)	Major	Minor	MAF	Source	Importance score	*P*-value
1	rs10936748	chr3:173283597	*LOC105374224*	T	G	0.151	recipient	0.562	0.020
2	rs2389923	chr4:119810542	*LINC01365***	A	G	0.246	donor	0.507	0.010
3	rs17172094	chr7:42622588	*LOC105375251*	G	A	0.435	recipient	0.474	0.010
4	rs3818283	chr9:27169126	TEK	C	T	0.252	recipient	0.473	0.010
5	rs4262322	chr8:14839455	*SGCZ*	G	T	0.244	donor	0.472	0.020
6	rs1410267	chr13:97614111	LOC105370324	A	C	0.306	donor	0.471	0.010
7	rs17161332	chr7:78675562	*MAGI2*	T	C	0.069	recipient	0.417	0.010
8	rs1341852	chr13:60350653	*LOC105370228*	A	G	0.449	recipient	0.416	0.010
9	rs7940835	chr11:3244721	*MPGPRE***	T	C	0.140	recipient	0.408	0.010
10	rs7187289	chr16:67933975	*PSMB10*; CTRL*; PSKH1*; LCAT*; SLC12A4**	A	C	0.320	donor	0.406	0.010
11	rs4638670	chr18:27701183	LOC105372042***	A	C	0.174	recipient	0.404	0.010

12	rs354843	chr4:141267650	*ZNF330**; RNF150***	T	C	0.269	donor	0.400	0.010
13	rs719910	chr7:42722647	*LINC01448***	T	C	0.399	recipient	0.395	0.010
14	rs6461551	chr7:21312333		G	A	0.383	recipient	0.391	0.010
15	rs17236893	chr3:170745833	*LOC101928583*	A	G	0.095	recipient	0.367	0.010
16	rs10974006	chr9:38738297		G	T	0.257	recipient	0.364	0.020
17	rs273592	chr11:30857302	*LOC107984419*	A	G	0.482	recipient	0.361	0.010
18	rs2481955	chr13:28009444	*FLT3*	G	A	0.378	donor	0.360	0.010
19	rs227130	chr20:8452312	*PLCB1*	G	A	0.370	recipient	0.350	0.020
20	rs11160228	chr14:95051806	DICER1**	G	A	0.242	recipient	0.336	0.010
21	rs4863533	chr4:137992400	*LOC107986315**; LOC105377447**; LINC00616**; SLC7A11****	G	A	0.270	recipient	0.333	0.010
22	rs2868956	chr19:28320511	*LOC107985269***	T	C	0.235	recipient	0.329	0.020
23	rs13035654	chr2:139894716		T	C	0.182	recipient	0.328	0.020
24	rs509012	chr13:2172052	*FGF9***	G	A	0.225	donor	0.316	0.020

		4							
25	rs10503960	chr8:34340069	*RPL10AP3****	A	C	0.181	donor	0.314	0.010
26	rs12203592	chr6:396321	*IRF4*	C	T	0.037	recipient	0.304	0.010
27	rs12763563	chr10:12989971	*CCDC3*	G	A	0.220	recipient	0.301	0.010
28	rs4685366	chr3:16614824	DAZL*	A	G	0.445	donor	0.301	0.030
29	rs504371	chr6:165310563	*C6orf118; LOC105378113*	G	T	0.431	recipient	0.298	0.010
30	rs2161495	chr5:103750446	*LOC105379107*	C	T	0.313	recipient	0.297	0.010

*: genes that are within 10 kb range of upstream or downstream from the marker

**: genes that are outside 10 kb but within 50 kb range from the marker

***: genes that are outside 50 kb but within 100 kb range from the marker

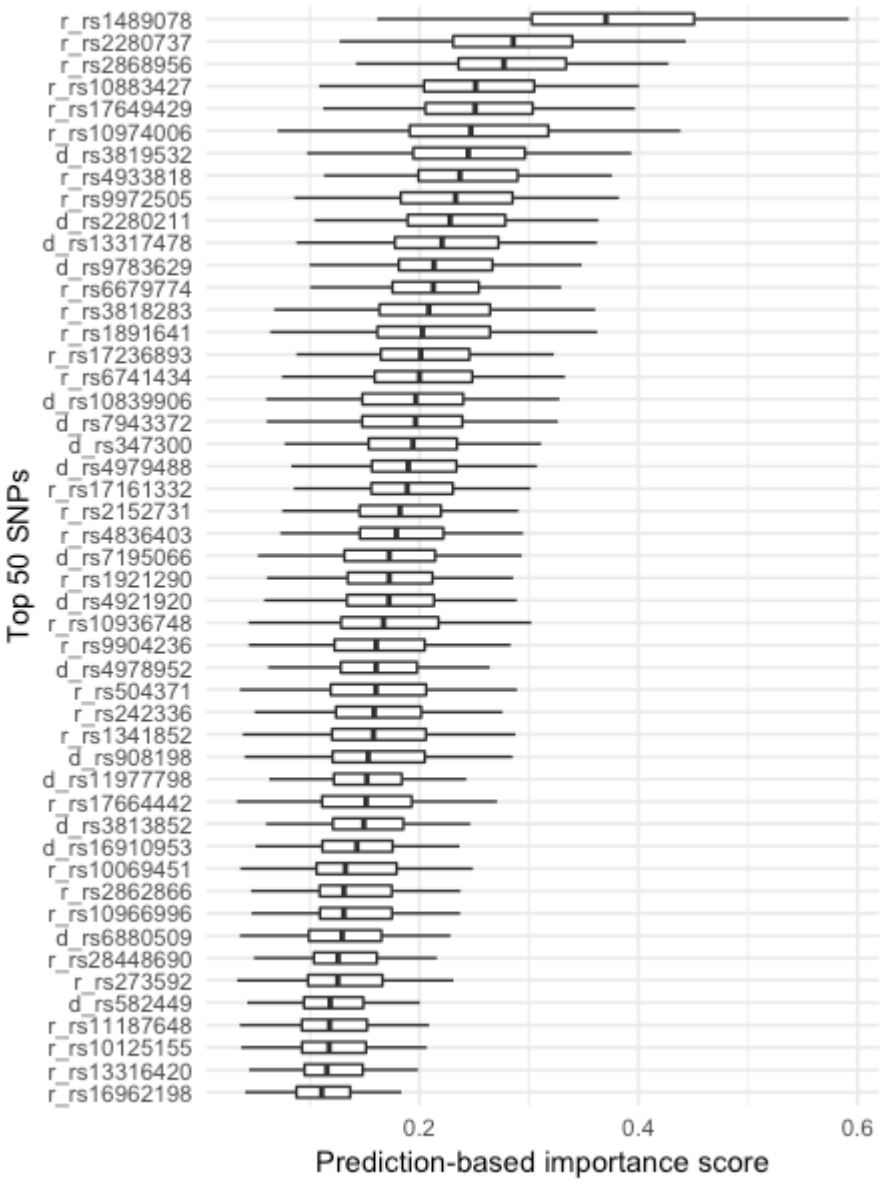
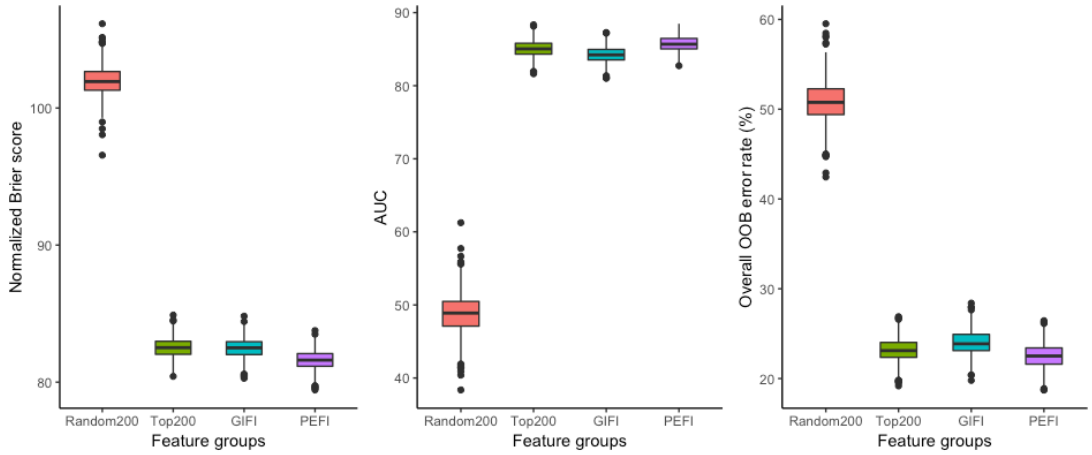


Figure 4.6 Delete-*d* Jackknife 95% asymptotic normal confidence intervals for the top 50 SNPs in the acute GHVD case-control scenario. The large positive variance importance values indicate the high predictability of the features, whereas zero and negative values suggest noise variables. 'r_' indicates SNPs from recipients, while 'd_' for SNPs from donors.

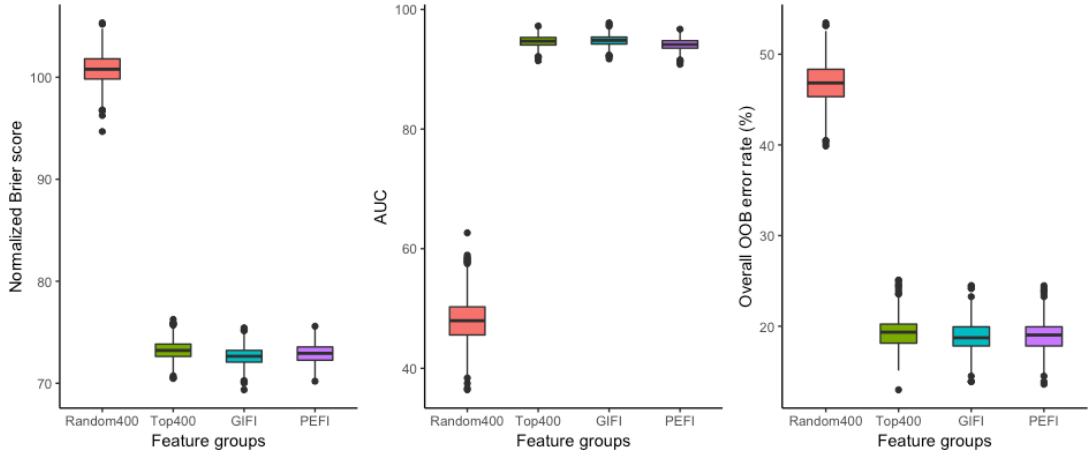
Discussions

Ideally, the features that are selected by iRBA-RF have the best predictability on the phenotypic outcomes and the related gene regions are actively involved in the pathways of the diseases in question. To assess the results, we first examined the predictability of the selected feature sets by comparing the classification performance to a random set of features with the same size. The classification performance was examined in three criteria: the normalized Brier score where a score of 100 indicates a random guess and the lower, the better classifier performance, the AUC, and the overall OOB error rate. Figure 4.7 shows the comparison among different feature sets. The random feature sets (Random200/Random400) were selected 1000 times, while the rest feature groups were trained and tested on OOB samples 1000 times. The selected features through iRBA-RF (Top200/Top400, GIFI, PEFI) in both scenarios show significantly superior classification performance in all three criteria (p < 2.2e-16). Within the selected groups (Top200/Top400, GIFI, PEFI), the pairwise t-tests showed a significant difference between each group using all three criteria (p < 0.001), except for the Brier scores between Top200 and GIFI groups. As shown in Figure 4.7, the classifiers using the PEFI-based feature sets generally showed better predictive performance, and this is mainly because PEFI-based features are ranked based on the classifier's performance.

From the functional point of view, the top-ranked features are not random and evidence can be found in the literature. In Scenario 1, multiple SNPs among the selected 200 SNPs are from the following functional gene groups that are reported to be linked to AML (Cancer Genome Atlas Research Network et al. 2013; Peker 2018). These gene



(a) Scenario 1: AML case-control



(b) Scenario 2: aGVHD vs non-GVHD

Figure 4.7 (a) Scenario 1: AML case control. (b) Scenario 2: aGVHD vs non-GVHD. Comparison of four different sets of features: (1) Random200 (or Random400): 200 (or 400) features randomly selected from the original feature set. (2) Top200 (or Top400): top ranking 200 (or 400) features selected by the iRBA-RF algorithm (3) GIFI: 176 (or 342) SNPs out of the selected 200 (or 400) SNPs using the GIFI score, (4) PEFI: 164 (or 297) features out of the selected 200 (or 400) SNPs using the PEFI score. Each feature sets were trained 1000 times and evaluated by the normalized Brier scores, AUC and overall error rate of OOB samples, respectively.

groups include spliceosome (rs10794031, rs3205166), cohesin complex (rs2025009), epigenetic modifiers (rs1987193), serine/threonine kinase (rs10002187, rs994502, rs13000880), protein tyrosine phosphatases (rs2185591) and other myeloid transcription

factors (rs285206). Most of these SNPs didn't show significant importance scores, however, *SLC39A11* (rs11869908), *MYBL2* (rs285206), *PTPRT* (rs2185591), *DHX32* (rs10794031), *RAD15B* (rs2025009) are ranked the top 30 GIFI with $p < 0.05$.

SLC39A11 (rs11869908), also known as *ZIP11*, is a zinc transporter gene that has been reported to be linked to multiple cancers (Pan et al. 2017). Specifically, a high expression of *ZIP4* and low expression of *ZIP11* are significantly associated with the higher grade of Glioma (Kang et al. 2015). In addition, mutations in *IDH1* is reported to be highly correlated with higher expression of *ZIP11*, suggesting a possible synergistic interaction between *IDH1* and *ZIP11* (Kang et al. 2015). *MYBL2* (rs285206) has an essential role in cell cycle progression, cell survival, and cell differentiation, and is found to be overexpressed in multiple cancer cases (Musa et al. 2017). Overexpression of *MYBL2* is suggested to have a prognostic value for disease-free survival and cumulative incidence of remission for AML patients (Fuster et al. 2013; Musa et al. 2017). *PTPRT* (rs2185591) encodes cellular signaling proteins that regulate cell growth, differentiation, and oncogenic transformation and is reported to be in the genetic interaction network of AML mutational landscape (J. W. Lee et al. 2007; Becker et al. 2014; Ibáñez et al. 2016). *DHX32* is an RNA helicase that is reported to associate with the acute lymphoblastic leukemia (Abdelhaleem 2002, 2005). Notably, its dysregulation is believed to contribute to carcinogenesis (Alli, Ho, and Abdelhaleem 2005). *RAD51B* (rs2025009) encodes one of the *RAD51* paralogs that participate in DNA repair, and the polymorphisms in the gene and the gene inactivation through chromosome translocation have been demonstrated to be linked to AML, breast cancer, head and neck cancer

(Nowacka-Zawisza et al. 2015; Miao et al. 2015; Rollinson et al. 2007; Cheng et al. 2014).

