GENETIC ASSOCIATIONS IN ACUTE LEUKEMIA PATIENTS AFTER MATCHED UNRELATED DONOR ALLOGENEIC HEMATOPOETIC STEM CELL TRANSPLANTATION

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Abbas A Rizvi

Graduate Program in Pharmaceutical Sciences

The Ohio State University

2019

Dissertation Committee:

Lara E Sucheston-Campbell, MS, PhD, Adviser

Guy Brock, PhD

Moray Campbell, PhD

Shili Lin, PhD

ABSTRACT

Here I will be writing an abstract that summarizes my dissertation results

DEDICATION

Dedicated to Ezgi, my parents, my siblings and their kids, and my sweet little pooch Bernie.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my adviser Dr. Lara Sucheston-Campbell. She has been an incredible mentor for me and has set me up with any and every opportunity that I had my eyes set on.

I am forever indebted to Dr. Barbara Foster, for taking a chance on me and giving me the opportunity to get an interview for the PhD program at RPCI, at the eleventh hour before the academic year began in 2015.

Thank you to Dr. Martin Morgan for taking time out of his extremely busy schedule to work with me one-on-one and teach me the inner workings of R.

I would not have gone done the road of computational biology/bioinformatics if it were not for two very important mentors of mine: Dr. Moray Campbell and Dr. Sebastiano Battaglia. Dr. Campbell set me up with a life defining opprtunity: the Cancer and Systems Biology program at the University of Luxembourg and VU University Amsterdam. And Dr. Battaglia for his mentorship during my Master's thesis.

And thank you to Ezgi Karaesmen – if it were not for you, I truly believe I would not had the level of success I have had for past four years. Your love, your support, your intelligence, and your ability to challenge me on my scientific thought, my coding, and my ability to think have been instrumental my development as a scientist – I cannot thank you enough.

This work was supported by the National Institute of Health/National Heart, Blood, and Lung Institute R01HL102278 and National Institute of Health/National Cancer Institute R03CA188733. Support provided by the Center for Computational Research at the University at Buffalo.

VITA

2012	B.S. Biology, SUNY Fredonia
2015	M.Sc. Integrated Systems Biology, University of Luxembourg
2015	M.S. Natural Sciences, University at Buffalo
2016-present	Graduate Research Associate, Department of Pharmaceutics, The Ohio State University

Publications

[1] Abbas A. Rizvi, Ezgi Karaesmen, Martin Morgan, Leah Preus, Junke Wang, Michael Sovic, Theresa Hahn, and Lara E. Sucheston-Campbell, *gwasurvivr: an R package for genome wide survival analysis*, Bioinformatics (2018).

[2] Mark D. Long, Prashant K. Singh, James R. Russell, Gerard Llimos, Spencer Rosario, Abbas A. Rizvi, Patrick R. van den Berg, Jason Kirk, Lara E. Sucheston-Campbell, Dominic J. Smiraglia, and Moray Campbell, *The miR-96 and RARγ signaling axis governs androgen signaling and prostate cancer progression*, Oncogene (2018).

[3] Lara E. Sucheston-Campbell, Alyssa I. Clay-Gilmour, William E. Barlow, G. Thomas Budd, Daniel O. Stram, Christopher A. Haiman, Xin Sheng, Li Yan, Gary Zirpoli, Song Yao, Chen Jiang, Owzar Kouros, Dawn Hershman, Kathy S. Albain, Daniel F. Hayes, Halle C. Moore, Timothy J. Hobday, James A. Stewart, Abbas A. Rizvi, Claudine Isaacs, Muhammad Salim, Jule R. Gralow, Gabriel N. Hortobagyi, Robert B. Livingston, Deanna L. Kroetz, and Christine B. Ambrosone, *Genome-wide meta-analyses identifies novel taxane-induced peripheral neuropathy-associated loci*, Pharmacogenetics and Genomics **28** (2018), no. 2, 49-55.

[4] Ezgi Karaesmen, Abbas A. Rizvi, Leah M. Preus, Philip L. McCarthy, Marcelo C. Pasquini, Kenan Onel, Xiaochun Zhu, Stephen Spellman, Christopher A. Haiman, Daniel O. Stram, Loreall Pooler, Xin Sheng, Qianqian Zhu, Li Yan, Qian Liu, Qiang Hu, Amy Webb, Guy Brock, Alyssa I. Clay-Gilmour, Sebastiano Battaglia, David Tritchler, Song Liu, Theresa Hahn, and Lara E. Sucheston-Campbell, *Replication and validation of genetic polymorphisms as-*

sociated with survival after allogeneic blood or marrow transplant, Blood **130** (2017), no. 13, 1585-1596.

[5] Alyssa I. Clay-Gilmour, Theresa Hahn, Leah M. Preus, Kenan Onel, Andrew Skol, Eric Hungate, Qianqian Zhu, Christopher A. Haiman, Daniel O. Stram, Loreall Pooler, Xin Sheng, Li Yan, Qian Liu, Qiang Hu, Song Liu, Sebastiano Battaglia, Xiaochun Zhu, AnneMarie W. Block, Sheila N. J. Sait, Ezgi Karaesmen, Abbas A. Rizvi, Daniel J. Weisdorf, Christine B. Ambrosone, David Tritchler, Eva Ellinghaus, David Ellinghaus, Martin Stanulla, Jacqueline Clavel, Laurent Orsi, Stephen Spellman, Marcelo C. Pasquini, Philip L. McCarthy, and Lara E. Sucheston-Campbell, *Genetic association with B-cell acute lymphoblastic leukemia in allogeneic transplant patients differs by age and sex*, Blood Advances **1** (2017), no. 20, 1717-1728.

[6] Abbas A. Rizvi, *Reprogramming androgen receptor and lysine-specific demethylase 1 transcriptome in castration-resistant prostate cancer*, University at Buffalo, 2015.

[7] Maxwell S. DeNies, Jordan Johnson, Amanda B. Maliphol, Michael Bruno, Annabelle Kim, Abbas A. Rizvi, Kevyn Rustici, and Scott Medler, *Diet-induced obesity alters skeletal muscle fiber types of male but not female mice*, Physiological Reports **2** (2014), no. 1, e00204.

Fields of Study

Major Field: Pharmaceutical Sciences

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

Broadly, this dissertation examines germline genetic variation in the context matched unrelated donor (MUD) hematopoetic stem cell transplantation (HSCT). This dissertation seeks to identify and characterize genetic variants in patients who have acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), or myelodysplastic syndrome (MDS) and received an HSCT from an HLA matched-unrelated donor (MUD). This dissertation also seeks to enhance the computational workflows that are used in these fields. The significance is three-fold, first, we can identify clinically relevant markers that may improve donor selection beyond traditional methods; second, we can characterize pre-transplant risk of disease or transplant related death within the first year; and third, we can help facilitate other researchers studying similar problems with similar data by developing open-source software.

The project that this dissertation contributes to has the opportunity to be a real life example of a translational study (from computational analysis of biological data to bedside). The dissertation is broken down into 6 chapters. This chapter (Chapter 1) first introduces genetic association studies and important related concepts (genetic and statistical) that are needed to appreciate and understand the underlying analysis. Chapter 1 also introduces allogeneic HSCT and DISCOVeRY-BMT genome wide-association study (GWAS), including the corresponding clinical data in detail. Chapter 2 will discuss a replication and validation study of all previous literature that examined genetic variation in the same context as our study, which we published on in *Blood*. Chapter 3 discusses the R package that we developed and it details the development and testing procedures executed. The package

is available on R/Bioconductor and was published in *Bioinformatics*. Chapter 4 is the application of the R package and custom pipeline developed to perform large automated GWAS, specialized to our lab, but can be generalized to other large scale projects. Chapter 5 discusses the discovery and inference of markers in ALL donor and recipient pairs. This dissertation ends with Chapter 6, which comprises preliminary data on genetic contributions to early death after transplant (the first 100 days) and the future directions that should be undertaken.

Genetic Association Studies

Genetic association studies test for correlations between genetic variation as it relates to disease risk or to physical quantitative traits (i.e. height or weight) (C. M. Lewis and Knight 2012). These studies have been successful in identifying certain variants as being predictive of disease susceptibility or drug response and have helped us understand that many diseases have complex genetic signatures that need to be further understood (Visscher et al. 2012). The human genome consists of over 3 billion base pairs, all of which are contained in every nucleated cell in the body. A genome sequence is the complete collection of all nucleotides (A, C, T, or G for DNA genomes) that make up all the chromosomes in individuals or species (Lander et al. 2001). The vast majority of nucleotides (>99.5%) are identical between individuals within a species, however, genetic variation arises within individuals and populations over time and different spaces. Indeed, the fundamental source of genetic variation is mutation, where permanent alterations occur to a single nucleotoide or larger structural changes in the genome of a species. A mutation at single position (locus) in a DNA sequence that occurs in at least 1% of a population is called a single nucleotide polymorphism (SNP). SNPs are the most widely used marker to describe genetic variation. Larger structural variations may include mi-

crosatellite regions, insertion/deletions (indels), copy number variations (CNVs), or variable-tandem repeats (VNTRs) all of which have importance in understanding the genetic architecture and disease etiology (Timpson et al. 2018). For purposes of this document, unless otherwise specified, genetic variation will refer to single nucleotide polymorphisms (SNPs). SNPs will also be referred to as simply as polymorphisms, genetic markers, or markers interchangeably.

Different forms of the same variant are called *alleles*. For diploid organisms, one allele is passed from each parent. When the same variation is passed both parents it is called homozygous and when they are different it is known as heterozygous. When mutations are passed between generations they are known as *germline mutations*. Importantly, in humans (and other eukaroytes), genetic recombination occurs during meiosis, where large chunks of genetic materials are exchanged and shuffled between parents and their offspring. A particular combination of alleles that lie on the same chromosome are called *haplotypes*.

While genetic variation only occurs in 0.1% to 0.5% of the human genome, modern genomic tools have revealed even in this small proportion of the genome, the underlying architecture is very complex. Sequence variations can occur coding regions of genes, non-coding regions of genes, or intergenic regions. SNPs that are within a coding regions are of two types: synonymous and non-synonymous SNPs. Synonymous SNPs do not change amino acid sequences and therefore do not change protein structure, while non-synonmyous SNPs change amino acid sequences. Non-synonymous SNPs are further stratified into two types: missense and nonsense polymorphisms. Missense SNPs result in codon changes that code for different amino acids (Shi and Moult 2011). Nonsense SNPs are genetic alterations that yield a premature stop codon that often yield a non-functional or truncated protein product. Polymorphisms that are in non-coding regions may alter impor-

tant transcriptional properties such as gene splicing, transcription factor binding, or messenger RNA (mRNA) decay (Green et al. 2003). SNPs may affect gene expression (and may be upstream or downstream from a gene) are called *expression quantiative trait loci* (eQTL).

Historically, gene mapping studies were used to determine associations between genomic DNA sequence variations and phenotypic variability (Visscher et al. 2012). These studies were quite successful, particularly in Mendelian traits (e.g. single gene disorders) (Botstein and Risch 2003). Over the past two decades, research has increasingly evolved from looking at specific regions of interest (candidate gene association studies) to more agnostic approaches that investigate larger portions of the genome, such as genome wide association studies (GWAS), whole-exome sequencing studies (WES studies) and whole-genome sequencing studies (WGS studies) (Timpson et al. 2018). This dissertation is primarily focused on GWAS, specifically the application of using GWAS in the context of identifying common variants in after hematopoetic stem cell transplantation (HSCT), where more details will be discussed in subsequent subsections of this chapter.

GWAS employ genotyping microarrays to measure genetic variation – and they have become the standard platform in academic and industry to test for association of phenotype with common genetic variants. Common genetic variants are defined as those with a *minor allele frequency* (MAF) of $\geq 1\%$ and rare variants are defined as those with a MAF $\leq$ than 1%. Genotyping microarrays are designed to contain common variants but optionally can contain rare variants. GWAS ask if the allele of a genetic variant is found more often than what would be expected than by random chance in individuals with the phenotype of interest (e.g. the disease being studied). If the variant (one allele) occurs more in those affected by the disease than those without the disease, then the variant deemed as being *associated* with

the disease.

Linkage Disequilibrium

GWASs are heavily based on the principal of linkage disequilbrium (LD) at the population level. LD is the non-random dependence of allele frequencies at two more loci in the general population (Jorde 2000). LD reflects the relationship between alleles at different loci. In other words, LD is a measure of two alleles or specific sequences being inherited together. The unit of measure for LD is r^2 (squared correlation coefficient) (Pritchard and Przeworski 2001). In general, loci that are in close proximity exhibit stronger LD (Pritchard and Przeworski 2001). Perfect LD ($r^2 = 1$) means that no recombination occurred on this chunk of genome. Regions that are further apart on a choromsome exhibit weaker LD (D. E. Reich et al. 2001). Low LD means that recombination occurred and that there are lots of possible rearrangements that may have occurred during meiosis. LD decay influences the number of SNPs needed to "tag" a haplotype, and that number of SNPs is just a small subset of the number of segregating polymorphisms in the population (D. E. Reich et al. 2001).

With the rapid growth of genetic association studies and statistical methods assessing genetic variation, researchers routinely exploit LD to map regions in the human genome. While costs have lowered over the past decade, a major barrier to overcome when conducting large scale studies, was the expense of searching the entire genome for disease associations (International HapMap Consortium 2005). Whole-genome sequencing (WGS) would have had to been performed on numerous patients to identify variants, where the cost would have been unfeasable.

The first major large scale multi-institutional international project to finely map genetic variation was the HapMap project (International HapMap Consortium

2005). HapMap developed a haplotype map of the human genome to finely describe patterns of human genetic variation. In fact, HapMap demonstrated that genomic blocks are shared in common areas across contintental populations.

Since, 1000 Genomes Project (Consortium 2015) has

Genetic Imputation

Genetic imputation (will also be simply referred to as imputation) has had significant contributions to genetic assocation studies. Imputation can be defined as predicting unobserved genotypes that were not directly assayed in a sample of individuals. The term refers to when a reference panel of haplotypes at set of a SNPs is used to impute SNPs that have been genotyped at a subset of that set of SNPs. Today, several reference panels are available, such as (but not limited to) HapMap2, 1000 Genomes Phase 3, Haplotype Reference Consortsium (HRC), which differ by the number of samples, sites (chromosomes 1-22, X), and number of haplotypes.

Knowledge of haplotype structure makes it possible to retreive more information from GWAS. Tagging SNPs with known haplotype block structure can capture much of the genetic information in a region.

Imputation can be performed across the entire genome (GWAS) or at targeted regions as part of a fine-mapping study.

Phasing (or haplotype estimation)

IMPUTE2. Sanger Imputation Server. Michigan Imputation Server

Hematopoetic Stem Cell Transplantation

HSCT is an established therapeutic procedure that is used as a potentially curative treatment for life-threatening congenital or acquired blood disorders (malignant or non-malignant) (Henig and Zuckerman 2014).

HSCT involves the intravenous infusion of autologous or allogeneic hematopoietic progenitor cells to restore normal function in patients whose bone marrow is compromised. Autologous HSCT involves self-donation of marrow stem cells, whereas allogeneic HSCT is when stem cells are transferred from a HLA-matched related donor (MRD) or a HLA-matched unrelated donor (MUD).

Although a matched sibling donor is preferred, only approximately 30% of patients who may benefit from HSCT have such a donor available. In the United States, the number of allogeneic transplants yearly has dramatically risen over the past decade, across all diseases (M. Pasquini et al. 2013).

Patients with acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), or myelodysplastic syndrome (MDS), represent the largest group treated with allogeneic HSCT.

While both patient care and matching has improved over the past few decades almost half of all high-resolution 10/10 HLA MUD-HSCT recipients die within one-year post-transplant due to either their disease or transplant-related causes (M. Pasquini et al. 2013).

These trends also show transplant-related causes are a larger contributor to mortality within the first 100-days post-transplant and shift towards primary disease after approximately six months post-transplant (D'Souza et al. 2017).

Reducing TRM without increasing risk of disease death and vice versa continue to represent a substantial clinical challenge.

The four elements of HSCT are:

1. Graft Source
2. Graft Type
3. HLA matching
4. Pre-transplant conditioning regimens

The graft sources are either from bone marrow or peripheral blood stem cells (PBSCs). For bone marrow grafts, hematopoetic stem cells (HSCs) are typically extracted from the center of the iliac crest (pelvis) using a large needle. The

DISCOVeRY-BMT

Determining the Influence of Susceptibility COnveying Variants Related to one-Year mortality after Bone Marrow Transplant (DISCOVeRY-BMT) is a GWAS. This GWAS aims at identifying and characterizing non-human leukocyte antigen (HLA) genetic variation in the context of survival outcomes in acute lymphoblastic leukemia (ALL), acute myelogenous leukemia (AML) and myelodysplastic syndrome (MDS) patients and their matched unrelated donor (MUD) after hematopoietic stem cell transplantation (HSCT). HSCT is also called bone marrow transplantation (BMT) and hence both terms will be used interchangeably throughout this document. Additionally, the terms recipients and patients will be used interchangeably.

Patient Characteristics

DISCOVeRY-BMT investigates donor and recipient genetic factors that contribute to 1 year cause-specific mortality after MUD-HSCT (Hahn et al. 2015, Clay-Gilmour et al. (2017)). This GWAS comprises two independent cohorts. All patients in DISCOVeRY-BMT were reported to the Center for International Blood and Marrow Research (CIBMTR) and had banked biorepository samples (blood samples) from recipients and donors at the National Marrow Donor Program (NMDP). Data that were reported to CIBMTR was collected from 151 transplant centers in the USA. Furthermore, all patients were de-identified and approved by the Roswell Park Cancer Institute Institutional Review Board and signed informed

consent for research. Although no patients were excluded based on gender, age, or race, over 90% of DISCOVeRY-BMT cohorts were of self-reported European American ancestry. Additional exclusion criteria included: umbilical cord blood graft, *ex vivo* T-cell depleted graft or the patient underwent a prior autologous or allogeneic transplant.