Other genes ranked high in the list have also shown evidence of roles in AML linked pathways. For instance, the top-ranked gene *USP34* (rs2694642) is reported to regulate the levels of axin and stabilize beta-catenin and further modulate Wnt signaling pathway positively (Lui et al. 2011). Wnt/*β*-catenin pathway has shown to be essential in AML for leukemia stem cells to develop and thus allow malignant progression (Y. Wang et al. 2010; Müller-Tidow et al. 2004; Holland et al. 2013; Tickenbrock et al. 2008; Reya et al. 2003; Reya and Clevers 2005). *KCNJ15* (rs928770) encodes potassium inwardly-rectifying channel on the cell membrane and is reportedly a susceptible gene for Type 2 diabetes(Okamoto et al. 2010, 2012) and linked to the hematological traits and clinical features of Down syndrome (Felipe et al. 2006; J.-B. Lee et al. 2016; Canzonetta et al. 2012). *NEURL1B* is a paralog of *NEURL1,* and the deletion of this gene region has been linked to adult *de novo* AML (Cancer Genome Atlas Research Network et al. 2013).

In the case of acute GVHD, most of the top-ranked SNPs do not lie in a gene region; however, they are within 50 kb range of downstream or upstream of the coding genes. Interestingly, multiple gene regions from donors are ranked high in both GIFI- and PEFI-based feature set, suggesting the potential role of genetic polymorphisms from graft stem cells in the transplant outcomes. Genes that are linked to abnormalities of skin and the gastrointestinal (GI) tract are also selected by the iRBA-RF algorithm, all of which are the main symptoms of acute GVHD.

Notably, rs7187289 (donor) is located on chromosome 16q22.1, where five genes (*PSMB10*, *CTRL*, *PSKH1*, *LCAT*, *SLC12A4*) tightly clustered together(Larsen et al. 1993). It is in the upstream of *PSMB10*, an immunoproteasome subunit that plays a vital role in major histocompatibility complex (MHC) class I restricted antigen processing and presentation(Nagata et al. 2003), T-cell polarization and differentiation, and cytokine production by macrophages(Kimura et al. 2015). Hence, *PSMB10* is believed to involve in the development of inflammatory autoimmune diseases and hematologic malignancies and to be a marker of cell damage and immunological activity (Csizmar, Kim, and Sachs 2016). In renal transplantation, it has recently reported being associated with chronic antibody-mediated rejection (AMR) and posed as a potential intragraft and peripheral blood marker of acute rejection (Ashton-Chess et al. 2010; Iwase et al. 2011). The impairment of immunoproteasome subunits is critical for malignant cells to escape immune recognition, suggesting its possible role in the graft-versus-tumor effect after allo-HCT. LCAT is secreted by the liver and generally believed to maintain the unesterified cholesterol gradient between peripheral cells and high-density lipoprotein (HDL)(Asztalos et al. 2007). *SLC12A4* encodes a human potassium chloride cotransporter 1 (KCC1) (Zhou et al. 2004), and the dysfunction of the membrane ion channels has been reported to link to several diseases, like sickle cell disease (Kato et al. 2018).

Several gene regions in the list have direct roles in the signs of acute GVHD in the GI tract. *CTRL* is a chymotrypsin-like protease expressed in the pancreas and secreted in pancreatic juice (Whitcomb and Lowe 2007) and is well known to be downregulated in pancreatic cancer (Laurell et al. 2006). PSKH1 protein is mainly found in Golgi

apparatus, endoplasmic reticulum (ER), nucleus, cell membrane, cytoskeleton (G. Brede et al. 2000) and believed to play a role in intranuclear serine/arginine-rich domain (SR protein) trafficking and pre-mRNA processing (Gaute Brede, Solheim, and Prydz 2002). A recent study suggested it possibly linked to the pathogenesis of Crohn's disease (Iborra et al. 2018). *MAGI2* (rs17161332) encodes a scaffolding protein that involved in epithelial integrity, and studies have shown that the genetic variation in *MAGI2* is linked to the inflammatory bowel disease (IBD), i.e., ulcerative colitis and Crohn's disease (McGovern et al. 2009). Moreover, *IRF4* (rs12203592) controls T_{H2} (type 2 T helper cell) responses and intestinal Th17 cell differentiation, mucosal cytokine IL-17 regulation, suggesting a central of *IRF4* in immune regulation in the gut (Messmann et al. 2015; Cretney et al. 2011; Zheng et al. 2009; Huber et al. 2008; Persson et al. 2013; Schlitzer et al. 2013).

The risk of getting a secondary solid cancer following allo-HCT is substantially higher than in general population, and the risk factors have been well documented (Curtis et al. 1997; Inamoto et al. 2015; Tichelli et al. 2018). Melanoma, breast cancer, thyroid cancer, prostate cancer, and cervix cancer are the most frequently occurring cancer types after allo-HCT in the recipient's later life. The genes that were selected in this study have evidence to link to the cancer progression and may be able to explain the incidents. The feature sets after iRBA-RF include multiple genes that play critical roles in the progression of carcinogenesis. For example, *SGCZ* (rs4262322) encodes a protein that plays a role in maintaining cell membrane stability and have been reported its role linked to cancer development and progression (Chi, Murphy, and Hu 2018). *DICER1* (rs11160228) encodes essential proteins for a micro-RNA processing pathway and plays

a central role in epigenetic modulation of gene expression, and downregulation of DICER1 expression has been reported to be linked to a wide range of cancer types(Bahubeshi, Tischkowitz, and Foulkes 2011; Radom-Aizik et al. 2010).

A few leukocyte-specific genes, such as *TEK*, *FLT3,* and *PLCB1* are also ranked high in the list. *TEK* (rs3818283) encodes angiopoietin-1 receptor that is critical to the induction and growth of new blood vessels and influence tumor growth. It has been reported that mutations in *TEK* are linked to AML suggesting its essential role in leukemogenesis depending on an uncharacterized cellular context (Tyner et al. 2009; De Palma et al. 2005). A more recent study demonstrated that angiogenesis precedes leukocyte infiltration during inflammation suggesting the essential involvement of angiogenesis in the initiation of inflammatory diseases, such as acute GVHD and IBD (Riesner et al. 2017). Proteins encoded by *FLT3* (rs2481955) stimulates hematopoiesis and is reportedly expressed at high levels in a spectrum of hematologic malignancies, including AML. Mutations in *FLT3* usually lead to a poor prognosis and is suggested to be a potential therapeutic target for kinase inhibitor (Gilliland and Griffin 2002). Similarly, *PLCB1* (rs227130) is recently proposed as a potential therapeutic target for AML patients, since the monoallelic deletion or increased PLCB1 expression is a prognostic factor and is reportedly linked to the transition from myelodysplastic syndromes (MDS) to AML (Ratti et al. 2018; Damm et al. 2010; Fiume et al. 2014).

These genes and their functional interpretation seem to explain potential underlying mechanisms; however, they by no means explain the actual biological functions nor do they provide a full picture of the genetic interactions in AML or acute GVHD. The SNPs

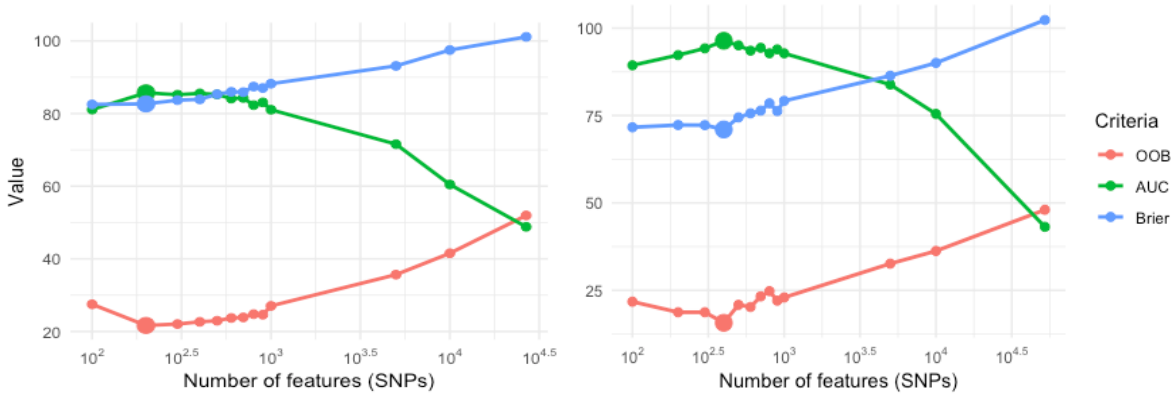
used in this study are common polymorphisms (MAF>0.005), and the locations are much sparser than the whole genome sequences. Thus, the representation of genes from the SNPs is merely a remote approximation. The highly ranked SNPs are not necessarily directly involved in the pathways linked to the disease; instead, some ungenotyped genes that share a high linkage disequilibrium with those SNPs may exert more significant influence on the disease status. On the other hand, the feature interactions captured through iRBA-RF suggest the statistical epistasis among these features or SNPs but not the biological epistasis. Therefore, functional interpretation from SNP sets needs much careful consideration. More rigorous experiments may be needed to validate the potential genetic interactions. Overall, it is a promising start to investigate the genetic interactions in transplant-related outcome studies while considering both donor's and recipient's genomes simultaneously.

One of the advantages of using GIFI and PEFI-ranked features in RF is that it removes the arbitrariness of choosing the number of features. Positive values of GIFI or PEFI indicate the features contribute positively to the predictive power of the predictive models and a value of zero is an appropriate cutoff. The *p*-values of GIFI scores add additional confidence level to the selection, as do the confidence intervals to the PEFI scores. On the other hand, GIFI ranking and PEFI ranking are not always consistent with each other, as they are using two different criteria to measure the feature importance. Moreover, as many researchers have point out (Amrhein, Korner-Nievergelt, and Roth 2017; Wellek 2017), a *p*-value is not a reliable metric to infer the significance, since it depends on assumptions of the models and the sample size and lacks reproducibility. Lu et al. (M. Lu and Ishwaran 2018) proposed standard error and confidence intervals of

PEFI as an alternative to the *p*-values of regression models and demonstrated the robustness of PEFI to the sample size and model assumptions. Especially, when the sample size is small, it is a more reliable indicator than *p*-values. Therefore, it is desirable to use the PEFI ranked feature set for further downstream analysis.

There are several parameters to determine for the iRBA-RF model. The first step of feature elimination using iRBA requires four parameters: the optimal subset size (*Ns*), the number of iterations (*Iteration*), the percentage of features that will be removed after each iteration (*pct*), and the number of features to keep after the last iteration (*featNum*). The original VLSReliefF algorithm suggested a large sample size and relatively small *Ns* achieve reliable results (Eppstein and Haake 2008). By default, top 50th percentile (*pct*=0.5) rank of SNPs is selected after each iteration. However, in our experiment, this removes many interactive and relevant variables, and the final feature sets have very little predictive power. The goal of iRBA is to remove as many irrelevant variants as possible while keeping all possible interacting ones. In this study, we chose the following parameters for both scenarios: *Ns*=1000, *Iteration*=5, *pct*=0.25. As for *featNum*, we employed a grid search strategy to find the optimum values for each of the scenarios. As shown in Figure 4.8, *featNum*=200 for Scenario 1 and *featNum*=400 for Scenario 2 achieved the best classification performance, measured by the normalized Brier score, AUC and overall OOB error rate. As for RF models, each forest has 300 trees (*ntree*=300) with the default *mtry*=sqrt(#dimensions).

One caveat of the iRBA-RF model is that, unlike regression or a simple associations test, it cannot tell the direction of a variable's impact on the disease (positive or negative,



(a) AML case-control scenario (b) Acute GVHD scenario

Figure 4.8 The effect of the number of SNPs. from 10,000, 5000, 1000, 900, 700, 500, 300, 200, 100: the OOB prediction error reaches its minimum between 300 and 500 SNPs for both AML and acute GVHD. Number of features are shown on a $\log_{10}$ scale. OOB: Out-of-bag sample average prediction error, shown in percentage (%); AUC: Area under the ROC; Brier: the normalized Brier score. The optimal feature size would produce the minimum OOB error rate (%), the minimum Brier score and the maximum AUC value.

protective or progressive) from the model. However, this may avoid or minimize the effect of Simpson's paradox, where the positive or negative association of variables reverse sign due to the change of a confounding factor. Simpson's paradox is a common issue in association studies, especially in a high-dimensional bioinformatics data set (Freitas 2019). Keep in mind that, the selected features collectively contribute to the predictive power and thus it is not an indication of the influence of a single SNPs on the disease. It is worth noting that the importance scores and feature ranking are a relative concept, and they have little indication of biological importance. In summary, the iRBA-RF model offers a computationally efficient and functionally effective method to find the candidate variable groups whose interactions may exert to the disease status. With a larger cohort size with broader genomic coverage, it may effectively find the genetic interaction networks that are directly linked to the disease development.

Conclusion

We developed a hybrid feature selection model, iRBA-RF, to reduce the feature space and ultimately rank and select the variants that may be linked to the diseases in question, AML, and acute GVHD. The proposed model successfully selected the most related SNPs out of over 600 K and 1 M SNPs and produced a reasonable predictive accuracy. The model was applied to genomic data in this study, but it can be extended to examine multi-omics data with other clinical characteristics, as well as the multi-class prediction problems.

As discussed above, evidence of the genes can be found in the literature to be linked to the disease in question; however, in order to determine their biological role and further assist optimized donor selection process and personalized therapeutic development, experiments on a larger cohort size, along with immunological wet lab validation experiments on the selected genes, are desired.

Chapter 5 Discussions and Conclusion

Allo-HCT is a curative option for many hematologic malignancies and inherited or acquired genetic disorders, yet side effects limit its broader application. GVHD remains a principal barrier to more effective treatment, and immune responses that contribute to therapeutic benefit and adverse events are physiologically coupled. As personalized therapeutic treatment options advance, it is essential to precisely identify genetic factors that affect clinical outcomes, either protectively or aggressively.

In this dissertation, we focused on acute GVHD after 10/10 HLA matched (HLA-A, -B, -C, -DRB1, and -DQB1) allo-HCT from a genomic perspective. MHC genes are the most polymorphic known genes known and HLA matching is one of the most critical factors in the process of optimal graft selection. Here, we showed the sequence diversity of the classical HLA genes that are used in current clinical settings and provided evidence of high sequence variations outside the ARD exons. Despite the relatively large number of HLA matched donor-recipient pairs with their full-length classical HLA gene sequences, the current study cohort was not able to provide statistically significant results to support transplant outcome associations of non-ARD region mismatches. However, due to the low frequency of amino acid mismatches in the non-ARD region and their reportedly weak alloimmune reactions, we suggest that the non-ARD sequence mismatches within the ARD-matched DR pairs have limited influence on the development of post-transplant complications, such as acute GVHD, and may not be a primary factor. Meanwhile, the HLA gene haplotype mismatches and their clinical association is another interesting area

to investigate, as many studies have shown evidence of better outcomes in HLA haploidentical allo-HCT.

Like many other common complex diseases, acute GVHD and other transplant-related complications involve many other genes. We first investigated the missense variant mismatches between donor and recipient from the whole genome sequences to identify genes that encode MiHAs. The clinical outcome study on the identified autosomal MiHAs found no statistically significant association with acute GVHD outcome, which may suggest that MHC mismatching outweighs other genetic mismatches as contributors to acute GVHD risk. More carefully stratified studies may be needed to confirm each specific MiHA cases; however, such randomized case studies are a big challenge in allo-HCT since the transplant cases each year are limited and each case is almost unique. On the other hand, we were able to identify multiple MiHAs encoded by genes from the Y chromosome that show statistically significant association with acute GVHD. Our limited study cohort of primarily ethnic Caucasians suggests that Y chromosome haplotypes in this population increase the risk of acute GVHD for male patients when paired with HLA matching female donors. This provides genetic evidence to support the clinical preferences of HLA-matched male donors over female donors with the same condition for male patients. For further genomics investigations with a more extensive and diverse cohort, we recommend controlling for confounding factors such as HLA-DPB1 T-cell epitope permissibility and tissue-specific gene expression.