Blood samples were genotyped and comprise the genotype data is the "genetic data" that is reported throughout this dissertation. In preparation for this GWAS, all post-mortem cause-specific deaths of the patients were reviewed and judged by a post-mortem by an expert panel (Hahn et al. 2015). The review and adjudicated of cause of death was conducted because many of the 151 centers the data are inconsistent in reporting and classification of outcome attributions. Thus the panel was established for concordance in cause of death, logic, and quality. Re-assessment by the panel was performed using autopsy reports and death certificates. The panel consisted of three HSCT oncologists and a HSCT clinical epidemiologist (Hahn et al. 2015).

Both cohorts comprise AML, ALL, or MDS patients who were T-cell replete and treated with myeloablative (MA) or reduced intensity (RIC) conditioning regimens prior to transplant. Cohort 1 included 2110 HSCT recipients and 2052 10/10 HLA matched unrelated donors (matched at HLA-A, -B, -C, -DRB1, and -DQB1) from 2000 to 2008 (**Table 1.1**). Cohort 2 included included 763 donors and 777 recipients from the years 2000 to 2011 (**Table 1.1**). A subset of patients (n=281) in cohort 2 were 8/8 HLA matched (matched at HLA-A, -B, -C, -DRB1), while the remaining patients (n=496) had the same matching criteria as Cohort 1 and are 10/10 HLA matched.

The proportion of patients with AML is relatively consistent between cohorts at approximately 60% of patients (**Table 1.2**). However, the proportion of ALL

Table 1.1: Distribution of Donors and Recipients in DISCOVeRY-BMT Cohorts.

Genome	Cohort 1: N (Percent %)	Cohort 2: N (Percent %)
Donor	2052 (72.9%)	763 (27.1%)
Recipient	2110 (73.1%)	777 (26.9%)

N denotes sample size. The percentages represent the proportion of donors or recipients in either cohort 1 or 2. The sample sizes reported here are based on completeness of follow up time intervals and event data.

Table 1.2: Donor and Recipients Disease Proportions by Cohort in DISCOVeRY-BMT.

	Donor (N/Percent %)		Recipient (N/Percent %)	
	Donor Cohort 1	Donor Cohort 2	Recipient Cohort 1	Recipient Cohort 2
ALL	468 (23%)	93 (12%)	483 (23%)	94 (12%)
AML	1247 (61%)	478 (63%)	1282 (61%)	488 (63%)
MDS	337 (16%)	192 (25%)	345 (16%)	195 (25%)

Shown here are disease proportions of ALL, AML and MDS in DISCOVeRY-BMT cohorts. The percentage in each cell is computed column-wise as the proportion of sample size (N) within a cohort across disease group.

and MDS patients shifts between cohorts. For ALL, in cohort 1 the proportion is 23%, while in cohort 2, the proportion is 12% (**Table 1.2**). Similarly, for MDS, in cohort 1 the proportion of patients with this disease is 16% and in cohort 2 it is 25% (**Table 1.2**).

Why do the ALL proportions go down in cohort 2

Note: Inquire about changes in MDS definition between cohorts

Survival Outcome Definitions:

The survival outcomes are cause-specific deaths during the first year post-HSCT that were reviewed and re-assessed with high confidence (described above). The primary outcomes included are: disease related mortality (DRM), and transplant related mortality (TRM). The secondary outcomes were: overall survival

(OS), progression-free survival (PFS), and relapse (REL).

TRM subtypes were further stratified into graft-versus-host disease (GVHD), organ failure (OF), infection (INF), or other. Other causes of death included more rare events, such as (but not limited to) secondary malignancies, primary or secondary graft failure, or hemorrhage.

GVHD deaths included acute or chronic GVHD, conditional upon a patient actively receiving treatment for GVHD at the time of death.

Deaths that were considered to be infection were identified as arising from bacterial, fungal, viral, and/or protozoan organisms that caused organ damage. OF deaths were defined as organ toxicity relating to the transplant and not due to disease progression, GVHD, or infection.

The cohorts had two primary outcomes:

Disease Related Mortality (DRM) – broadly defined as deaths relating to leukemia/MDS relapse/progression, including death attributed to toxicity or infection from anti-leukemic treatments post-HSCT.

Transplant Related Mortality (TRM) – defined as any cause of death except the underlying disease, pre-existing disease, accidental death or suicide, or death unrelated to the transplant.

Secondary outcomes:

Overall Survival (OS) – defined as any patient (recipient) that died at any point within the first 12 month window post-HSCT of this observational study.

Progression Free Survival (PFS) is defined as the time to relapse. All patients were analyzed as time to progression of disease

patients who were in complete remission (CR) before HSCT had documented disease progression after HSCT and succumb to the disease.

Relapse (REL) – patients who were not in CR pre-HSCT and the disease

Table 1.3: Mixed Disease: Proportion of Events by Outcome and Cohort in Donor and Recipients

	Donor (N/Percent %)		Recipient (N/Percent %)	
	Donor Cohort 1	Donor Cohort 2	Recipient Cohort 1	Recipient Cohort 2
OS	861 (41.96%)	303 (39.71%)	879 (41.66%)	309 (39.77%)
DRM	464 (22.61%)	163 (21.36%)	474 (22.46%)	168 (21.62%)
PFS	1031 (50.24%)	368 (48.23%)	1055 (50%)	374 (48.13%)
REL	623 (30.36%)	228 (29.88%)	639 (30.28%)	233 (29.99%)
TRM	397 (19.35%)	140 (18.35%)	405 (19.19%)	141 (18.15%)
GVHD	132 (6.43%)	60 (7.86%)	134 (6.35%)	59 (7.59%)
INF	118 (5.75%)	34 (4.46%)	122 (5.78%)	36 (4.63%)
OF	102 (4.97%)	26 (3.41%)	104 (4.93%)	26 (3.35%)

Shown here are the proportion of events for mixed disease (ALL, AML, and MDS) in both DISCOVeRY-BMT cohorts. The survival outcomes are overall survival (OS), progression free survival (PFS), disease related mortality (DRM), transplant related mortality (TRM), graft-versus-host disease (GVHD), infection (INF), or organ failure (OF).

returns (relapse) after HSCT.

Proportions of Events Mixed disease with outcomes

Explain that discrepancies between TRM and TRM subtypes are because some patients adjudication was not well differentiated between these subtypes.

Proportion of Events AML/MDS with outcomes

Proportions of Events in AMLonly with outcomes

Proportions of Events in ALLonly with outcomes

Proportions of Events in MDSonly with outcomes

Survival Curves for One year Outcomes

Genotyping and Quality Control

All samples were genotyped on a Illumina Human OmniExpress BeadChip whole-genome genotyping microarray. This chip has approximately 750,000 tagged

Figure 1.1: Histogram of Donor and Recipient Age Distributions in Mixed Disease (AML, ALL, and MDS)

Table 1.4: AML + MDS: Proportion of Events by Outcome and Cohort in Donor and Recipients

	Donor (N/Percent %)		Recipient (N/Percent %)	
	Donor Cohort 1	Donor Cohort 2	Recipient Cohort 1	Recipient Cohort 2
OS	662 (41.79%)	264 (39.4%)	680 (41.79%)	270 (39.53%)
DRM	373 (23.55%)	144 (21.49%)	383 (23.54%)	149 (21.82%)
PFS	790 (49.87%)	323 (48.21%)	810 (49.78%)	329 (48.17%)
REL	494 (31.19%)	203 (30.3%)	506 (31.1%)	208 (30.45%)
TRM	289 (18.24%)	120 (17.91%)	297 (18.25%)	121 (17.72%)
GVHD	95 (6%)	56 (8.36%)	97 (5.96%)	55 (8.05%)
INF	86 (5.43%)	25 (3.73%)	90 (5.53%)	27 (3.95%)
OF	74 (4.67%)	21 (3.13%)	76 (4.67%)	21 (3.07%)

Shown here are the proprtion of events for AML+MDS in both DISCOVeRY-BMT cohorts. The survival outcomes are overall survival (OS), progression free survival (PFS), disease related mortality (DRM), transplant related mortality (TRM), graft-versus-host disease (GVHD), infection (INF), or organ failure (OF).

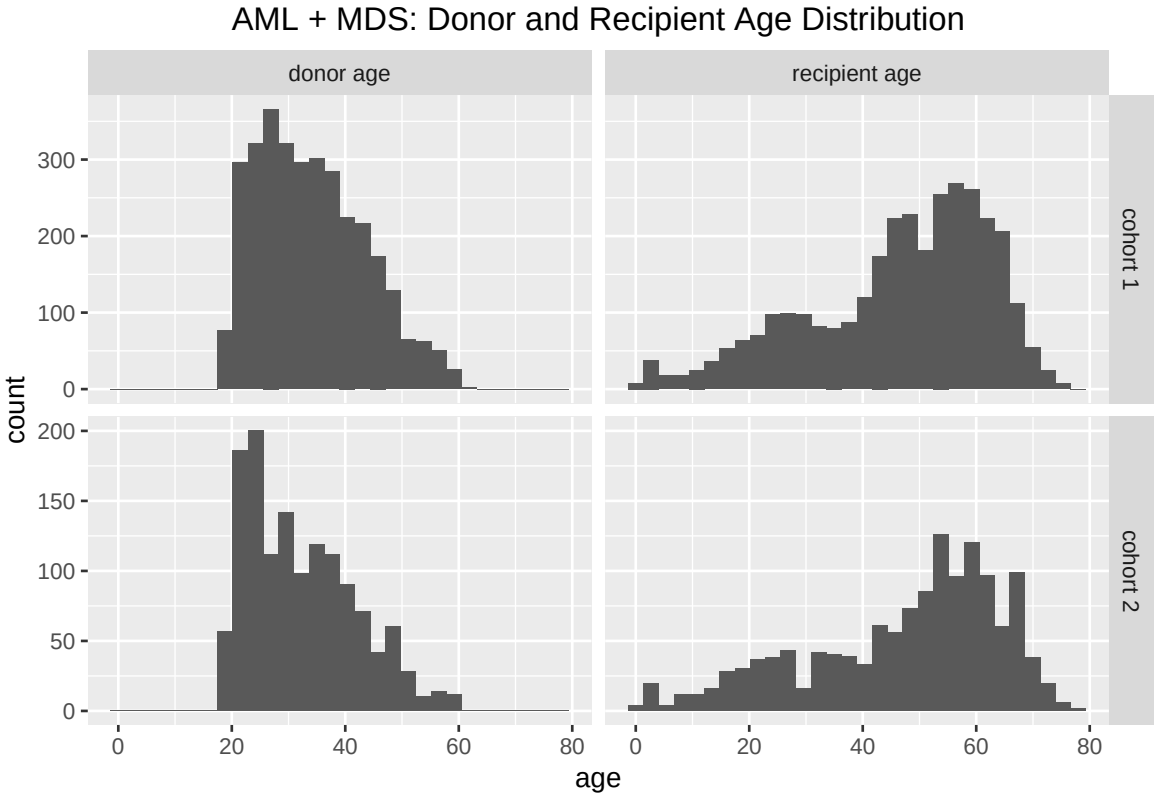

Figure 1.2: Histogram of Donor and Recipient Age Distributions in AML and MDS patients

Table 1.5: AML only: Proportion of Events by Outcome and Cohort in Donor and Recipients

	Donor (N/Percent %)		Recipient (N/Percent %)	
	Donor Cohort 1	Donor Cohort 2	Recipient Cohort 1	Recipient Cohort 2
OS	532 (42.66%)	184 (38.49%)	544 (42.43%)	188 (38.52%)
DRM	324 (25.98%)	111 (23.22%)	333 (25.98%)	115 (23.57%)
PFS	634 (50.84%)	230 (48.12%)	648 (50.55%)	234 (47.95%)
REL	420 (33.68%)	157 (32.85%)	431 (33.62%)	161 (32.99%)
TRM	208 (16.68%)	73 (15.27%)	211 (16.46%)	73 (14.96%)
GVHD	61 (4.89%)	26 (5.44%)	61 (4.76%)	26 (5.33%)
INF	66 (5.29%)	14 (2.93%)	69 (5.38%)	14 (2.87%)
OF	59 (4.73%)	18 (3.77%)	59 (4.6%)	18 (3.69%)

Shown here are the proprtion of events for AML only in both DISCOVeRY-BMT cohorts. The survival outcomes are overall survival (OS), progression free survival (PFS), disease related mortality (DRM), transplant related mortality (TRM), graft-versus-host disease (GVHD), infection (INF), or organ failure (OF).

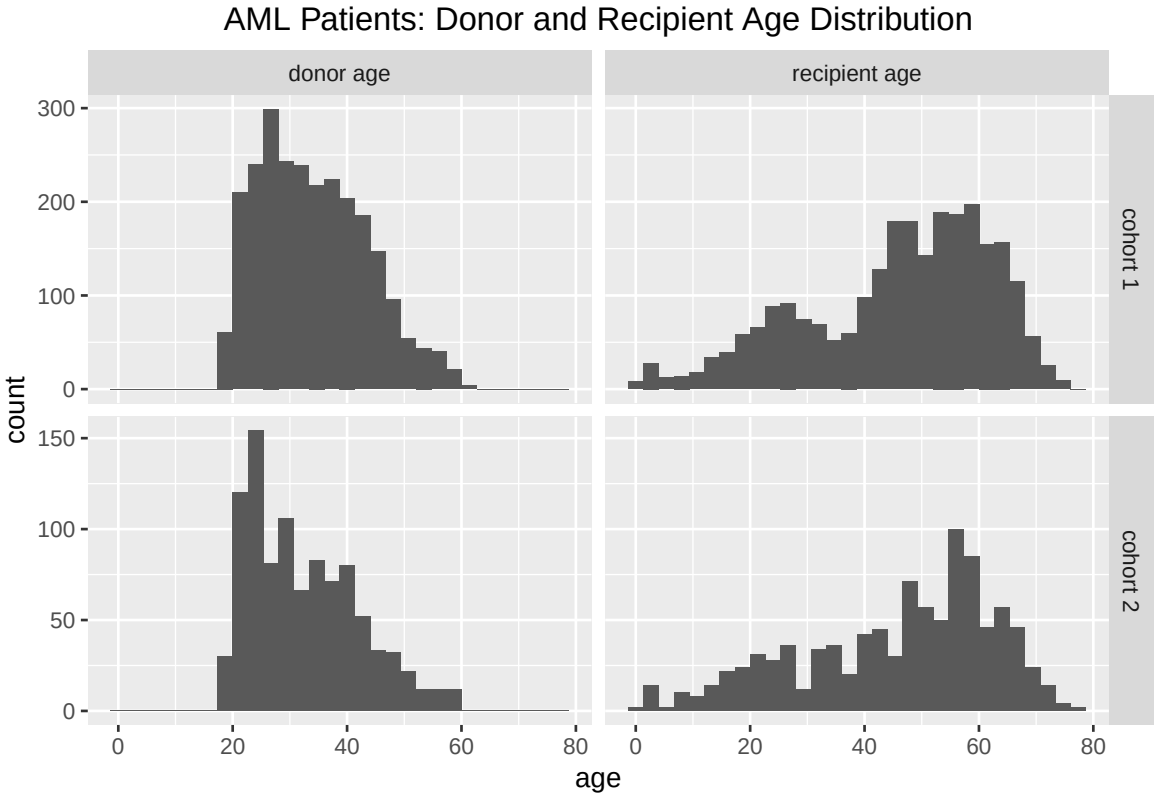

Figure 1.3: Histogram of Donor and Recipient Age Distribution in AML and MDS patients

Table 1.6: ALL only: Proportion of Events by Outcome and Cohort in Donor and Recipients

	Donor (N/Percent %)		Recipient (N/Percent %)	
	Donor Cohort 1	Donor Cohort 2	Recipient Cohort 1	Recipient Cohort 2
OS	199 (42.52%)	39 (41.94%)	199 (41.2%)	39 (41.49%)
DRM	91 (19.44%)	19 (20.43%)	91 (18.84%)	19 (20.21%)
PFS	241 (51.5%)	45 (48.39%)	245 (50.72%)	45 (47.87%)
REL	129 (27.56%)	25 (26.88%)	133 (27.54%)	25 (26.6%)
TRM	108 (23.08%)	20 (21.51%)	108 (22.36%)	20 (21.28%)
GVHD	37 (7.91%)	4 (4.3%)	37 (7.66%)	4 (4.26%)
INF	32 (6.84%)	9 (9.68%)	32 (6.63%)	9 (9.57%)
OF	28 (5.98%)	5 (5.38%)	28 (5.8%)	5 (5.32%)

Shown here are the proprtion of events for ALL only in both DISCOVeRY-BMT cohorts. The survival outcomes are overall survival (OS), progression free survival (PFS), disease related mortality (DRM), transplant related mortality (TRM), graft-versus-host disease (GVHD), infection (INF), or organ failure (OF).