One of the challenges in transplant genomic studies is the ultra-high dimensional genomic data from both donors and recipients. Here we developed a hybrid feature

selection model, iRBA-RF, to provide a bioinformatics data mining tool for such investigations and clinical outcome studies. The proposed iRBA-RF model successfully selected the most related SNPs that have evidence to be linked to the diseases in question, out of over 600 K and 1 M SNPs, and produced a reasonable predictive accuracy. The model may be extended to investigating multi-omics data (epigenomics, proteomics, microbiome, and metabolomics) with other clinical characteristics, as well as the multi-class prediction problems, which would be the future work to assess. Genetic factor discovery *in silico* may provide insightful interpretation of biological pathways; however, the clinical implications need further assessment and validation through immunological wet lab experiments.

In summary, this work offers evidence and guidance for further research in acute GVHD and allo-HCT and provides useful bioinformatics and data mining tools for transplant genomic studies, and can be extended to broader investigations including epigenetics in acute GVHD, pre-transplantation screening of potential donors, biomarker discovery and risk assessment of transplant outcomes.

Bibliography

Abdelhaleem, Mohamed. 2002. “The Novel Helicase Homologue DDX32 Is down-Regulated in Acute Lymphoblastic Leukemia.” *Leukemia Research* 26 (10): 945–54.

———. 2005. “RNA Helicases: Regulators of Differentiation.” *Clinical Biochemistry* 38 (6): 499–503.

Aha, David W., Dennis Kibler, and Marc K. Albert. 1991. “Instance-Based Learning Algorithms.” *Machine Learning* 6 (1): 37–66.

Albrechtsen, Anders, Ida Moltke, and Rasmus Nielsen. 2010. “Natural Selection and the Distribution of Identity-by-Descent in the Human Genome.” *Genetics* 186 (1): 295–308.

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel A. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, et al. 2013. “Signatures of Mutational Processes in Human Cancer.” *Nature* 500 (7463): 415–21.

Alli, Zaman, Michael Ho, and Mohamed Abdelhaleem. 2005. “Expression of DHX32 in Lymphoid Tissues.” *Experimental and Molecular Pathology* 79 (3): 219–23.

Altmann, André, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. “Permutation Importance: A Corrected Feature Importance Measure.” *Bioinformatics* 26 (10): 1340–47.

Amrhein, Valentin, Fränzi Korner-Nievergelt, and Tobias Roth. 2017. “The Earth Is Flat (p> 0.05): Significance Thresholds and the Crisis of Unreplicable Research.” *PeerJ* 5: e3544.

Anderson, Carl A., Fredrik H. Pettersson, Geraldine M. Clarke, Lon R. Cardon, Andrew P. Morris, and Krina T. Zondervan. 2010. “Data Quality Control in Genetic Case-Control Association Studies.” *Nature Protocols* 5 (9): 1564–73.

Antin, J. H., and J. L. Ferrara. 1992. “Cytokine Dysregulation and Acute Graft-versus-Host Disease.” *Blood* 80 (12): 2964–68.

Ashton-Chess, Joanna, Hoa Le Mai, Vojislav Jovanovic, Karine Renaudin, Yohann Foucher, Magali Giral, Anne Moreau, et al. 2010. “Immunoproteasome Beta Subunit 10 Is Increased in Chronic Antibody-Mediated Rejection.” *Kidney International* 77 (10): 880–90.

Asztalos, Bela F., Ernst J. Schaefer, Katalin V. Horvath, Shizuya Yamashita, Michael Miller, Guido Franceschini, and Laura Calabresi. 2007. “Role of LCAT in HDL Remodeling: Investigation of LCAT Deficiency States.” *Journal of Lipid Research* 48 (3): 592–99.

Bader, Peter, Hermann Kreyenberg, Walter Hoelle, Gregor Dueckers, Rupert Handgretinger, Peter Lang, Bernhard Kremens, et al. 2004. “Increasing Mixed Chimerism Is an Important Prognostic Factor for Unfavorable Outcome in Children with Acute Lymphoblastic Leukemia after Allogeneic Stem-Cell Transplantation: Possible Role for Pre-Emptive Immunotherapy?” *Journal of Clinical Oncology:*

Official Journal of the American Society of Clinical Oncology 22 (9): 1696–1705.

Bahubeshi, Amin, Marc Tischkowitz, and William D. Foulkes. 2011. "miRNA Processing and Human Cancer: DICER1 Cuts the Mustard." *Science Translational Medicine* 3 (111): 111ps46.

Bakker, Paul I. W. de, Gil McVean, Pardis C. Sabeti, Marcos M. Miretti, Todd Green, Jonathan Marchini, Xiayi Ke, et al. 2006. "A High-Resolution HLA and SNP Haplotype Map for Disease Association Studies in the Extended Human MHC." *Nature Genetics* 38 (10): 1166–72.

Ballen, Karen K., Eliane Gluckman, and Hal E. Broxmeyer. 2013. "Umbilical Cord Blood Transplantation: The First 25 Years and beyond." *Blood* 122 (4): 491–98.

Ball, L. M., R. M. Egeler, and EBMT Paediatric Working Party. 2008. "Acute GvHD: Pathogenesis and Classification." *Bone Marrow Transplantation* 41 Suppl 2 (June): S58–64.

Barker, Juliet N., and John E. Wagner. 2003. "Umbilical Cord Blood Transplantation: Current Practice and Future Innovations." *Critical Reviews in Oncology/hematology* 48 (1): 35–43.

Becker, Heiko, Kenichi Yoshida, Nadja Blagitko-Dorfs, Rainer Claus, Milena Pantic, Mahmoud Abdelkarim, Christoph Niemöller, et al. 2014. "Tracing the Development of Acute Myeloid Leukemia in CBL Syndrome." *Blood* 123 (12): 1883–86.

Bell, David A., and Hui Wang. 2000. "A Formalism for Relevance and Its Application in Feature Subset Selection." *Machine Learning* 41 (2): 175–95.

Benito, A. I., M. A. Diaz, M. González-Vicent, J. Sevilla, and L. Madero. 2004. "Hematopoietic Stem Cell Transplantation Using Umbilical Cord Blood Progenitors: Review of Current Clinical Results." *Bone Marrow Transplantation* 33 (7): 675–90.

Bergström, Tomas F., Steven J. Mack, Ulf Gyllensten, and Henry A. Erlich. 2000. "Evolution of HLA-DRB Loci, DRB1 Lineages, and Alleles: Analyses of Intron-1 and -2 Sequences." In *Major Histocompatibility Complex*, 329–46. Springer Japan.

Besse, Kelsey, Martin Maiers, Dennis Confer, and Mark Albrecht. 2016. "On Modeling Human Leukocyte Antigen–Identical Sibling Match Probability for Allogeneic Hematopoietic Cell Transplantation: Estimating the Need for an Unrelated Donor Source." *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 22 (3): 410–17.

Bolón-Canedo, V., N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera. 2014. "A Review of Microarray Datasets and Applied Feature Selection Methods." *Information Sciences* 282 (October): 111–35.

Boulesteix, Anne-Laure, Andreas Bender, Justo Lorenzo Bermejo, and Carolin Strobl. 2012. "Random Forest Gini Importance Favours SNPs with Large Minor Allele Frequency: Impact, Sources and Recommendations." *Briefings in Bioinformatics* 13 (3): 292–304.

Brede, Gaute, Jorun Solheim, and Hans Prydz. 2002. "PSKH1, a Novel Splice Factor Compartment-associated Serine Kinase." *Nucleic Acids Research* 30 (23): 5301–9.

Brede, G., J. Solheim, G. Tröen, and H. Prydz. 2000. “Characterization of PSKH1, a Novel Human Protein Serine Kinase with Centrosomal, Golgi, and Nuclear Localization.” *Genomics* 70 (1): 82–92.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. New York: Routledge.

Broen, Kelly, Henriette Levenga, Johanna Vos, Kees van Bergen, Hanny Fredrix, Annelies Greupink-Draaisma, Michel Kester, et al. 2011. “A Polymorphism in the Splice Donor Site of ZNF419 Results in the Novel Renal Cell Carcinoma-Associated Minor Histocompatibility Antigen ZAPHIR.” *PloS One* 6 (6): e21699.

Browning, Brian L., and Sharon R. Browning. 2013. “Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data.” *American Journal of Human Genetics* 93 (5): 840–51.

Budczies, Jan, Michael Bockmayr, Frederick Klauschen, Volker Endris, Stefan Fröhling, Peter Schirmacher, Carsten Denkert, and Albrecht Stenzinger. 2017. “Mutation Patterns in Genes Encoding Interferon Signaling and Antigen Presentation: A Pan-Cancer Survey with Implications for the Use of Immune Checkpoint Inhibitors.” *Genes, Chromosomes & Cancer* 56 (8): 651–59.

Burkhardt, Ute E., and Catherine J. Wu. 2013. “Boosting Leukemia-Specific T Cell Responses in Patients Following Stem Cell Transplantation.” *Oncoimmunology* 2 (11): e26587.

Cai, Ann, Derin B. Keskin, David S. DeLuca, Anselmo Alonso, Wandi Zhang, Guang Lan Zhang, Naa Norkor Hammond, et al. 2012. “Mutated BCR-ABL Generates Immunogenic T-Cell Epitopes in CML Patients.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 18 (20): 5761–72.

Callan, James P., Tom Fawcett, and Edwina L. Rissland. 1991. “CABOT: An Adaptive Approach to Case-Based Search.” In *IJCAI*, 12:803–8. pdfs.semanticscholar.org.

Cancer Genome Atlas Research Network, Timothy J. Ley, Christopher Miller, Li Ding, Benjamin J. Raphael, Andrew J. Mungall, A. Gordon Robertson, et al. 2013. “Genomic and Epigenomic Landscapes of Adult de Novo Acute Myeloid Leukemia.” *The New England Journal of Medicine* 368 (22): 2059–74.

Canzonetta, Claudia, Alexander Hoischen, Emanuela Giarin, Guiseppe Basso, Joris A. Veltman, Elisabeth Nacheva, Dean Nizetic, and Jürgen Groet. 2012. “Amplified Segment in the ‘Down Syndrome Critical Region’on HSA21 Shared between Down Syndrome and Euploid AML-M0 Excludes RUNX1, ERG and ETS2.” *British Journal of Haematology* 157 (2): 197–200.

Cereb, N., A. L. Hughes, and S. Y. Yang. 1997. “Locus-Specific Conservation of the HLA Class I Introns by Intra-Locus Homogenization.” *Immunogenetics* 47 (1): 30–36.

Cheng, Dan, Huimin Shi, Kan Zhang, Lingling Yi, and Guohua Zhen. 2014. “RAD 51 Gene 135G/C Polymorphism and the Risk of Four Types of Common Cancers: A Meta-Analysis.” *Diagnostic Pathology* 9 (1): 18.

Chi, Chen, Leigh C. Murphy, and Pingzhao Hu. 2018. “Recurrent Copy Number Alterations in Young Women with Breast Cancer.” *Oncotarget* 9 (14): 11541–58.

Chojecki, Aleksander. 2017. “Does MHC Class I Chain-Related Gene A Matter?” *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 23 (3): 365–66.

Christiansen, Ole Bjarne, Rudi Steffensen, and Henriette Svarre Nielsen. 2011. “Anti-HY Responses in Pregnancy Disorders.” *American Journal of Reproductive Immunology* 66 Suppl 1 (July): 93–100.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain w1118; Iso-2; Iso-3.” *Fly* 6 (2): 80–92.

Ciurea, Stefan O., Rima M. Saliba, Gabriela Rondon, Poliana A. Patah, Fleur Aung, Pedro Cano, Borje S. Andersson, et al. 2011. “Outcomes of Patients with Myeloid Malignancies Treated with Allogeneic Hematopoietic Stem Cell Transplantation from Matched Unrelated Donors Compared with One Human Leukocyte Antigen Mismatched Related Donors Using HLA Typing at 10 Loci.” *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 17 (6): 923–29.

Cleary, John G., Ross Braithwaite, Kurt Gaastra, Brian S. Hilbush, Stuart Inglis, Sean A. Irvine, Alan Jackson, et al. 2014. “Joint Variant and de Novo Mutation Identification on Pedigrees from High-Throughput Sequencing Data.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 21 (6): 405–19.

Copelan, Edward A. 2006. “Hematopoietic Stem-Cell Transplantation.” *The New England Journal of Medicine* 354 (17): 1813–26.

Cordell, Heather J. 2009. “Detecting Gene–gene Interactions That Underlie Human Diseases.” *Nature Reviews. Genetics* 10 (6): 392–404.

Couriel, Daniel, Humberto Caldera, Richard Champlin, and Krishna Komanduri. 2004. “Acute Graft-versus-Host Disease: Pathophysiology, Clinical Manifestations, and Management.” *Cancer* 101 (9): 1936–46.

Cretney, Erika, Annie Xin, Wei Shi, Martina Minnich, Frederick Masson, Maria Miasari, Gabrielle T. Belz, et al. 2011. “The Transcription Factors Blimp-1 and IRF4 Jointly Control the Differentiation and Function of Effector Regulatory T Cells.” *Nature Immunology* 12 (4): 304–11.

Crivello, Pietro, Laura Zito, Federico Sizzano, Elisabetta Zino, Martin Maiers, Arend Mulder, Cristina Toffalori, et al. 2015. “The Impact of Amino Acid Variability on Alloreactivity Defines a Functional Distance Predictive of Permissive HLA-DPB1 Mismatches in Hematopoietic Stem Cell Transplantation.” *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 21 (2): 233–41.

Crocchiolo, Roberto, Elisabetta Zino, Luca Vago, Rosi Oneto, Barbara Bruno, Simona Pollichieni, Nicoletta Sacchi, et al. 2009. “Nonpermissive HLA-DPB1 Disparity Is a Significant Independent Risk Factor for Mortality after Unrelated Hematopoietic Stem Cell Transplantation.” *Blood* 114 (7): 1437–44.

Csizmar, C. M., D. H. Kim, and Z. Sachs. 2016. “The Role of the Proteasome in AML.” *Blood Cancer Journal* 6 (12): e503.

Curtis, R. E., P. A. Rowlings, H. J. Deeg, D. A. Shriner, G. Socíe, L. B. Travis, M. M. Horowitz, et al. 1997. “Solid Cancers after Bone Marrow Transplantation.” *The New England Journal of Medicine* 336 (13): 897–904.

Cutler, C., and J. H. Antin. 2001. “Peripheral Blood Stem Cells for Allogeneic Transplantation: A Review.” *Stem Cells* 19 (2): 108–17.

Damm, Frederik, Kathrin Lange, Michael Heuser, Tina Oberacker, Michael Morgan, Katharina Wagner, Jürgen Krauter, Brigitte Schlegelberger, Arnold Ganser, and Gudren Göhring. 2010. “Phosphoinositide Phospholipase Cβ1 (PI-PLC β 1) Gene in Myelodysplastic Syndromes and Cytogenetically Normal Acute Myeloid Leukemia: Not a Deletion, but Increased PI-PLC β 1 Expression Is an Independent Prognostic Factor.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 28 (22): e384–87.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58.