Figure 1.4: Histogram of Donor and Recipient Age Distributions in ALL patients

Table 1.7: MDS only: Proportion of Events by Outcome and Cohort in Donor and Recipients

	Donor (N/Percent %)		Recipient (N/Percent %)	
	Donor Cohort 1	Donor Cohort 2	Recipient Cohort 1	Recipient Cohort 2
OS	130 (38.58%)	80 (41.67%)	136 (39.42%)	82 (42.05%)
DRM	49 (14.54%)	33 (17.19%)	50 (14.49%)	34 (17.44%)
PFS	156 (46.29%)	93 (48.44%)	162 (46.96%)	95 (48.72%)
REL	74 (21.96%)	46 (23.96%)	75 (21.74%)	47 (24.1%)
TRM	81 (24.04%)	47 (24.48%)	86 (24.93%)	48 (24.62%)
GVHD	34 (10.09%)	30 (15.62%)	36 (10.43%)	29 (14.87%)
INF	20 (5.93%)	11 (5.73%)	21 (6.09%)	13 (6.67%)
OF	15 (4.45%)	3 (1.56%)	17 (4.93%)	3 (1.54%)

Shown here are the proprtion of events for MDS only in both DISCOVeRY-BMT cohorts. The survival outcomes are overall survival (OS), progression free survival (PFS), disease related mortality (DRM), transplant related mortality (TRM), graft-versus-host disease (GVHD), infection (INF), or organ failure (OF).

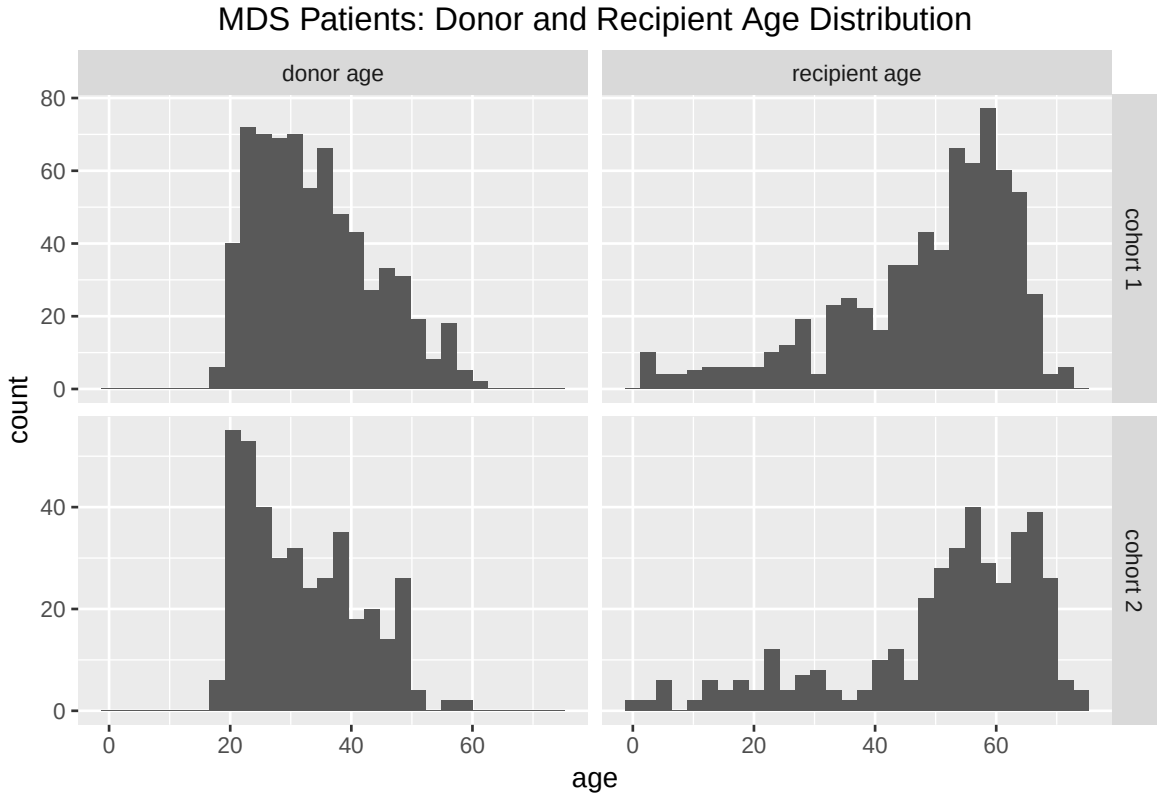

Figure 1.5: Histogram of Donor and Recipient Age Distributions in MDS patients

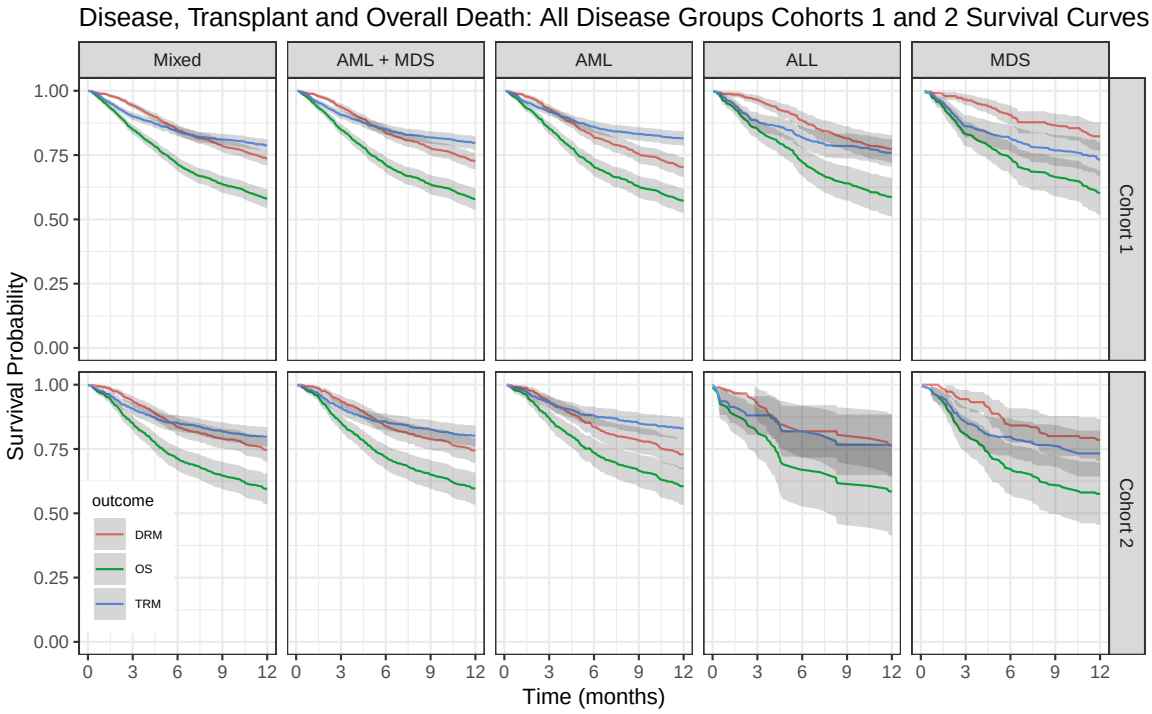

Figure 1.6: Survival curves for all disease groups being tested (Mixed disease, AML+MDS, AML only, and ALL only). The x-axis is survival probability. The y-axis is time in months. Cohorts 1 and 2 are shown in the left and right panel respectively. Red is disease-related mortality (DRM), green is overall survival (OS), and blue is transplant related mortality (TRM). The gray shaded areas are 95 percent confidence intervals.

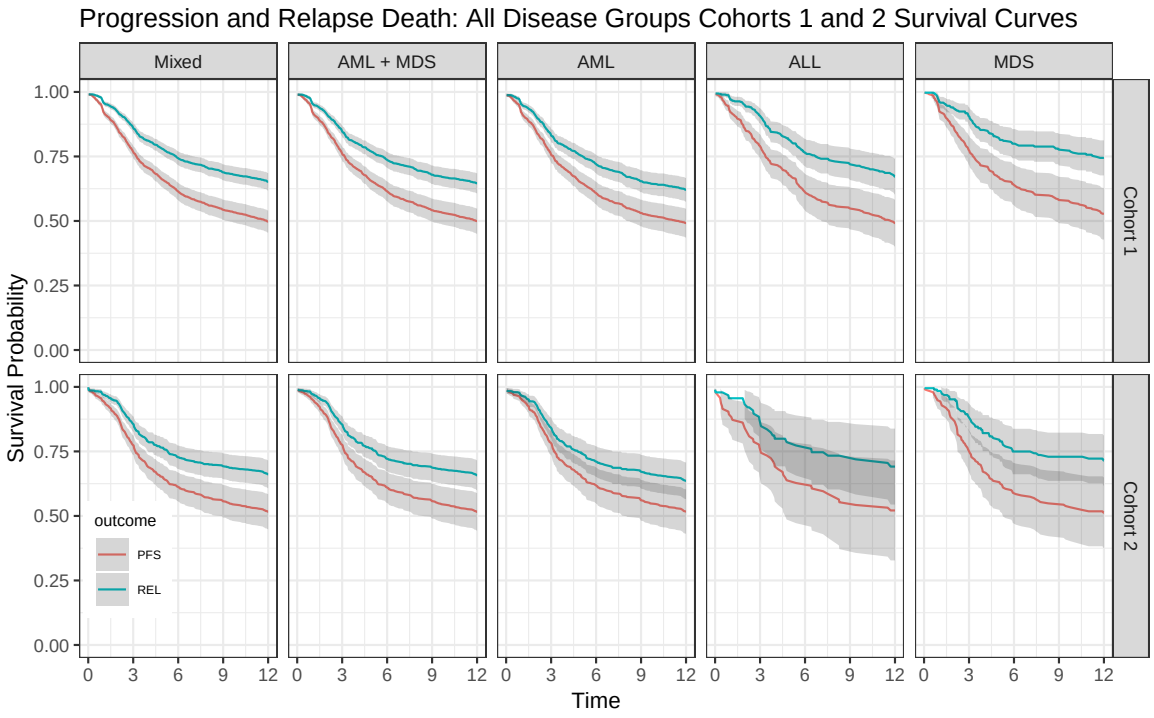

Figure 1.7: Survival curves for all disease groups being tested (Mixed disease, AML+MDS, AML only, and ALL only). Cohorts 1 and 2 are shown in the left and right panel respectively. Red is progression free survival (PFS) and teal is relapse (REL). The gray shaded areas are 95 percent confidence intervals.

SNPs that were strategically selected from all three phases of the International HapMap project to capture the greatest amount of common SNP variation (>5% MAF).

Samples were assigned to plates to ensure the even distribution of patient characteristics and potential confounding variables using Optimal Sample Assignment Tool (OSAT), an R/Bioconductor software package (Yan et al. 2012).

Over 90% of DISCOVeRY-BMT patients self-reported as European American, Caucasian or White and thus replication and validation analyses are performed on these recipient-donor pairs.

An important concept in statistical genetics is *population stratification* **Define population stratification**

Stringent quality control was performed on both samples and SNPs within this population. Population outliers were removed using EIGENSTRAT (Price et al. 2006) (n=73).

Additional sample quality control removed samples with missing call rate 2% (n=54), sex mismatch (n=9), abnormal inbreeding coefficients (n=20), and evidence of cryptic relatedness (n=17), yielding 2106 and 777 donor-recipient pairs in cohorts 1 and 2, respectively. Typed SNPs were removed if the call rate was <98%, there was deviation from Hardy-Weinberg equilibrium proportions or discordance between duplicate samples was >2%.

In total 637,655 and 632,823 SNPs from the OmniExpress BeadChip were available for imputation in cohorts 1 and cohort 2, respectively, using 1000 Genomes Project Phase 3 (Consortium 2015).

IMPUTE2 (B. N. Howie, Donnelly, and Marchini 2009) software was used for Imputation and QCTOOL was used to remove imputed genotypes with info score <0.7, certainty <0.7 and a minor allele frequency <0.005. The recipient-donor mis-

match genome dosage calculations, described above, were done as the absolute value of the recipient minus donor minor allele dosages. Rs2066847 (SNP13) in NOD2/CARD15 was the only variant analyzed from the Illumina HumanExome as it was not typed on the OmniExpress chip or available following imputation.

Statistical Analysis

Cox Proportional Hazards Model

Survival models examine the time it takes for events to occur. Specifically, survival models examine the relationship between survival (time that passes before some event occurs) and one or more *covariates* (predictors) that may be associated with that quantity of time. The Cox proportional hazards regression model (D. R. Cox 1972) is used for survival analysis and is our main statistical model of choice.

Mathematical concepts and notations

The Cox model is as follows:

Consider a population of subjects, i, we observe either time to event or censoring. For the censored individuals, we know that time to event is greater than censoring time. So the survival function is S(t)

Let T represent survival time.

T is a random variable:

Cumulative distribution function (CDF):

$$P(t) = Pr(T \leq t) \tag{1.1}$$

Probability density function (PDF):

$$p(t) = \frac{dP(t)}{dt} \tag{1.2}$$

The survival function $S(t)$ is the complement of the CDF:

$$S(t) = Pr(T \geq t) = 1 - P(t) \tag{1.3}$$

and the hazara function is $\lambda(t)$ (or age specific failure rate)

The hazard function, $\lambda(t)$, is the distribution of survival times, which assesses the instantaneous risk of dying at time t, conditional on survival to that time:

** make a sentence about assuming no ties **

** baseline hazard is nuissance paramter and is completely removed**

covariate information on different items easily incorporated

** data censoring patterns often encountered in life tests (such as those in senior citizens studies) do not affect the partial likelihood equation **

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0} \frac{Pr\Big[t \leq T < t + \Delta t) | t \geq T\Big]}{\Delta t} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{1.4}$$

The Cox model:

Let $X_i = X_{i1}, ..., X_{ip}$ be realized values of covariates for subject i.

The hazard function for the Cox model has the form:

$$\begin{aligned} \lambda(t|X_i) &= \lambda_0(t) \cdot \exp(\beta_1 X_{i1} + ... + \beta_p X_{ip}) \\ &= \lambda_0(t) \cdot \exp(X_i^T \cdot \beta) \end{aligned} \tag{1.5}$$

This expression gives us the hazard function at time t for subject i with covariate vector X_i.

The probability of the event to be observed occuring with subject i at time Y_i can be written as:

$$\begin{aligned} L_i(\beta) &= \frac{\lambda(Y_i|X_i)}{\sum_{j:Y_j \geq Y_i} \lambda(Y_i|X_j)} \\ &= \frac{\lambda_0(Y_i)\theta_i}{\sum_{j:Y_j \geq Y_i} \lambda_0(Y_i)\theta_j} \\ &= \frac{\theta_i}{\sum_{j:Y_j \geq Y_i} \theta_j} \end{aligned} \tag{1.6}$$

We describe how model equations already developed to express the effects of exposure on disease rates calculated from grouped data are adapted to the continuous case. Section 5.2 introduces the 'partial likelihood' methodology for estimating regression coefficients in models in which the exposure variables are assumed to act multiplicatively on the background rates.

** Don't forget to mention hazard ratio computation **

Feature Selection

Prior to genetic analyses, clinical covariates for inclusion in genome-wide survival models were selected using bidirectional stepwise regression (Venables and Ripley 2002) on Cox proportional hazard models (D. R. Cox 1972) of OS, PFS, TRM and DRM using R statistical software (R Core Team 2018). Cox proportional hazard models of OS, TRM and DRM evaluated SNPs associated with time to death with all survivors censored at 1 year post-BMT. PFS was defined as the time to disease progression or death (Therneau and Grambsch 2013).

Deaths from TRM and DRM were treated as competing risks and analyzed

accordingly (Fine and Gray 1999). SNP models for OS adjusted for recipient age, disease status (early/intermediate or advanced), and graft source (blood or marrow); PFS and DRM SNP models adjusted for recipient age and disease status; TRM SNP models adjusted for recipient age, graft source and body mass index (underweight/normal, overweight, or obese). Dosage data accounting for the probability of each genotype were used in all analyses of imputed data. Effect size estimates and standard errors from DISCOVeRY-BMT Cohorts 1 and 2 were compared and combined using a fixed-effects inverse variance meta-analyses in METAL (Willer, Li, and Abecasis 2010). For each SNP, heterogeneity of effect size estimates between cohorts 1 and 2 was assessed using p-values from significance tests of heterogeneity (P_{het}) and I^2. Variants with $P_{het} < 0.05$ and $I^2 > 50$ were meta-analyzed with a random effects models using meta in R (Schwarzer 2007).

Covariates Included In Models

1. OS – age, disease status, graft source
2. DRM – age, disease status
3. TRM – age, BMI (obesity, overweight), graft source
4. PFS – age, disease status
5. OF – disease status, graft source
6. INF – age, BMI, CMVpn, CMVp, CMVn
7. GVHD – age, donor age, BMI

Disease types are included in the model, depending on the disease group the analysis was conducted on. E.g. for analyses without ALL (AML and MDS), AML dummy or MDS dummy were used.