Deeg, H. Joachim, and Joseph H. Antin. 2006. “The Clinical Spectrum of Acute Graft-versus-Host Disease.” *Seminars in Hematology* 43 (1): 24–31.

Dehn, Jason, Michelle Setterholm, Kelly Buck, Jane Kempenich, Beth Beduhn, Loren Gragert, Abeer Madbouly, Stephanie Fingerson, and Martin Maiers. 2016. “HapLogic: A Predictive Human Leukocyte Antigen–Matching Algorithm to Enhance Rapid Identification of the Optimal Unrelated Hematopoietic Stem Cell Sources for Transplantation.” *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 22 (11): 2038–46.

De Palma, Michele, Mary Anna Venneri, Rossella Galli, Lucia Sergi Sergi, Letterio S. Politi, Maurilio Sampaolesi, and Luigi Naldini. 2005. “Tie2 Identifies a Hematopoietic Lineage of Proangiogenic Monocytes Required for Tumor Vessel Formation and a Mesenchymal Population of Pericyte Progenitors.” *Cancer Cell* 8 (3): 211–26.

Ding, Chris, and Hanchuan Peng. 2005. “Minimum Redundancy Feature Selection from Microarray Gene Expression Data.” *Journal of Bioinformatics and Computational Biology* 3 (2): 185–205.

Domingos, Pedro. 1998. “Occam’s Two Razors: The Sharp and the Blunt.” In *KDD*, 37–43. aaai.org.

———. 1999. “The Role of Occam’s Razor in Knowledge Discovery.” *Data Mining and Knowledge Discovery* 3 (4): 409–25.

D’Souza, A., and X. Zhu. 2017. “Current Uses and Outcomes of Hematopoietic Cell

Transplantation (HCT): CIBMTR Summary Slides, 2017." http://www.cibmtr.org.
Eapen, Mary, Pablo Rubinstein, Mei-Jie Zhang, Cladd Stevens, Joanne Kurtzberg, Andromachi Scaradavou, Fausto R. Loberiza, et al. 2007. "Outcomes of Transplantation of Unrelated Donor Umbilical Cord Blood and Bone Marrow in Children with Acute Leukaemia: A Comparison Study." *The Lancet* 369 (9577): 1947–54.
Edgar, Robert C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research* 32 (5): 1792–97.
Eom, Jae-Hong, and Byoung-Tak Zhang. 2004. "PubMiner: Machine Learning-Based Text Mining for Biomedical Information Analysis." *Genomics & Informatics* 2 (2): 99–106.
Eppstein, M. J., and P. Haake. 2008. "Very Large Scale ReliefF for Genome-Wide Association Analysis." In *2008 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 112–19. ieeexplore.ieee.org.
European Bioinformatics Institute. 2018. "IPD-IMGT/HLA Database." January 2018. https://www.ebi.ac.uk/ipd/imgt/hla/.
Falchook, Gerald S., Rom Leidner, Elizabeth Stankevich, Brian Piening, Carlo Bifulco, Israel Lowy, and Matthew G. Fury. 2016. "Responses of Metastatic Basal Cell and Cutaneous Squamous Cell Carcinomas to Anti-PD1 Monoclonal Antibody REGN2810." *Journal for Immunotherapy of Cancer* 4 (November): 70.
Felipe, Antonio, Rubén Vicente, Núria Villalonga, Meritxell Roura-Ferrer, Ramón Martínez-Mármol, Laura Solé, Joan C. Ferreres, and Enric Condom. 2006. "Potassium Channels: New Targets in Cancer Therapy." *Cancer Detection and Prevention* 30 (4): 375–85.
Ferrara, James L. M., John E. Levine, Pavan Reddy, and Ernst Holler. 2009. "Graft-versus-Host Disease." *The Lancet* 373 (9674): 1550–61.
Fiume, Roberta, Xu Huang, Giulia Ramazzotti, Sara Mongiorgi, Patrizia Santi, Tim Somervaille, and Nullin Divecha. 2014. "Phospholipase c Beta 1 (PLCb1) in Acute Myeloid Leukemia (AML): A Novel Potential Therapeutic Target." *Italian Journal of Anatomy and Embryology = Archivio Italiano Di Anatomia Ed Embriologia* 119 (1): 88.
Fleischhauer, Katharina, Bronwen E. Shaw, Theodore Gooley, Mari Malkki, Peter Bardy, Jean-Denis Bignon, Valérie Dubois, et al. 2012. "Effect of T-Cell-Epitope Matching at HLA-DPB1 in Recipients of Unrelated-Donor Haemopoietic-Cell Transplantation: A Retrospective Study." *The Lancet Oncology* 13 (4): 366–74.
Fleischhauer, K., M. A. Fernandez-Viña, T. Wang, M. Haagenson, M. Battiwalla, L. A. Baxter-Lowe, F. Ciceri, et al. 2014. "Risk Associations between HLA-DPB1 T-Cell Epitope Matching and Outcome of Unrelated Hematopoietic Cell Transplantation Are Independent of HLA-DPA1." *Bone Marrow Transplantation* 49 (9): 1176–83.
Freitas, Alex A. 2019. "Investigating the Role of Simpson's Paradox in the Analysis of Top-Ranked Features in High-Dimensional Bioinformatics Datasets." *Briefings in*

Bioinformatics, January. https://doi.org/10.1093/bib/bby126.

Friedman, Jerome H. 1997. “On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality.” *Data Mining and Knowledge Discovery* 1 (1): 55–77.

Fuerst, Daniel, Christine Neuchel, Dietger Niederwieser, Donald Bunjes, Martin Gramatzki, Eva Wagner, Gerald Wulf, et al. 2016. “Matching for the MICA-129 Polymorphism Is Beneficial in Unrelated Hematopoietic Stem Cell Transplantation.” *Blood* 128 (26): 3169–76.

Fürst, Daniel, Carlheinz Müller, Vladan Vucinic, Donald Bunjes, Wolfgang Herr, Martin Gramatzki, Rainer Schwerdtfeger, et al. 2013. “High-Resolution HLA Matching in Hematopoietic Stem Cell Transplantation: A Retrospective Collaborative Analysis.” *Blood* 122 (18): 3220–29.

Fuster, Oscar, Marta Llop, Sandra Dolz, Paloma García, Esperanza Such, Mariam Ibáñez, Irene Luna, et al. 2013. “Adverse Prognostic Value of MYBL2 Overexpression and Association with microRNA-30 Family in Acute Myeloid Leukemia Patients.” *Leukemia Research* 37 (12): 1690–96.

Gao, Liquan, Ilaria Bellantuono, Annika Elsässer, Stephen B. Marley, Myrtle Y. Gordon, John M. Goldman, and Hans J. Stauss. 2000. “Selective Elimination of Leukemic CD34+ Progenitor Cells by Cytotoxic T Lymphocytes Specific for WT1.” *Blood* 95 (7): 2198–2203.

Garofalo, Andrea, Lynette Sholl, Brendan Reardon, Amaro Taylor-Weiner, Ali Amin-Mansour, Diana Miao, David Liu, et al. 2016. “The Impact of Tumor Profiling Approaches and Genomic Data Strategies for Cancer Precision Medicine.” *Genome Medicine* 8 (1): 79.

Gilliland, D. Gary, and James D. Griffin. 2002. “The Roles of FLT3 in Hematopoiesis and Leukemia.” *Blood* 100 (5): 1532–42.

Goldman, J. M., K. H. Th’Ng, D. S. Park, A. S. D. Spiers, R. M. Lowenthal, and T. Ruutu. 1978. “Collection, Cryopreservation and Subsequent Viability of Haemopoietic Stem Cells Intended for Treatment of Chronic Granulocytic Leukaemia in Blast-Cell Transformation.” *British Journal of Haematology* 40 (2): 185–95.

Gooley, Ted A., Jason W. Chien, Steven A. Pergam, Sangeeta Hingorani, Mohamed L. Sorror, Michael Boeckh, Paul J. Martin, et al. 2010. “Reduced Mortality after Allogeneic Hematopoietic-Cell Transplantation.” *The New England Journal of Medicine* 363 (22): 2091–2101.

Goulder, Philip J. R., and Bruce D. Walker. 2012. “HIV and HLA Class I: An Evolving Relationship.” *Immunity* 37 (3): 426–40.

Gourraud, Pierre-Antoine, Pouya Khankhanian, Nezih Cereb, Soo Young Yang, Michael Feolo, Martin Maiers, John D. Rioux, Stephen Hauser, and Jorge Oksenberg. 2014. “HLA Diversity in the 1000 Genomes Dataset.” *PloS One* 9 (7): e97282.

Goyal, R. K., S. J. Lee, T. Wang, M. Trucco, M. Haagenson, S. R. Spellman, M. Verneris, and R. E. Ferrell. 2017. “Novel HLA-DP Region Susceptibility Loci Associated with Severe Acute GvHD.” *Bone Marrow Transplantation* 52 (1): 95–

100.
Granizo-Mackenzie, Delaney, and Jason H. Moore. 2013. “Multiple Threshold Spatially Uniform ReliefF for the Genetic Analysis of Complex Human Diseases.” In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, 1–10. Springer Berlin Heidelberg.
Gratwohl, A. 2012. “The EBMT Risk Score.” *Bone Marrow Transplantation* 47 (6): 749–56.
Gratwohl, Alois, Martin Stern, Ronald Brand, Jane Apperley, Helen Baldomero, Theo de Witte, Giorgio Dini, et al. 2009. “Risk Score for Outcome after Allogeneic Hematopoietic Stem Cell Transplantation: A Retrospective Analysis.” *Cancer* 115 (20): 4715–26.
Greene, Casey S., Daniel S. Himmelstein, Jeff Kiralis, and Jason H. Moore. 2010. “The Informative Extremes: Using Both Nearest and Farthest Individuals Can Improve Relief Algorithms in the Domain of Human Genetics.” In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, 182–93. Springer Berlin Heidelberg.
Greene, Casey S., Nadia M. Penrod, Jeff Kiralis, and Jason H. Moore. 2009. “Spatially Uniform Relieff (SURF) for Computationally-Efficient Filtering of Gene-Gene Interactions.” *BioData Mining* 2 (1): 5.
Griffioen, Marieke, Cornelis A. M. van Bergen, and J. H. Frederik Falkenburg. 2016. “Autosomal Minor Histocompatibility Antigens: How Genetic Variants Create Diversity in Immune Targets.” *Frontiers in Immunology* 7 (March): 100.
Griffioen, Marieke, M. Willy Honders, Edith D. van der Meijden, Simone A. P. van Luxemburg-Heijs, Ellie G. A. Lurvink, Michel G. D. Kester, Cornelis A. M. van Bergen, and J. H. Frederik Falkenburg. 2012. “Identification of 4 Novel HLA-B*40:01 Restricted Minor Histocompatibility Antigens and Their Potential as Targets for Graft-versus-Leukemia Reactivity.” *Haematologica* 97 (8): 1196–1204.
GTEx Consortium. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature Genetics* 45 (6): 580–85.
Halenius, Anne, Carolin Gerke, and Hartmut Hengel. 2015. “Classical and Non-Classical MHC I Molecule Manipulation by Human Cytomegalovirus: So Many Targets—but How Many Arrows in the Quiver?” *Cellular & Molecular Immunology* 12 (2): 139–53.
Hansen, John A., Jason W. Chien, Edus H. Warren, Lue Ping Zhao, and Paul J. Martin. 2010. “Defining Genetic Risk for Graft-versus-Host Disease and Mortality Following Allogeneic Hematopoietic Stem Cell Transplantation.” *Current Opinion in Hematology* 17 (6): 483–92.
Holland, Jane D., Alexandra Klaus, Alistair N. Garratt, and Walter Birchmeier. 2013. “Wnt Signaling in Stem and Cancer Stem Cells.” *Current Opinion in Cell Biology* 25 (2): 254–64.
Horowitz, M. M., R. P. Gale, P. M. Sondel, J. M. Goldman, J. Kersey, H. J. Kolb, A. A. Rimm, O. Ringdén, C. Rozman, and B. Speck. 1990. “Graft-versus-Leukemia

Reactions after Bone Marrow Transplantation." *Blood* 75 (3): 555–62.
Horton, Roger, Laurens Wilming, Vikki Rand, Ruth C. Lovering, Elspeth A. Bruford, Varsha K. Khodiyar, Michael J. Lush, et al. 2004. "Gene Map of the Extended Human MHC." *Nature Reviews. Genetics* 5 (12): 889–99.
Hou, Lihua, Cynthia Vierra-Green, Ana Lazaro, Colleen Brady, Michael Haagenson, Stephen Spellman, and Carolyn Katovich Hurley. 2017. "Limited HLA Sequence Variation outside of Antigen Recognition Domain Exons of 360 10 of 10 Matched Unrelated Hematopoietic Stem Cell Transplant Donor-Recipient Pairs." *Hladnikia* 89 (1): 39–46.
Huang, Alexander C., Michael A. Postow, Robert J. Orlowski, Rosemarie Mick, Bertram Bengsch, Sasikanth Manne, Wei Xu, et al. 2017. "T-Cell Invigoration to Tumour Burden Ratio Associated with Anti-PD-1 Response." *Nature* 545 (7652): 60–65.
Huang, Hu, Cynthia A. Vierra-Green, Colleen Brady, Jayesh Iyer, Caleb J. Kennedy, and Stephen R. Spellman. 2018. "P071 Limited Exon Sequence Mismatches Outside the Antigen Recognition Domain in a Cohort of 4646 High Resolution 10/10 HLA-Matched Donor and Recipient." *Human Immunology* 79 (October): 114.
Huang, Hu, Wei Wang, Yung-Tsi Bolon, Craig Malmberg, Caleb Kennedy, and Martin Maiers. 2016. "P065 Information Theory-Based Analysis of Classical HLA Genes." *Human Immunology* 77 (Supplement): 85.
Huber, Magdalena, Anne Brüstle, Katharina Reinhard, Anna Guralnik, Gina Walter, Azita Mahiny, Eberhard von Löw, and Michael Lohoff. 2008. "IRF4 Is Essential for IL-21-Mediated Induction, Amplification, and Stabilization of the Th17 Phenotype." *Proceedings of the National Academy of Sciences of the United States of America* 105 (52): 20846–51.
Hwang, William Ying Khee, Miny Samuel, Daryl Tan, Liang Piu Koh, Winston Lim, and Yeh Ching Linn. 2007. "A Meta-Analysis of Unrelated Donor Umbilical Cord Blood Transplantation versus Unrelated Donor Bone Marrow Transplantation in Adult and Pediatric Patients." *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 13 (4): 444–53.
Ibáñez, Mariam, José Carbonell-Caballero, Luz García-Alonso, Esperanza Such, Jorge Jiménez-Almazán, Enrique Vidal, Eva Barragán, et al. 2016. "The Mutational Landscape of Acute Promyelocytic Leukemia Reveals an Interacting Network of Co-Occurrences and Recurrent Mutations." *PloS One* 11 (2): e0148346.
Iborra, Marisa, Inés Moret, Francisco Rausell, Enrique Busó, Elena Cerrillo, Esteban Sáez-González, Pilar Nos, and Belén Beltrán. 2018. "Different Genetic Expression Profiles of Oxidative Stress and Apoptosis-Related Genes in Crohn's Disease." *Digestion*, October, 1–10.
Inamoto, Y., N. N. Shah, B. N. Savani, B. E. Shaw, A. A. Abraham, I. A. Ahmed, G. Akpek, et al. 2015. "Secondary Solid Cancer Screening Following Hematopoietic Cell Transplantation." *Bone Marrow Transplantation* 50 (8): 1013–23.
Ishwaran, Hemant. 2007. "Variable Importance in Binary Regression Trees and Forests."

Electronic Journal of Statistics 1: 519–37.

Ishwaran, Hemant, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. 2008. "Random Survival Forests." *The Annals of Applied Statistics* 2 (3): 841–60.

Ishwaran, Hemant, and Min Lu. 2018. "Standard Errors and Confidence Intervals for Variable Importance in Random Forest Regression, Classification, and Survival." *Statistics in Medicine*, June. https://doi.org/10.1002/sim.7803.

Iwase, Hayato, Takaaki Kobayashi, Yasuhiro Kodera, Yuko Miwa, Takafumi Kuzuya, Kenta Iwasaki, Masataka Haneda, et al. 2011. "Clinical Significance of Regulatory T-Cell–Related Gene Expression in Peripheral Blood After Renal Transplantation." *Transplantation* 91 (2): 191.

Jacobsohn, David A., and Georgia B. Vogelsang. 2007. "Acute Graft versus Host Disease." *Orphanet Journal of Rare Diseases* 2 (September): 35.

Jagasia, Madan, Mukta Arora, Mary E. D. Flowers, Nelson J. Chao, Philip L. McCarthy, Corey S. Cutler, Alvaro Urbano-Ispizua, et al. 2012. "Risk Factors for Acute GVHD and Survival after Hematopoietic Cell Transplantation." *Blood* 119 (1): 296–307.

Jagasia, Madan H., Hildegard T. Greinix, Mukta Arora, Kirsten M. Williams, Daniel Wolff, Edward W. Cowen, Jeanne Palmer, et al. 2015. "National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: I. The 2014 Diagnosis and Staging Working Group Report." *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 21 (3): 389–401.e1.

Jameson-Lee, Max, Vishal Koparde, Phil Griffith, Allison F. Scalora, Juliana K. Sampson, Haniya Khalid, Nihar U. Sheth, et al. 2014. "In Silico Derivation of HLA-Specific Alloreactivity Potential from Whole Exome Sequencing of Stem-Cell Transplant Donors and Recipients: Understanding the Quantitative Immunobiology of Allogeneic Transplantation." *Frontiers in Immunology* 5: 529.

Janeway, Charles A., Jr, Paul Travers, Mark Walport, and Mark J. Shlomchik. 2001. "The Major Histocompatibility Complex and Its Functions." In *Immunobiology: The Immune System in Health and Disease. 5th Edition*. Garland Science.

Janitza, Silke, Ender Celik, and Anne-Laure Boulesteix. 2016. "A Computationally Fast Variable Importance Test for Random Forests for High-Dimensional Data." *Advances in Data Analysis and Classification*, November. https://doi.org/10.1007/s11634-016-0276-4.

Jones, Siân, Valsamo Anagnostou, Karli Lytle, Sonya Parpart-Li, Monica Nesselbush, David R. Riley, Manish Shukla, et al. 2015. "Personalized Genomic Analyses for Cancer Mutation Discovery and Interpretation." *Science Translational Medicine* 7 (283): 283ra53.

Kang, Xing, Rong Chen, Jie Zhang, Gang Li, Peng-Gao Dai, Chao Chen, and Hui-Juan Wang. 2015. "Expression Profile Analysis of Zinc Transporters (ZIP4, ZIP9, ZIP11, ZnT9) in Gliomas and Their Correlation with IDH1 Mutation Status." *Asian Pacific Journal of Cancer Prevention: APJCP* 16 (8): 3355–60.

Kato, Gregory J., Frédéric B. Piel, Clarice D. Reid, Marilyn H. Gaston, Kwaku Ohene-Frempong, Lakshmanan Krishnamurti, Wally R. Smith, et al. 2018. “Sickle Cell Disease.” *Nature Reviews. Disease Primers* 4 (March): 18010.

Keen, Judy C., and Helen M. Moore. 2015. “The Genotype-Tissue Expression (GTEx) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine.” *Journal of Personalized Medicine* 5 (1): 22–29.

Kekre, Natasha, Kimberley S. Mak, Konrad H. Stopsack, Moritz Binder, Kazusa Ishii, Elsa Brånvall, and Corey S. Cutler. 2016. “Impact of HLA-Mismatch in Unrelated Donor Hematopoietic Stem Cell Transplantation: A Meta-Analysis.” *American Journal of Hematology* 91 (6): 551–55.

Kent, W. James. 2002. “BLAT—The BLAST-Like Alignment Tool.” *Genome Research* 12 (4): 656–64.

Keşmir, Can, Alexander K. Nussbaum, Hansjörg Schild, Vincent Detours, and Søren Brunak. 2002. “Prediction of Proteasome Cleavage Motifs by Neural Networks.” *Protein Engineering* 15 (4): 287–96.

Kim, S. Joseph, and John S. Gill. 2009. “H-Y Incompatibility Predicts Short-Term Outcomes for Kidney Transplant Recipients.” *Journal of the American Society of Nephrology: JASN* 20 (9): 2025–33.

Kimura, Hiroaki, Patrizio Caturegli, Masafumi Takahashi, and Koichi Suzuki. 2015. “New Insights into the Function of the Immunoproteasome in Immune and Nonimmune Cells.” *Journal of Immunology Research* 2015 (October): 541984.

Kollman, Craig, Stephen R. Spellman, Mei-Jie Zhang, Anna Hassebroek, Claudio Anasetti, Joseph H. Antin, Richard E. Champlin, et al. 2016. “The Effect of Donor Characteristics on Survival after Unrelated Donor Transplantation for Hematologic Malignancy.” *Blood* 127 (2): 260–67.

Kononenko, Igor. 1994. “Estimating Attributes: Analysis and Extensions of RELIEF.” In *Machine Learning: ECML-94*, 171–82. Springer Berlin Heidelberg.

Koo, Ching Lee, Mei Jing Liew, Mohd Saberi Mohamad, and Abdul Hakim Mohamed Salleh. 2013. “A Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology.” *BioMed Research International* 2013 (October): 432375.

Kotsch, K., and R. Blasczyk. 2000. “The Noncoding Regions of HLA-DRB Uncover Interlineage Recombinations as a Mechanism of HLA Diversification.” *Journal of Immunology* 165 (10): 5664–70.

Kryuchkova-Mostacci, Nadezda, and Marc Robinson-Rechavi. 2017. “A Benchmark of Gene Expression Tissue-Specificity Metrics.” *Briefings in Bioinformatics* 18 (2): 205–14.

Lahn, Bruce T., Nathaniel M. Pearson, and Karin Jegalian. 2001. “The Human Y Chromosome, in the Light of Evolution.” *Nature Reviews. Genetics* 2 (3): 207.

Larsen, F., J. Solheim, T. Kristensen, A. B. Kolstø, and H. Prydz. 1993. “A Tight Cluster of Five Unrelated Human Genes on Chromosome 16q22.1.” *Human Molecular*

Genetics 2 (10): 1589–95.

Laurell, Henrik, Michele Bouisson, Philippe Berthelemy, Philippe Rochaix, Sebastien Dejean, Philippe Besse, Christiane Susini, Lucien Pradayrol, Nicole Vaysse, and Louis Buscail. 2006. "Identification of Biomarkers of Human Pancreatic Adenocarcinomas by Expression Profiling and Validation with Gene Expression Analysis in Endoscopic Ultrasound-Guided Fine Needle Aspiration Samples." *World Journal of Gastroenterology: WJG* 12 (21): 3344–51.

Lazarian, Gregory, Romain Guièze, and Catherine J. Wu. 2017. "Clinical Implications of Novel Genomic Discoveries in Chronic Lymphocytic Leukemia." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 35 (9): 984–93.

Lee, Jae-Bong, Chae-Kyoung Yoo, Hee-Bok Park, In-Cheol Cho, and Hyun-Tae Lim. 2016. "Association of the Single Nucleotide Polymorphisms in RUNX1, DYRK1A, and KCNJ15 with Blood Related Traits in Pigs." *Asian-Australasian Journal of Animal Sciences* 29 (12): 1675–81.

Lee, Jong Woo, Eun Goo Jeong, Sung Hak Lee, Suk Woo Nam, Sang Ho Kim, Jung Young Lee, Nam Jin Yoo, and Sug Hyung Lee. 2007. "Mutational Analysis of PTPRT Phosphatase Domains in Common Human Cancers." *APMIS: Acta Pathologica, Microbiologica, et Immunologica Scandinavica* 115 (1): 47–51.

Lee, Stephanie J., John Klein, Michael Haagenson, Lee Ann Baxter-Lowe, Dennis L. Confer, Mary Eapen, Marcelo Fernandez-Vina, et al. 2007. "High-Resolution Donor-Recipient HLA Matching Contributes to the Success of Unrelated Donor Marrow Transplantation." *Blood* 110 (13): 4576–83.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Lindsley, R. Coleman, Wael Saber, Brenton G. Mar, Robert Redd, Tao Wang, Michael D. Haagenson, Peter V. Grauman, et al. 2017. "Prognostic Mutations in Myelodysplastic Syndrome after Stem-Cell Transplantation." *The New England Journal of Medicine* 376 (6): 536–47.

Lin, Ming-Tseh, Barry Storer, Paul J. Martin, Li-Hui Tseng, Bryan Grogan, Pei-Jer Chen, Lue P. Zhao, and John A. Hansen. 2005a. "Genetic Variation in the IL-10 Pathway Modulates Severity of Acute Graft-versus-Host Disease Following Hematopoietic Cell Transplantation: Synergism between IL-10 Genotype of Patient and IL-10 Receptor β Genotype of Donor." *Blood* 106 (12): 3995–4001.

———. 2005b. "Genetic Variation in the IL-10 Pathway Modulates Severity of Acute Graft-versus-Host Disease Following Hematopoietic Cell Transplantation: Synergism between IL-10 Genotype of Patient and IL-10 Receptor Beta Genotype of Donor." *Blood* 106 (12): 3995–4001.

Littera, Roberto, Nicola Orrù, Giovanni Caocci, Marco Sanna, Marina Mulargia, Eugenia

Piras, Adriana Vacca, et al. 2012. “Interactions between Killer Immunoglobulin-like Receptors and Their Human Leucocyte Antigen Class I Ligands Influence the Outcome of Unrelated Haematopoietic Stem Cell Transplantation for Thalassaemia: A Novel Predictive Algorithm.” *British Journal of Haematology* 156 (1): 118–28.

Loren, Alison W., Greta R. Bunin, Christian Boudreau, Richard E. Champlin, Avital Cnaan, Mary M. Horowitz, Fausto R. Loberiza, and David L. Porter. 2006. “Impact of Donor and Recipient Sex and Parity on Outcomes of HLA-Identical Sibling Allogeneic Hematopoietic Stem Cell Transplantation.” *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 12 (7): 758–69.

Lou, Xiao, Chuanhua Zhao, and Hu Chen. 2018. “Unrelated Donor Umbilical Cord Blood Transplant versus Unrelated Hematopoietic Stem Cell Transplant in Patients with Acute Leukemia: A Meta-Analysis and Systematic Review.” *Blood Reviews* 32 (3): 192–202.

Lui, Tony T. H., Celine Lacroix, Syed M. Ahmed, Seth J. Goldenberg, Craig A. Leach, Avais M. Daulat, and Stephane Angers. 2011. “The Ubiquitin-Specific Protease USP34 Regulates Axin Stability and Wnt/β-Catenin Signaling.” *Molecular and Cellular Biology* 31 (10): 2053–65.

Lu, Min, and Hemant Ishwaran. 2018. “A Prediction-Based Alternative to P Values in Regression Models.” *The Journal of Thoracic and Cardiovascular Surgery* 155 (3): 1130–36.e4.

Lu, Yong-Chen, and Paul F. Robbins. 2016. “Targeting Neoantigens for Cancer Immunotherapy.” *International Immunology* 28 (7): 365–70.

Mack, Steven J., Pedro Cano, Jill A. Hollenbach, Jun He, Carolyn Katovich Hurley, Derek Middleton, Maria Elisa Moraes, et al. 2013. “Common and Well-Documented HLA Alleles: 2012 Update to the CWD Catalogue.” *Tissue Antigens* 81 (4): 194–203.

Madbouly, Abeer, Tao Wang, Michael Haagenson, Vanja Paunic, Cynthia Vierra-Green, Katharina Fleischhauer, Katharine C. Hsu, et al. 2017. “Investigating the Association of Genetic Admixture and Donor/Recipient Genetic Disparity with Transplant Outcomes.” *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 23 (6): 1029–37.

Marsh, S. G. E., and WHO Nomenclature Committee for Factors of the HLA System. 2018. “Nomenclature for Factors of the HLA System, Update January 2018.” *Hladnikia* 91 (6): 549–55.

Martin, Paul J., David M. Levine, Barry E. Storer, Edus H. Warren, Xiuwen Zheng, Sarah C. Nelson, Anajane G. Smith, Bo K. Mortensen, and John A. Hansen. 2017. “Genome-Wide Minor Histocompatibility Matching as Related to the Risk of Graft-versus-Host Disease.” *Blood* 129 (6): 791–98.

McGovern, Dermot P. B., Kent D. Taylor, Carol Landers, Carrie Derkowski, Deb Dutridge, Marla Dubinsky, Andy Ippoliti, et al. 2009. “MAGI2 Genetic Variation and

Inflammatory Bowel Disease." *Inflammatory Bowel Diseases* 15 (1): 75–83.
McKinney, Brett A., David M. Reif, Marylyn D. Ritchie, and Jason H. Moore. 2006. "Machine Learning for Detecting Gene-Gene Interactions: A Review." *Applied Bioinformatics* 5 (2): 77–88.
Mehta, J., S. Singhal, A. P. Gee, K-Y Chiang, K. Godder, F. van Rhee Fv, S. DeRienzo, W. O'Neal, L. Lamb, and P. J. Henslee-Downey. 2004. "Bone Marrow Transplantation from Partially HLA-Mismatched Family Donors for Acute Leukemia: Single-Center Experience of 201 Patients." *Bone Marrow Transplantation* 33 (4): 389–96.
Messmann, Joanna J., Tanja Reisser, Frank Leithäuser, Manfred B. Lutz, Klaus-Michael Debatin, and Gudrun Strauss. 2015. "In Vitro-Generated MDSCs Prevent Murine GVHD by Inducing Type 2 T Cells without Disabling Anti-Tumor Cytotoxicity." *Blood*, January, blood – 2015–01 – 624163.
Meyer, Diogo, and Rainer Blasczyk. 2000. "The Effect of Mutation, Recombination and Selection on HLA Non-Coding Sequences." In *Major Histocompatibility Complex*, 398–411. Springer Japan.
Miao, Lei, X. F. Qian, G. H. Yang, and L. D. Zhao. 2015. "Relationship between RAD51-G135C and XRCC3-C241T Single Nucleotide Polymorphisms and Onset of Acute Myeloid Leukemia." *Zhongguo Shi Yan Xue Ye Xue Za Zhi / Zhongguo Bing Li Sheng Li Xue Hui = Journal of Experimental Hematology / Chinese Association of Pathophysiology* 23 (3): 605–11.
Miklos, David B., Haesook T. Kim, Katherine H. Miller, Luxuan Guo, Emmanuel Zorn, Stephanie J. Lee, Ephraim P. Hochberg, et al. 2005. "Antibody Responses to H-Y Minor Histocompatibility Antigens Correlate with Chronic Graft-versus-Host Disease and Disease Remission." *Blood* 105 (7): 2973–78.
Miller, Jeffrey S., Edus H. Warren, Marcel R. M. van den Brink, Jerome Ritz, Warren D. Shlomchik, William J. Murphy, A. John Barrett, et al. 2010. "NCI First International Workshop on the Biology, Prevention, and Treatment of Relapse after Allogeneic Hematopoietic Stem Cell Transplantation: Report from the Committee on the Biology Underlying Recurrence of Malignant Disease Following Allogeneic HSCT: Graft-versus-Tumor/leukemia Reaction." *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 16 (5): 565–86.
Molldrem, J., S. Dermime, K. Parker, Y. Z. Jiang, D. Mavroudis, N. Hensel, P. Fukushima, and A. J. Barrett. 1996. "Targeted T-Cell Therapy for Human Leukemia: Cytotoxic T Lymphocytes Specific for a Peptide Derived from Proteinase 3 Preferentially Lyse Human Myeloid Leukemia Cells." *Blood* 88 (7): 2450–57.
Molldrem, J. J., P. P. Lee, C. Wang, R. E. Champlin, and M. M. Davis. 1999. "A PR1-Human Leukocyte Antigen-A2 Tetramer Can Be Used to Isolate Low-Frequency Cytotoxic T Lymphocytes from Healthy Donors That Selectively Lyse Chronic Myelogenous Leukemia." *Cancer Research* 59 (11): 2675–81.