Model Diagnostics

Schoenfield residuals

Simulating hazard function

Power Calculations

We conducted meta-analyses to combine the effects of both cohorts (discussed in next section below), as such, power calculations were done considering the sample size of both cohorts combined. Minimum detectable hazard ratios of recipient and/or donor depend on three variables: 1.) proportion of individuals experiencing an event, 2.) frequency of a causal variant, and 3.) the quality of genotyping SNPs that capture the genetic variation underlying the hazard of an event. Events are measured at 100 days and at the most updated observation time (most recent phenotype data available is May 5th, 2017) post-HSCT. The events that will be measured are death due to transplant (TRM) and specific causes of death (organ failure, infection, GVHD) attributable to TRM. OS is a function of TRM and thus we will present minimum detectable hazard ratios for OS. The proportion of events (Figure 3) ranges from infrequent events to (i.e. TRM subtypes) to frequent events (i.e. OS). We will assume lower and upper bound causal variant frequency between 5% and 40%, respectively. For Aim 1, we assumed SNPs selected for genotyping capture 85% of the variation across each gene; thus, all power calculations are corrected by setting our effective total sample size equal to 0.85 x 3532 = ~3000. We present the range of hazard ratios detectable for varying proportions of events and allele frequencies in a univariate model assuming 80% power to detect genome-wide significance at 5x10-8. With 3,532 recipient-donor pairs, the minimum detectable hazard ratio under these assumptions is identical for recipient genotype, donor

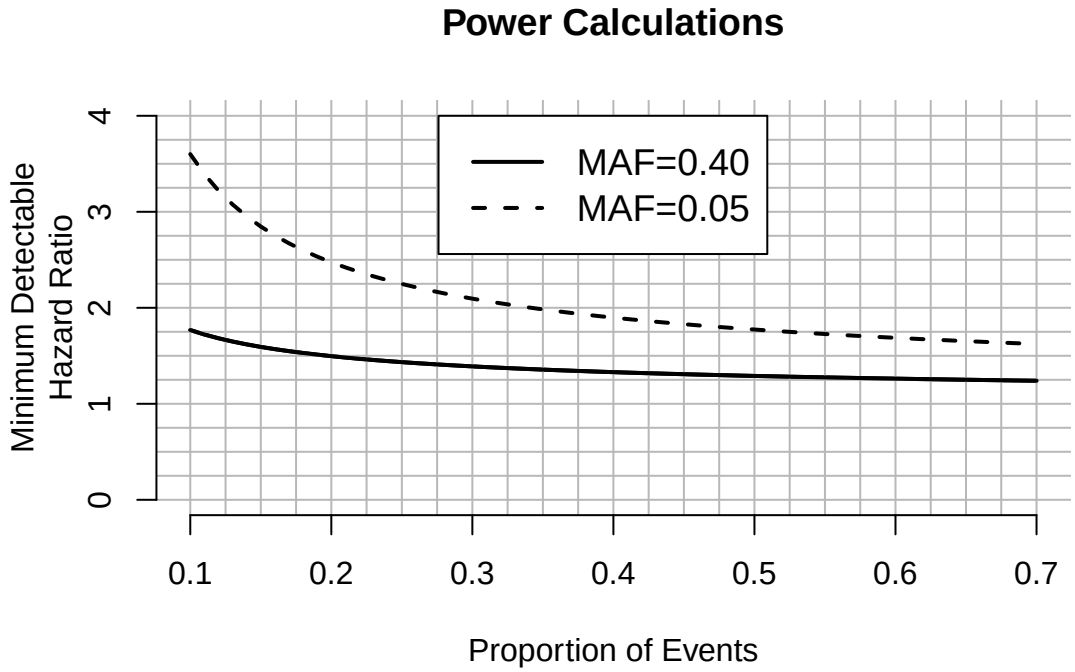

Figure 1.8: Hazard ratios associated with survival models.

genotype, and the mismatch between donor and recipients. Given the minimum proportion of events experienced in TRM subtypes and overall TRM are between 0.10 and 0.30 with a common allele (MAF=0.40), we have power to detect hazard ratios between 1.69 and 1.35, respectively. Under these same proportion of events, with more rare variants (MAF=0.05), we have the power to detect hazard ratios between 3.3 and 2.0, respectively. For OS models, assuming the overall rate of death is 0.50, we can detect SNPs in with hazard ratios between 1.26 (MAF=0.40) and 1.7 (MAF=0.05).

Lower bound is based off each TRM subtype being at least 0.10 (10%) of all patients.

Meta-analysis

CHAPTER 2: Literature Review and Replication/Validation

Literature Review

Genetic associations studies usually fall under two main subcategories, candidate gene association studies (CGAS) or genome wide association studies (GWAS). CGAS are association studies that investigate specific genes or regions of interest. Typically these are performed when researchers believe that the underlying biology is understood and they want to identify specific markers that contribute to genetic variation in these 'known' regions.

An extensive literature search of PubMed was performed using to identify peer-reviewed scientific studies (published on or before December 30, 2016) that reported non-HLA genetic polymorphisms associated with survival outcomes after allogeneic BMT, including disease-related mortality (DRM), progression-free survival (PFS), transplant-related mortality (TRM) and/or overall survival (OS) (Karaesmen et al. 2017). The PubMed search terms, filtering approach and link to all articles described herein are provided in the Supplemental Methods.

For over a decade, researchers have conducted candidate gene association studies of patient survival outcomes after allogeneic blood or marrow transplantation (BMT). The intent of these studies was to identify genetic variants outside of the human leukocyte antigen (HLA) region that would increase knowledge about clinical management or serve as a potential target for novel therapeutics. [1–70]

The majority of these studies tested for associations in small datasets, ranging from a few dozen to a few hundred patients and donors, included heterogeneous diseases spanning benign to malignant hematological diseases, related and/or un-

related donors with various degrees of HLA-matching and patients treated across multiple decades, from the 1980s through early 2000s.

We conducted the first adequately powered evaluation of these candidate SNP and gene hypotheses using typed and imputed data from an existing genome-wide association study (GWAS) named Determining the Influence of Susceptibility COnveying Variants Related to one-Year mortality after BMT (DISCOVeRY-BMT) to replicate or validate these published associations.[71−72] In addition, we leveraged the available genome-wide data from DISCOVeRY-BMT and measured the aggregate association of all SNPs in the candidate genes with survival outcomes to determine how many of these candidate genes play a significant role in survival after transplant. Lastly, using publically available data, we characterized the potential functionality of each candidate SNP in relation to the gene of interest.

Replication and Validation of Candidate Gene Studies

Results from genetic association studies should be reproduced in independent samples in order to confirm findings (Colhoun, McKeigue, and Smith 2003). Researchers have defined two distinctive terms to describe the reproducibility based on differences between the original study population and the confirmation studies: replication and validation (Igl, Konig, and Ziegler 2009). Replication is defined as the original and confirmation studies both having similar inclusion criteria (including the same ethnic/ancestral population) so that any differences between the study populations can be attributed to random variation [(Igl, Konig, and Ziegler 2009). Validation is defined as the original and confirmation study populations having different inclusion criteria (including different ethnic/ancestral populations) so that any differences between the original and confirmation study could be due to systematic variation (Igl, Konig, and Ziegler 2009).

Thus, replication analyses were conducted when the original study included HLA-matched unrelated donor BMTs in patients of European ancestry. Validation analyses were performed on studies of leukemia patients of non-European ancestry, patient populations who received a BMT from a matched related donor, or patient populations that were mixed between those who received a BMT from related and unrelated donor. For studies of outcomes involving multiple hematologic malignancies, the entire DISCOVeRY-BMT study population was analyzed. If the original study population was specified as AML, ALL and/or MDS, the same disease inclusion criteria were applied so that the replication/validation study population aligned with that of the original study population.

Gene-Based Association Testing

VErsatile Gene-based Association Study 2 (VEGAS2) software was used for gene-based association testing (Mishra and Macgregor 2015). VEGAS2 uses 10^6 Monte Carlo simulations to test the global significance of an association for sets of SNPs in defined genomic regions. VEGAS2 reports a gene-based P-value for each gene determined using individual SNP association P-values. Directional effects are not incorporated into analyses; thus, all SNPs can be aggregated without dampening an association signal. For the gene-based replication or validation analyses, the P-values from typed and imputed SNPs in DISCOVeRY-BMT (+/- a 10kb flanking region) meta-analyses of OS, PFS, TRM and DRM were used as input into the VEGAS2 software. Gene-based P-values were calculated for donor, recipient, and R-D mismatch analyses of the full cohort (ALL, AML and MDS patients) or homogenous disease subgroups (ALL or AML or MDS patients) corresponding to the analyses performed in the original studies.

Functional Annotation

RegulomeDB Blood expression quantitative trait loci (eQTL) Browser, and Variant Effect Predictor (VEP)88 were used to provide functional annotation of the candidate SNPs. For each database, the raw data scores, P-values and annotations, respectively were downloaded from each website and assigned to each SNP in our list. RegulomeDB scores are categorized as follows: 1a-1f are likely to affect transcription factor binding and linked to expression of a gene target; 2a-2c are likely to affect transcription factor binding; 3a-3b are less likely to affect transcription factor binding, and > 3 has minimal binding evidence. A RegulomeDB score is assigned based on the level and evidence of functional modification attributable to the SNP in multiple cell lines from a range of tissues, with scores from 1 to 7, with 1 having the highest functional effect, supported by experimental evidence and 7 having no modifying effect.

RegulomeDB database derives these annotations using the publically available data sets from Gene Expression Omnibus (GEO), the Encyclopedia of DNA elements (ENCODE) project and the Roadmap Epigenome Consortium. The Blood eQTL data are derived from a study of correlations between genetic variants and gene expression in over 5000 patients, with replication in almost 3000 individuals. Herein, we consider only cis-eQTLs, defined as < 250KB distance between the SNP chromosomal position and the probe midpoint for gene expression. VEP was used to determine the hypothetical functional importance of missense and nonsense variants based on SIFT, Mutation Taster and PolyPhen-2.