Moore, Jason H., and Marylyn D. Ritchie. 2004. “STUDENTJAMA. The Challenges of Whole-Genome Approaches to Common Diseases.” *JAMA: The Journal of the American Medical Association* 291 (13): 1642–43.

Moore, Jason H., and Bill C. White. 2007. “Tuning ReliefF for Genome-Wide Genetic Analysis.” In *Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics*, 166–75. Springer Berlin Heidelberg.

Morishima, Yasuo, Koichi Kashiwase, Keitaro Matsuo, Fumihiro Azuma, Satoko Morishima, Makoto Onizuka, Toshio Yabe, et al. 2015. “Biological Significance of HLA Locus Matching in Unrelated Donor Bone Marrow Transplantation.” *Blood* 125 (7): 1189–97.

Mullally, Ann, and Jerome Ritz. 2007. “Beyond HLA: The Significance of Genomic Variation for Allogeneic Hematopoietic Stem Cell Transplantation.” *Blood* 109 (4): 1355–62.

Müller-Tidow, Carsten, Björn Steffen, Thomas Cauvet, Lara Tickenbrock, Ping Ji, Sven Diederichs, Bülent Sargin, et al. 2004. “Translocation Products in Acute Myeloid Leukemia Activate the Wnt Signaling Pathway in Hematopoietic Cells.” *Molecular and Cellular Biology* 24 (7): 2890–2904.

Musa, Julian, Marie-Ming Aynaud, Olivier Mirabeau, Olivier Delattre, and Thomas Gp Grünewald. 2017. “MYBL2 (B-Myb): A Central Regulator of Cell Proliferation, Cell Survival and Differentiation Involved in Tumorigenesis.” *Cell Death & Disease* 8 (6): e2895.

Mutis, Tuna, Rob Verdijk, Ellen Schrama, Bennie Esendam, Anneke Brand, and Els Goulmy. 1999. “Feasibility of Immunotherapy of Relapsed Leukemia With Ex Vivo–Generated Cytotoxic T Lymphocytes Specific for Hematopoietic System-Restricted Minor Histocompatibility Antigens.” *Blood* 93 (7): 2336–41.

Nagata, Naoko, Tadahilo Oshida, Ning Lu Yoshida, Noriko Yuyama, Yuji Sugita, Gozoh Tsujimoto, Toshio Katsunuma, Akira Akasawa, and Hirohisa Saito. 2003. “Analysis of Highly Expressed Genes in Monocytes from Atopic Dermatitis Patients.” *International Archives of Allergy and Immunology* 132 (2): 156–67.

Nakamae, Hirohisa, Barry E. Storer, Rainer Storb, Jan Storek, Thomas R. Chauncey, Michael A. Pulsipher, Finn B. Petersen, et al. 2010. “Low-Dose Total Body Irradiation and Fludarabine Conditioning for HLA Class I-Mismatched Donor Stem Cell Transplantation and Immunologic Recovery in Patients with Hematologic Malignancies: A Multicenter Trial.” *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 16 (3): 384–94.

Nakasone, Hideki, Bita Sahaf, and David B. Miklos. 2015. “Therapeutic Benefits Targeting B-Cells in Chronic Graft-versus-Host Disease.” *International Journal of Hematology* 101 (5): 438–51.

Nakasone, Hideki, Bita Sahaf, Lu Tian, Tao Wang, Michael D. Haagenson, Kelsi Schoenrock, Spenser Perloff, et al. 2016. “Presensitization to HY Antigens in

Female Donors prior to Transplant Is Not Associated with Male Recipient Post-Transplant HY Antibody Development nor with Clinical Outcomes." *Haematologica* 101 (1): e30–33.

Nakasone, Hideki, Lu Tian, Bita Sahaf, Takakazu Kawase, Kelsi Schoenrock, Spenser Perloff, Christine E. Ryan, et al. 2015. "Allogeneic HY Antibodies Detected 3 Months after Female-to-Male HCT Predict Chronic GVHD and Nonrelapse Mortality in Humans." *Blood* 125 (20): 3193–3201.

NCBI Resource Coordinators. 2016. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 44 (D1): D7–19.

Nembrini, Stefano, Inke R. König, and Marvin N. Wright. 2018. "The Revival of the Gini Importance?" *Bioinformatics* . https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty373/4994791.

Nicodemus, Kristin K. 2011. "Letter to the Editor: On the Stability and Ranking of Predictors from Random Forest Variable Importance Measures." *Briefings in Bioinformatics* 12 (4): 369–73.

Nielsen, H. S., R. Steffensen, M. Lund, L. Egestad, L. H. Mortensen, A-M N. Andersen, Ø. Lidegaard, and O. B. Christiansen. 2010. "Frequency and Impact of Obstetric Complications Prior and Subsequent to Unexplained Secondary Recurrent Miscarriage." *Human Reproduction* 25 (6): 1543–52.

Nielsen, Morten, and Massimo Andreatta. 2016. "NetMHCpan-3.0; Improved Prediction of Binding to MHC Class I Molecules Integrating Information from Multiple Receptor and Peptide Length Datasets." *Genome Medicine* 8 (1): 33.

Nowacka-Zawisza, Maria, Ewelina Wiśnik, Andrzej Wasilewski, Milena Skowrońska, Ewa Forma, Magdalena Bryś, Waldemar Różański, and Wanda M. Krajewska. 2015. "Polymorphisms of Homologous Recombination RAD51, RAD51B, XRCC2, and XRCC3 Genes and the Risk of Prostate Cancer." *Analytical Cellular Pathology* 2015 (August): 828646.

Nunes, Eduardo, Helen Heslop, Marcelo Fernandez-Vina, Cynthia Taves, Dawn R. Wagenknecht, A. Bradley Eisenbrey, Gottfried Fischer, et al. 2011. "Definitions of Histocompatibility Typing Terms." *Blood* 118 (23): e180–83.

Ofran, Y., H. T. Kim, V. Brusic, L. Blake, and M. Mandrell. 2010. "Diverse Patterns of T-Cell Response against Multiple Newly Identified Human Y Chromosome–encoded Minor Histocompatibility Epitopes." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*. http://clincancerres.aacrjournals.org/content/early/2010/02/12/1078-0432.CCR-09-2701.abstract.

Okamoto, Koji, Naoko Iwasaki, Kent Doi, Eisei Noiri, Yasuhiko Iwamoto, Yasuko Uchigata, Toshiro Fujita, and Katsushi Tokunaga. 2012. "Inhibition of Glucose-Stimulated Insulin Secretion by KCNJ15, a Newly Identified Susceptibility Gene for Type 2 Diabetes." *Diabetes* 61 (7): 1734–41.

Okamoto, Koji, Naoko Iwasaki, Chisa Nishimura, Kent Doi, Eisei Noiri, Shinko

Nakamura, Miho Takizawa, et al. 2010. “Identification of KCNJ15 as a Susceptibility Gene in Asian Patients with Type 2 Diabetes Mellitus.” *American Journal of Human Genetics* 86 (1): 54–64.

Oostvogels, R., H. M. Lokhorst, and T. Mutis. 2016. “Minor Histocompatibility Ags: Identification Strategies, Clinical Results and Translational Perspectives.” *Bone Marrow Transplantation* 51 (2): 163–71.

Ott, P. A., Z. Hu, D. B. Keskin, S. A. Shukla, J. Sun, and D. J. Bozym. 2017. “An Immunogenic Personal Neoantigen Vaccine for Patients with Melanoma.” *Nature*. https://www.nature.com/articles/nature22991.

Pan, Zui, Sangyong Choi, Halima Ouadid-Ahidouch, Jin-Ming Yang, John H. Beattie, and Irina Korichneva. 2017. “Zinc Transporters and Dysregulated Channels in Cancers.” *Frontiers in Bioscience* 22 (January): 623–43.

Park, Yongjung, June-Won Cheong, Myoung Hee Park, Myoung Soo Kim, Jong Sun Kim, and Hyon-Suk Kim. 2016. “Effect of Major Histocompatibility Complex Haplotype Matching by C4 and MICA Genotyping on Acute Graft versus Host Disease in Unrelated Hematopoietic Stem Cell Transplantation.” *Human Immunology* 77 (2): 176–83.

Pasquini, M. C., and X. Zhu. 2015. “Current Uses and Outcomes of Hematopoietic Stem Cell Transplantation: CIBMTR Summary Slides, 2015.” In *Center for International Blood & Marrow Transplant Research Conference*. Vol. 2014. Available at: http://www.cibmtr.org.

Peker, Deniz. 2018. “Navigating through Mutations in Acute Myeloid Leukemia. What Do We Know and What Do We Do with It?” *Erciyes Medical Journal* 40 (4): 183–87.

Peng, Hanchuan, Fuhui Long, and Chris Ding. 2005. “Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (8): 1226–38.

Perkey, Eric, and Ivan Maillard. 2018. “New Insights into Graft-Versus-Host Disease and Graft Rejection.” *Annual Review of Pathology* 13 (January): 219–45.

Persson, Emma K., Heli Uronen-Hansson, Monika Semmrich, Aymeric Rivollier, Karin Hägerbrand, Jan Marsal, Sigurdur Gudjonsson, et al. 2013. “IRF4 Transcription-Factor-Dependent CD103+ CD11b+ Dendritic Cells Drive Mucosal T Helper 17 Cell Differentiation.” *Immunity* 38 (5): 958–69.

Petersdorf, Effie W. 2013. “The Major Histocompatibility Complex: A Model for Understanding Graft-versus-Host Disease.” *Blood* 122 (11): 1863–72.

———. 2015. “HLA Mismatching in Transplantation.” *Blood* 125 (7): 1058–59.

———. 2017. “Which Factors Influence the Development of GVHD in HLA-Matched or Mismatched Transplants?” *Best Practice & Research. Clinical Haematology* 30 (4): 333–35.

Petersdorf, Effie W., Claudio Anasetti, Paul J. Martin, and John A. Hansen. 2018. “HLA Typing in Support of Hematopoietic Cell Transplantation from Unrelated Donors.” In

Neoplastic Diseases of the Blood, edited by Peter H. Wiernik, Janice P. Dutcher, and Morie A. Gertz, 1193–1209. Cham: Springer International Publishing.

Petersdorf, Effie W., Mari Malkki, Colm O'hUigin, Mary Carrington, Ted Gooley, Michael D. Haagenson, Mary M. Horowitz, Stephen R. Spellman, Tao Wang, and Philip Stevenson. 2015. "High HLA-DP Expression and Graft-versus-Host Disease." *The New England Journal of Medicine* 373 (7): 599–609.

Pfeffer, P. F., and E. Thorsby. n.d. "HLA-RESTRICTED CYTOTOXICITY AGAINST MALE-SPECIFIC (H-Y) ANTIGEN AFTER ACUTE REJECTION OF AN HLA-IDENTICAL SIBLING KIDNEY CLONAL DISTRIBUTION OF THE CYTOTOXIC CELLS." *Transplantation* 33 (1): 52–56.

Pidala, Joseph, Stephanie J. Lee, Kwang Woo Ahn, Stephen Spellman, Hai-Lin Wang, Mahmoud Aljurf, Medhat Askar, et al. 2014. "Nonpermissive HLA-DPB1 Mismatch Increases Mortality after Myeloablative Unrelated Allogeneic Hematopoietic Cell Transplantation." *Blood* 124 (16): 2596–2606.

Pont, Margot J., Dyantha I. van der Lee, Edith D. van der Meijden, Cornelis A. M. van Bergen, Michel G. D. Kester, Maria W. Honders, Martijn Vermaat, et al. 2016. "Integrated Whole Genome and Transcriptome Analysis Identified a Therapeutic Minor Histocompatibility Antigen in a Splice Variant of ITGB2." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 22 (16): 4185–96.

Popli, Rakesh, Bita Sahaf, Hideki Nakasone, Joyce Yeuk Yu Lee, and David B. Miklos. 2014. "Clinical Impact of H-Y Alloimmunity." *Immunologic Research* 58 (2-3): 249–58.

Przepiorka, D., D. Weisdorf, P. Martin, H. G. Klingemann, P. Beatty, J. Hows, and E. D. Thomas. 1995. "1994 Consensus Conference on Acute GVHD Grading." *Bone Marrow Transplantation* 15 (6): 825–28.

Raczy, Come, Roman Petrovski, Christopher T. Saunders, Ilya Chorny, Semyon Kruglyak, Elliott H. Margulies, Han-Yu Chuang, et al. 2013. "Isaac: Ultra-Fast Whole-Genome Secondary Analysis on Illumina Sequencing Platforms." *Bioinformatics* 29 (16): 2041–43.

Radom-Aizik, Shlomit, Frank Zaldivar Jr, Stacy Oliver, Pietro Galassetti, and Dan M. Cooper. 2010. "Evidence for microRNA Involvement in Exercise-Associated Neutrophil Gene Expression Changes." *Journal of Applied Physiology* 109 (1): 252–61.

Rajasagi, Mohini, Sachet A. Shukla, Edward F. Fritsch, Derin B. Keskin, David DeLuca, Ellese Carmona, Wandi Zhang, et al. 2014. "Systematic Identification of Personal Tumor-Specific Neoantigens in Chronic Lymphocytic Leukemia." *Blood* 124 (3): 453–62.

Ratti, Stefano, Matilde Yung Follo, Giulia Ramazzotti, Irene Faenza, Roberta Fiume, Pann Ghill Suh, James A. McCubrey, Lucia Manzoli, and Lucio Cocco. 2018. "Nuclear Phospholipase C Isoenzyme Imbalance Leads to Pathologies in Brain,

Hematologic, Neuromuscular and Fertility Disorders." *Journal of Lipid Research*, October. https://doi.org/10.1194/jlr.R089763.

Razzaq, Badar Abdul, Allison Scalora, Vishal N. Koparde, Jeremy Meier, Musa Mahmood, Salman Salman, Max Jameson-Lee, et al. 2016. "Dynamical System Modeling to Simulate Donor T Cell Response to Whole Exome Sequencing-Derived Recipient Peptides Demonstrates Different Alloreactivity Potential in HLA-Matched and-Mismatched Donor--Recipient Pairs." *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 22 (5): 850–61.

Reya, Tannishtha, and Hans Clevers. 2005. "Wnt Signalling in Stem Cells and Cancer." *Nature* 434 (7035): 843–50.

Reya, Tannishtha, Andrew W. Duncan, Laurie Ailles, Jos Domen, David C. Scherer, Karl Willert, Lindsay Hintz, Roel Nusse, and Irving L. Weissman. 2003. "A Role for Wnt Signalling in Self-Renewal of Haematopoietic Stem Cells." *Nature* 423 (6938): 409–14.

Riesner, Katarina, Yu Shi, Angela Jacobi, Martin Kräter, Martina Kalupa, Aleixandria McGearey, Sarah Mertlitz, et al. 2017. "Initiation of Acute Graft-versus-Host Disease by Angiogenesis." *Blood* 129 (14): 2021–32.

Robinson, James, Jason A. Halliwell, James D. Hayhurst, Paul Flicek, Peter Parham, and Steven G. E. Marsh. 2015. "The IPD and IMGT/HLA Database: Allele Variant Databases." *Nucleic Acids Research* 43 (Database issue): D423–31.

Roelen, Dave, Yvonne de Vaal, Cynthia Vierra-Green, Stephanie Waldvogel, Stephen Spellman, Frans Claas, and Machteld Oudshoorn. 2018. "HLA Mismatches That Are Identical for the Antigen Recognition Domain Are Less Immunogenic." *Bone Marrow Transplantation* 53 (6): 729–40.

Rollinson, Sara, Alexandra G. Smith, James M. Allan, Peter J. Adamson, Kathryn Scott, Christine F. Skibola, Martyn T. Smith, and Gareth J. Morgan. 2007. "RAD51 Homologous Recombination Repair Gene Haplotypes and Risk of Acute Myeloid Leukaemia." *Leukemia Research* 31 (2): 169–74.

Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. 2007. "A Review of Feature Selection Techniques in Bioinformatics." *Bioinformatics* 23 (19): 2507–17.

Sahaf, B., Y. Yang, and S. Arai. 2013. "H–Y Antigen-Binding B Cells Develop in Male Recipients of Female Hematopoietic Cells and Associate with Chronic Graft vs. Host Disease." *Proceedings of the*. http://www.pnas.org/content/early/2013/01/30/1222900110.short.

Sahin, Ugur, Evelyna Derhovanessian, Matthias Miller, Björn-Philipp Kloke, Petra Simon, Martin Löwer, Valesca Bukur, et al. 2017. "Personalized RNA Mutanome Vaccines Mobilize Poly-Specific Therapeutic Immunity against Cancer." *Nature* 547 (7662): 222–26.

Sampson, Juliana K., Nihar U. Sheth, Vishal N. Koparde, Allison F. Scalora, Myrna G. Serrano, Vladimir Lee, Catherine H. Roberts, et al. 2014. "Whole Exome

Sequencing to Estimate Alloreactivity Potential between Donors and Recipients in Stem Cell Transplantation." *British Journal of Haematology* 166 (4): 566–70.

Sanchez-Mazas, Alicia, and Diogo Meyer. 2014. "The Relevance of HLA Sequencing in Population Genetics Studies." *Journal of Immunology Research* 2014 (July): 971818.

Schlitzer, Andreas, Naomi McGovern, Pearline Teo, Teresa Zelante, Koji Atarashi, Donovan Low, Adrian W. S. Ho, et al. 2013. "IRF4 Transcription Factor-Dependent CD11b+ Dendritic Cells in Human and Mouse Control Mucosal IL-17 Cytokine Responses." *Immunity* 38 (5): 970–83.

Schmitz, N., P. Dreger, M. Suttorp, E. B. Rohwedder, T. Haferlach, H. Loffler, A. Hunter, and N. H. Russell. 1995. "Primary Transplantation of Allogeneic Peripheral Blood Progenitor Cells Mobilized by Filgrastim (granulocyte Colony-Stimulating Factor) [see Comments]." *Blood* 85 (6): 1666–72.

Sellami, M. H., A. Ben Ahmed, H. Kaabi, A. Jridi, A. Dridi, and S. Hmida. 2010. "HA-1 and HA-2 Minor Histocompatibility Antigens in Tunisians." *Tissue Antigens* 75 (6): 720–23.

Shaw, B. E., N. P. Mayor, R. M. Szydlo, W. P. Bultitude, C. Anthias, K. Kirkland, J. Perry, et al. 2017. "Recipient/donor HLA and CMV Matching in Recipients of T-Cell-Depleted Unrelated Donor Haematopoietic Cell Transplants." *Bone Marrow Transplantation* 52 (5): 717–25.

Shaw, B. E., J. Robinson, K. Fleischhauer, J. A. Madrigal, and S. G. E. Marsh. 2013. "Translating the HLA-DPB1 T-Cell Epitope-Matching Algorithm into Clinical Practice." *Bone Marrow Transplantation* 48 (12): 1510–12.

Shiina, Takashi, Kazuyoshi Hosomichi, Hidetoshi Inoko, and Jerzy K. Kulski. 2009. "The HLA Genomic Loci Map: Expression, Interaction, Diversity and Disease." *Journal of Human Genetics* 54 (1): 15–39.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. "Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega." *Molecular Systems Biology* 7 (1): 539.

Smith, Albert V., Daryl J. Thomas, Heather M. Munro, and Gonçalo R. Abecasis. 2005. "Sequence Features in Regions of Weak and Strong Linkage Disequilibrium." *Genome Research* 15 (11): 1519–34.

Socié, Gerard, and Bruce R. Blazar. 2009. "Acute Graft-versus-Host Disease: From the Bench to the Bedside." *Blood* 114 (20): 4327–36.

Spellman, Stephen, Michelle Setterholm, Martin Maiers, Harriet Noreen, Machteld Oudshoorn, Marcelo Fernandez-Viña, Effie Petersdorf, et al. 2008. "Advances in the Selection of HLA-Compatible Donors: Refinements in HLA Typing and Matching over the First 20 Years of the National Marrow Donor Program Registry." *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 14 (9 Suppl): 37–44.

Srinivasan, Mythily, Daniel Sedmak, and Scott Jewell. 2002. “Effect of Fixatives and Tissue Processing on the Content and Integrity of Nucleic Acids.” *The American Journal of Pathology* 161 (6): 1961–71.

Stern, M., R. Brand, T. De Witte, and A. Sureda. 2008. “Female-versus-male Alloreactivity as a Model for Minor Histocompatibility Antigens in Hematopoietic Stem Cell Transplantation.” *American Journal of*. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-6143.2008.02374.x.

Stokes, Matthew E., and Shyam Visweswaran. 2012. “Application of a Spatially-Weighted Relief Algorithm for Ranking Genetic Predictors of Disease.” *BioData Mining* 5 (1): 20.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.” *BMC Bioinformatics* 8 (January): 25.

Strønen, Erlend, Mireille Toebes, Sander Kelderman, Marit M. van Buuren, Weiwen Yang, Nienke van Rooij, Marco Donia, et al. 2016. “Targeting of Cancer Neoantigens with Donor-Derived T Cell Receptor Repertoires.” *Science* 352 (6291): 1337–41.

Sullivan, Michael J. 2008. “Banking on Cord Blood Stem Cells.” *Nature Reviews. Cancer* 8 (7): 555–63.

The GTEx Consortium. 2015. “The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans.” *Science* 348 (6235): 648–60.

Tichelli, André, Eric Beohou, Myriam Labopin, Gérard Socié, Alicia Rovó, Manuela Badoglio, Anja van Biezen, et al. 2018. “Evaluation of Second Solid Cancers After Hematopoietic Stem Cell Transplantation in European Patients.” *JAMA Oncology*, November. https://doi.org/10.1001/jamaoncol.2018.4934.

Tickenbrock, Lara, Sina Hehn, Bülent Sargin, Chunaram Choudhary, Nicole Bäumer, Horst Buerger, Bernd Schulte, et al. 2008. “Activation of Wnt Signalling in Acute Myeloid Leukemia by Induction of Frizzled-4.” *International Journal of Oncology* 33 (6): 1215–21.

Tram, Kevin, Gretta Stritesky, Kim Wadsworth, Jennifer Ng, Claudio Anasetti, and Jason Dehn. 2017. “Identification of DPB1 Permissive Unrelated Donors Is Highly Likely.” *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 23 (1): 81–86.

Tran, Eric, Mojgan Ahmadzadeh, Yong-Chen Lu, Alena Gros, Simon Turcotte, Paul F. Robbins, Jared J. Gartner, et al. 2015. “Immunogenicity of Somatic Mutations in Human Gastrointestinal Cancers.” *Science* 350 (6266): 1387–90.

Tseng, Li-Hui, Barry Storer, Effie Petersdorf, Ming-Tseh Lin, Jason W. Chien, Bryan M. Grogan, Mari Malkki, et al. 2009. “IL10 and IL10 Receptor Gene Variation and Outcomes after Unrelated and Related Hematopoietic Cell Transplantation.” *Transplantation* 87 (5): 704–10.

Tyner, J. W., M. L. Rutenberg-Schoenberg, H. Erickson, S. G. Willis, T. O’Hare, M. W.

Deininger, B. J. Druker, and M. M. Loriaux. 2009. “Functional Characterization of an Activating TEK Mutation in Acute Myeloid Leukemia: A Cellular Context-Dependent Activating Mutation.” *Leukemia* 23 (7): 1345–48.

Uinuk-Ool, Tatiana S., Naoko Takezaki, and Jan Klein. 2003. “Ancestry and Kinships of Native Siberian Populations: The HLA Evidence.” *Evolutionary Anthropology: Issues, News, and Reviews* 12 (5): 231–45.

Urbanowicz, Ryan J., Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. 2018. “Relief-Based Feature Selection: Introduction and Review.” *Journal of Biomedical Informatics* 85 (September): 189–203.

Urbanowicz, Ryan J., Randal S. Olson, Peter Schmitt, Melissa Meeker, and Jason H. Moore. 2018. “Benchmarking Relief-Based Feature Selection Methods for Bioinformatics Data Mining.” *Journal of Biomedical Informatics* 85 (September): 168–88.

Vogt, Mario H. J., Els Goulmy, Freke M. Kloosterboer, Els Blokland, Roel A. de Paus, Roel Willemze, and J. H. Frederik Falkenburg. 2000. “UTY Gene Codes for an HLA-B60--Restricted Human Male-Specific Minor Histocompatibility Antigen Involved in Stem Cell Graft Rejection: Characterization of the Critical Polymorphic Amino Acid Residues for T-Cell Recognition.” *Blood* 96 (9): 3126–32.

Wagner, Steven. 2012. “H-Y Antigen in Kidney Transplant: Does Gender Matter?” *Gender Medicine* 9 (5): 387–88.

Wang, Wei, Hu Huang, Michael Halagan, Cynthia Vierra-Green, Michael Heuer, Jason E. Brelsford, Michael Haagenson, et al. 2018. “Chromosome Y--Encoded Antigens Associate with Acute Graft-versus-Host Disease in Sex-Mismatched Stem Cell Transplant.” *Blood Advances* 2 (19): 2419–29.

Wang, Y., A. V. Krivtsov, A. U. Sinha, and T. E. North. 2010. “The Wnt/β-Catenin Pathway Is Required for the Development of Leukemia Stem Cells in AML.” http://science.sciencemag.org/content/327/5973/1650.short.

Wellek, Stefan. 2017. “A Critical Evaluation of the Current ‘p-Value Controversy.’” *Biometrical Journal. Biometrische Zeitschrift* 59 (5): 854–72.

Welniak, Lisbeth A., Bruce R. Blazar, and William J. Murphy. 2007. “Immunobiology of Allogeneic Hematopoietic Stem Cell Transplantation.” *Annual Review of Immunology* 25: 139–70.

Whitcomb, David C., and Mark E. Lowe. 2007. “Human Pancreatic Digestive Enzymes.” *Digestive Diseases and Sciences* 52 (1): 1–17.

Wilson Sayres, Melissa A., Kirk E. Lohmueller, and Rasmus Nielsen. 2014. “Natural Selection Reduced Diversity on Human Y Chromosomes.” *PLoS Genetics* 10 (1): e1004064.

Wright, Marvin N., and Andreas Ziegler. 2017. “Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.” *Journal of Statistical Software* 77 (i01). http://arxiv.org/abs/1508.04409.

Xie, Tao, Lee Rowen, Begoña Aguado, Mary Ellen Ahearn, Anup Madan, Shizhen Qin,

R. Duncan Campbell, and Leroy Hood. 2003. “Analysis of the Gene-Dense Major Histocompatibility Complex Class III Region and Its Comparison to Mouse.” *Genome Research* 13 (12): 2621–36.

Xing, Eric P., Michael I. Jordan, Richard M. Karp, and Others. 2001. “Feature Selection for High-Dimensional Genomic Microarray Data.” In *ICML*, 1:601–8. Citeseer.

Yang, Joshua Y. C., and Minnie M. Sarwal. 2017. “Transplant Genetics and Genomics.” *Nature Reviews. Genetics* 18 (5): 309–26.

Zeiser, Robert, Reinhard Marks, Hartmut Bertz, and Jürgen Finke. 2004. “Immunopathogenesis of Acute Graft-versus-Host Disease: Implications for Novel Preventive and Therapeutic Strategies.” *Annals of Hematology* 83 (9): 551–65.

Zheng, Ye, Ashutosh Chaudhry, Arnold Kas, Paul deRoos, Jeong M. Kim, Tin-Tin Chu, Lynn Corcoran, Piper Treuting, Ulf Klein, and Alexander Y. Rudensky. 2009. “Regulatory T-Cell Suppressor Program Co-Opts Transcription Factor IRF4 to Control T(H)2 Responses.” *Nature* 458 (7236): 351–56.

Zhou, Guo-Ping, Clara Wong, Robert Su, Scott C. Crable, Kathleen P. Anderson, and Patrick G. Gallagher. 2004. “Human Potassium Chloride Cotransporter 1 (SLC12A4) Promoter Is Regulated by AP-2 and Contains a Functional Downstream Promoter Element.” *Blood* 103 (11): 4302–9.

Zino, Elisabetta, Guido Frumento, Sarah Marktel, Maria Pia Sormani, Francesca Ficara, Simona Di Terlizzi, Anna Maria Parodi, et al. 2004. “A T-Cell Epitope Encoded by a Subset of HLA-DPB1 Alleles Determines Nonpermissive Mismatches for Hematologic Stem Cell Transplantation.” *Blood* 103 (4): 1417–24.

Zino, Elisabetta, Luca Vago, Simona Di Terlizzi, Benedetta Mazzi, Laura Zito, Elisabetta Sironi, Silvano Rossini, et al. 2007. “Frequency and Targeted Detection of HLA-DPB1 T Cell Epitope Disparities Relevant in Unrelated Hematopoietic Stem Cell Transplantation.” *Biology of Blood and Marrow Transplantation: Journal of the American Society for Blood and Marrow Transplantation* 13 (9): 1031–40.

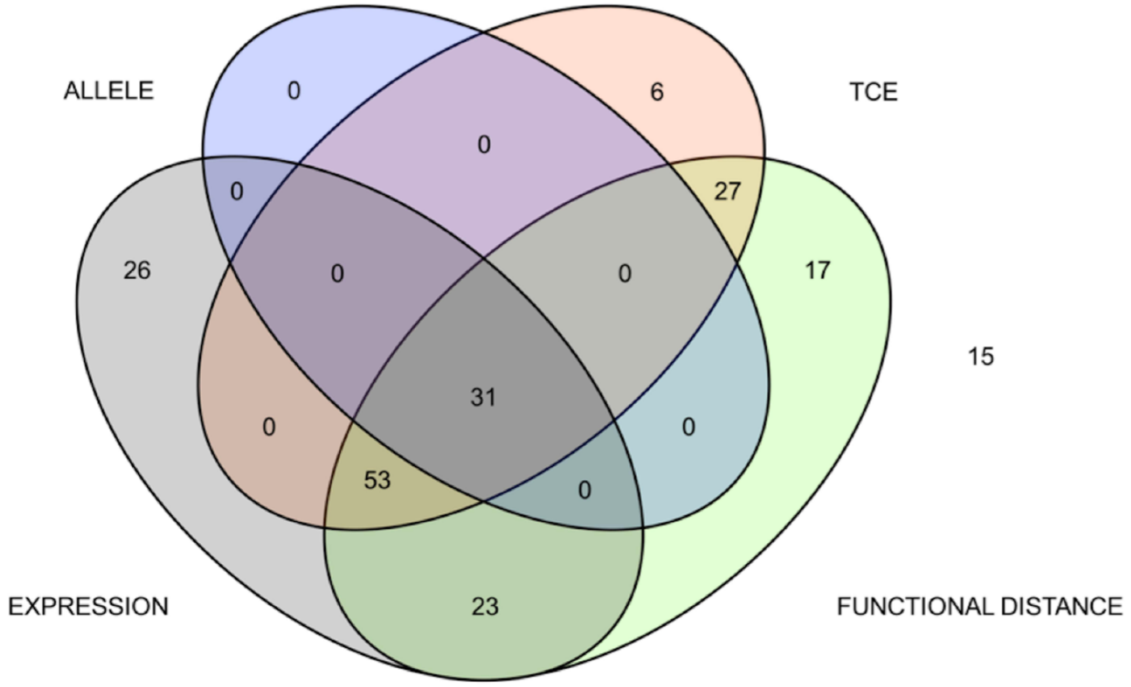
Appendix

A.1 Supplementary tables and figures for Chapter 3

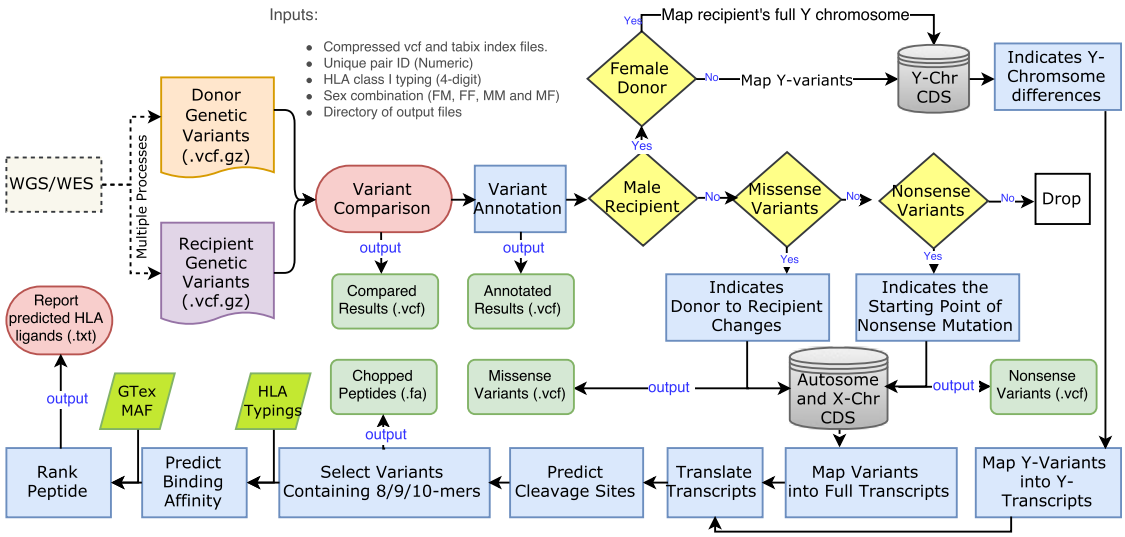
[*Supplementary Table 3.1.xlsx*]

Autosomal MiHAs do not associate with acute GVHD. Each row represents a single MiHA present in the whole genome sequencing (WGS) or microarray cohort (or both) with characterized restricted HLA allele, encoding single nucleotide polymorphism (SNP), gene, and counts in acute GVHD and non-GVHD groups. *P*-values were calculated, and multiple hypotheses corrected for, as described in Methods. MAF – minor allele frequency.

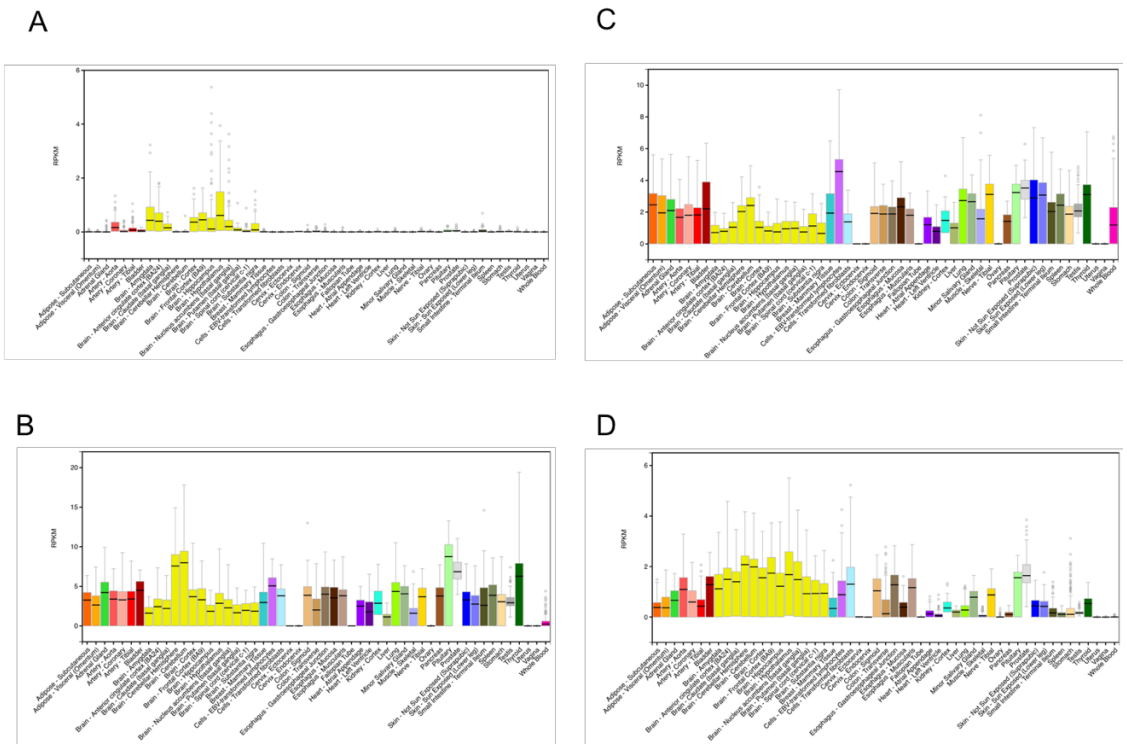
Supplementary Figure 3.1: A Venn diagram shows individual counts in overlapping categories for donor-recipient pairs that were retrospectively matched at HLA-DPB1 by allele, T-cell epitope permissibility (TCE), expression, and functional distance.



Supplementary Figure 3.2: A genomic annotation workflow to identify known and novel outcomes-associated variants. Raw sequence data were processed as described (Methods) to generate a single binary alignment (BAM) and variant call format (VCF) file per sample. Comparative analysis of donor-recipient pairs resulted in a single VCF file containing patient-specific variants, which were annotated further. Male recipients with female donors were treated separately to analyze variants on the Y chromosome.



Supplementary Figure 3.3: Tissue-specific gene expression for PCDH11Y (A), USP9Y (B), UTY (C), and NLGN4Y (D).



A.2 Supplementary materials for Chapter 4

Supplementary tables 4.1-4.4

[*Supplementary tables Chapter 4.xlsx*]

Supplementary Table 4.1 Top ranking SNPs from GIFI for Scenario 1: AML case-control
Supplementary Table 4.2 Top ranking SNPs from PEFI for Scenario 1: AML case-control
Supplementary Table 4.3 Top ranking SNPs from GIFI for Scenario 2: acute GVHD case-control
Supplementary Table 4.4 Top ranking SNPs from PEFI for Scenario 2: acute GVHD case-control

Supplementary Table 4.5. Characteristics of acute myeloid leukemia patients transplant cases

Variable	Value	
No. of transplant cases	331	
No. of transplant centers	97	
Recipient age, median (range), years	42 (0~65 yrs)	
Age at transplant, yr		
<20	42	(12.7%)
20-59	276	(83.4%)
≥60	13	(3.9%)
Recipient race group		
CAU	319	(96.4%)
AFA	2	(0.6%)
API	8	(2.4%)
HIS	0	(0 %)
Native American	1	(0.3%)
Other/multiple/declined/unknown	1	(0.3%)
Donor age, yrs		
18-29	152	(45.9%)
30-39	89	(26.8%)
40-49	71	(21.5%)
≥50	19	(5.7%)
Donor race group		
CAU	284	(85.8%)
AFA	3	(0.9%)
API	9	(2.7%)
HIS	0	(0 %)
Native American	0	(0 %)
Other/multiple/declined/unknown	35	(10.6%)
AML disease status at transplant		
Early	243	(73.4%)
Intermediate	10	(3.0%)
Advanced	77	(23.3%)
Other	1	(0.3%)
Graft Type		
Bone marrow	132	(39.9%)
Peripheral blood	199	(60.1%)
In vivo T cell depletion		
No	238	(71.9%)
Yes	93	(28.1%)
ATG given		
No	243	(73.4%)
Yes	88	(26.6%)
Donor/recipient CMV match		
Negative/negative	107	(32.3%)
Negative/positive	114	(34.4%)
Positive/negative	33	(10.0%)
Positive/positive	69	(20.9%)

Unknown	8	(2.4%)
GVHD prophylaxis		
Ex vivo T cell depletion	20	(6.0%)
CD34 Selection	0	(0%)
Cyclophosphamide	5	(1.5%)
Tacrolimus + (MTX or MMF) ± other	206	(62.2%)
Tacrolimus ± other	15	(4.5%)
Tacrolimus alone	4	(1.2%)
CsA + (MMF or MTX) ± other (except	74	(22.4%)
Tacrolimus)	2	(0.6%)
CsA ± other (no MTX nor MMF)	3	(0.9%)
CsA alone	2	(0.6%)
Other		
HLA-DPB1 typing	58	(17.5%)
Double mismatch	115	(34.7%)
Single mismatch	31	(9.4%)
Matched	127	(38.4%)
Missing/not typed		
Donor/recipient sex match	127	(38.4%)
Male/male	97	(29.3%)
Male/female	40	(12.1%)
Female/male	67	(20.2%)
Female/female		
GVHD outcome	185	(55.9%)
Grades 0~I acute GVHD	146	(44.1%)
Grades II~IV acute GVHD		

Supplementary Table 4.6. HLA typing summaries

HLA-A	Freq
A*02:01	174
A*01:01	115
A*03:01	94
A*24:02	50
A*11:01	32
A*29:02	21
A*68:01	21
A*26:01	18
A*32:01	17
A*31:01	15
A*25:01	13
A*23:01	12
A*30:01	12
A*33:01	6
A*68:02	6
A*02:05	3
A*30:02	3
A*03:02	2
A*24:03	1
A*24:17	1
A*30:04	1
A*36:01	1

HLA-B	Freq
B*07:02	114
B*08:01	107
B*44:02	73
B*40:01	41
B*15:01	40
B*35:01	37
B*18:01	36
B*27:05	21
B*13:02	19
B*44:03	19
B*57:01	16
B*14:02	15
B*38:01	13
B*37:01	11
B*52:01	8
B*49:01	7
B*51:01	6
B*55:01	6
B*56:01	6
B*14:01	5
B*40:02	5
B*35:03	3
B*39:01	3
B*41:01	3
B*45:01	3
B*50:01	3
B*15:17	2
B*27:02	2
B*35:02	2
B*47:01	2
B*53:01	2
B*58:01	2
B*15:02	1
B*15:03	1
B*15:10	1
B*15:18	1
B*35:08	1
B*58:02	1

HLA-C	Freq
C*07:01	126
C*07:02	115
C*05:01	71
C*03:04	56
C*06:02	56
C*04:01	54
C*03:03	29
C*12:03	26
C*08:02	20
C*01:02	19
C*02:02	16
C*16:01	12
C*07:04	10
C*12:02	8
C*14:02	3
C*17:01	2
C*02:10	1
C*08:01	1
C*15:02	1

HLA-DRB1	Freq
DRB1*15:01	116
DRB1*03:01	105
DRB1*07:01	74
DRB1*01:01	60
DRB1*04:01	56
DRB1*13:01	37
DRB1*11:01	30
DRB1*13:02	24
DRB1*04:04	23
DRB1*08:01	13
DRB1*11:04	12
DRB1*12:01	12
DRB1*01:02	11
DRB1*14:01	11
DRB1*01:03	10
DRB1*15:02	8
DRB1*04:02	4
DRB1*09:01	4
DRB1*13:03	4
DRB1*04:07	3
DRB1*10:01	3
DRB1*16:01	3
DRB1*04:03	2
DRB1*04:05	2
DRB1*11:03	2
DRB1*08:02	1
DRB1*08:04	1
DRB1*11:02	1
DRB1*12:02	1
DRB1*13:04	1
DRB1*13:05	1
DRB1*14:02	1
DRB1*16:02	1

HLA-DQB1	Freq
DQB1*06:02	116
DQB1*02:01	106
DQB1*03:01	103
DQB1*05:01	77
DQB1*02:02	59
DQB1*03:02	48
DQB1*06:03	37
DQB1*06:04	21
DQB1*03:03	18
DQB1*04:02	15
DQB1*05:03	11
DQB1*06:01	8
DQB1*05:02	4
DQB1*06:09	3

Supplementary Material 4.7: Microarray genotype data preprocessing

The original SNP genotype data was obtained and processed by Madbouly et al. and described in (Madbouly et al. 2017). In this study, we performed the quality control separately and did not use the imputed genotypes.

We have removed individuals that show ambiguous sex genotype than their reported sex. The rest of parameters used in the SNP filtering is as follows. 1) If a SNP minor allele frequency (MAF) is less than 0.005 or showed up in less than 10 individuals, then those SNPs are filtered out. 2) SNPs that have less than 95% call rate are removed. 3) It is recommended to use the control-only samples for Hardy-Weinberg equilibrium (HWE) test (Anderson et al. 2010). Here we used the healthy donor-only samples and excluded the SNPs that have P-values lower than 0.001 after the HWE test. After these steps, we obtained 630,793 SNPs for the 331 donor-recipient pairs.