Results

DISCOVeRY-BMT Patient Characteristics

DISCOVeRY-BMT cohorts 1 and 2 include mostly 10/10 HLA-matched unrelated donors, with 281 8/8 HLA-matched donor-recipient pairs in cohort 2; all patients are of European continental ancestry. Cohorts do not differ by intensity of conditioning regimen, recipient or donor sex proportions, KPS/LPS scores. However, cohort 1 includes more ALL patients whereas cohort 2 includes more recipients with MDS. AML disease status also differs between cohorts at $p<0.01$ (Table 1).

Candidate Gene Studies of Survival Outcomes

The literature search identified 70 publications that studied a total 458 SNPs and 2 multi-allelic polymorphisms in 171 genes (Figure 1, Table S1). Studies included patients who received a transplant from an HLA-matched unrelated donor (19 articles), an HLA-matched related donor (23 articles), or both (28 articles) (Table S1). Study populations included patients and donors of European ancestry (53 articles), Asian ancestry (15 articles), or mixed genomic ancestry (2 articles) (Table S1).

A total of 14 articles assessed genetic variation in HLA matched unrelated donor (URD) BMT patients of European ancestry, but only 7 of these articles reported significant associations ($P<0.05$ or an author specified significance threshold) and thus comprise our replication study (Table S2, Table S3). A total of 56 articles tested associations in either a combination of related and unrelated donors (RD-URD), only related donors (RD) and/or in non-European populations; 39 of these 56 articles reported at least one significant SNP association with survival out-

come and we attempted to validate the significant findings from these 39 articles (Table S2, Table S4).

Replication

DISCOVeRY-BMT cohorts were used to replicate published studies of European American acute leukemia or MDS patients treated with an unrelated donor BMT.1-14 Of the 7 articles whose findings we attempted to replicate, 2 articles tested multi-allelic models in NOD2/CARD155 and CCR56; 5 articles tested single SNP associations in TGFB11, CD2743, CD403, TNFSF43, HMGB14,IL1A7, IL1B7, and NOD2/CARD152 (Table 2, Figure 2, Table S3).1-7

The two NOD2/CARD15 associations were based on a three-variant R-D pair model [rs2066844 (SNP8), rs2066845 (SNP12) and rs2066847 (SNP13)] and single SNP associations with SNP13.27 The null type is when the R-D pair are homozygous common allele for all three SNPs and the effect allele combination is the presence of 1 or more minor alleles at any of the three SNPs within the R-D pair. In a study of 196 patients who received an unrelated donor BMT for AML or ALL, the NOD2/CARD15 multi-SNP model was significantly associated with OS (RR: 1.6, 95% CI 1.1-2.4, P=0.02) and TRM (RR: 1.6, 95% CI 1.1-2.4, P=0.02).5 However, in the DISCOVeRY-BMT AML and ALL patients (n=1597) treated with an unrelated donor BMT, there was no association with OS (HR: 1.03, 95% CI 0.9-1.2, P=0.72) or TRM (HR: 1.1, 95% CI 0.8-1.4, P=0.6, Figure 2, Table S3). In a study of 342 unrelated donor genotypes matched with AML or ALL patients, rs2066847 (SNP13) alone significantly increased risk of TRM and OS approximately 3-fold (P=0.001) and 2.5 (P=0.001), respectively2, however DISCOVeRY-BMT donor genotypes, did not associate with either TRM (HR: 1.17, 95% CI 0.78-1.74, P=0.45) or OS (HR: 0.98, 95% CI 0.73-1.31, P=0.89, in ALL or AML patients (Table 2, Fig-

ure 2, Table S3).

One of the largest candidate gene studies (N=1370) showed significant associations between PFS and recipient CCR5 H1/H1 genotype (n=163), as well as with author defined genotype risk subgroups and OS.6 In DISCOVeRY-BMT, neither the CCR5 H1/H1 genotype (n=294) nor the genotype risk groups defined by H1/H16 status were significantly associated with PFS or OS (Figure 2, Table S3). The genotype risk groups tested by the authors were substantially smaller than the full cohort (Table 2). In DISCOVeRY-BMT these subgroups were approximately twice as large as those in the original study and adequately powered to detect these associations. Attempts to replicate single SNP associations in TNFSF4,3 TGFB1,1 HMGB15, IL1A7, and IL1B7 also failed (Table 2, Figure 2, Table S3).

Validation

We attempted to validate 36 polymorphisms in 26 genes from 39 candidate gene articles (Table S2, Table S4),15-52 including: ABCB129,32, CD1442, CTLA428,40,43-46,51, CYP2C1938, DAAM252, EP30036, ESR117, GSTA219, GZMB24, ICAM148, IL23R20,22, IL615-17, IRF337, KLRK123, LIG348, MTHFR31,35,41, MUTYH48, NOD2/CARD1525,27,30,33,50, NOS130, P2RX734, TDG48, TIRAP17, TLR442, TYMP26, and VDR18,21,39,47. These studies reported significant genetic associations with survival after transplant in patients who received a HLA-matched related donor BMT (19 articles) or had a study population including HLA-matched related and unrelated donor BMT patients, without stratification of results (17 articles). We also attempted to validate survival associations seen in non-European leukemia patients who received an unrelated donor BMT (3 articles). We present results of variants reported significant in at least two separate publications in Table 3 and Figure 3.

Our validation analyses identified only one variant associated at P<0.05. Donor variation in rs1800795 (IL-6) associated with OS (HR: 1.11, 95% CI 1.0-1.2, P=0.02) (Figure 3, Table S4). This SNP association was initially reported in a single study by Balavarca et al., 2015, (HR: 1.29, 95% CI 1.07-1.55, P=0.007) in patients with acute leukemia, CML, or lymphoma treated with a matched related or unrelated donor BMT (n=743).

SNPs within NOD2/CARD15 were the most frequently studied and reported of all candidate gene association studies in our validation set (Table S2). NOD2/CARD15 is a susceptibility gene for inflammatory bowel disease and may be involved in Crohn's disease.27 We attempted to validate studies that reported an association of NOD2/CARD15 and survival outcomes in HLA-matched related and unrelated donor BMT patients27,30,33 or HLA-matched related donor BMT patients.25,50 Three studies reported significant findings between the presence of the NOD2/CARD15 multi-SNP polymorphism in either donor or recipient with TRM27,50 or PFS,25 however this did not validate in the DISCOVeRY-BMT cohorts (Figure 3, Table 3). There was also no significant association of the single variant rs2066842 in related/unrelated donors with PFS,30 or the single variant rs2066847 (SNP13) in recipients of related/unrelated donor BMTs with TRM (Figure 3, Table 3)33 in the DISCOVeRY-BMT cohorts.

Due to its known functions and perceived implications in transplant biology,43 associations with multiple SNPs in CTLA4 have been tested in numerous transplant populations (Table S2), with 4 CTLA4 SNPs (rs3087243, rs231775, rs4553808, rs5742909) reported as significantly associated with survival after related or unrelated donor allogeneic BMT in acute leukemias, CML, lymphomas, MDS, and other hematological disorders (Table 3). Attempts to validate CTLA4 SNPs with DRM, PFS, OS, and TRM were unsuccessful in the DISCOVeRY-BMT

cohorts (Table 3, Figure 3, Table S4).

The remaining results of the 25 additional candidate genes containing SNPs that were tested in the DISCOVeRY-BMT cohorts are summarized in Tables S4 and 3 as well as Figure 3; no SNP associations were found at $P<0.05$. Importantly, the P-value distribution of the single SNP associations showed no deviation from the null expectation with 95% confidence intervals (Figure S2), suggesting we cannot reject the null hypothesis of no association with survival outcome.

Gene based replication and validation of previous studies

The reviewed candidate gene studies first selected genes based on their hypothesized or known function, and subsequently selected variants within that gene for single SNP or haplotype testing. Thus, while SNPs and haplotypes were tested individually for association, the hypotheses from the literature can be considered gene-based. The density of typed and imputed markers in the DISCOVeRY-BMT recipients and donors allows us to measure the aggregate effect of all SNPs within each candidate gene on survival. Genes were selected for testing from the same literature summarized above for the replication and validation SNP and haplotype analyses. VEGAS2 gene-based testing did not reveal any associations at $P<0.05$ with any of the survival outcomes in either the replication or validation groups (Table S5).

Candidate polymorphism annotation

Candidate gene SNPs were analyzed using the RegulomeDB,86 VEP88 and Blood eQTL Browser87 databases to assess their functional characteristics and better understand their biological framework. Eighty percent of previously reported SNPs had RegulomeDB scores greater than 3 (Figure 4, Table S6), indicating that

these SNPs have minimal to no effect on modifying transcription. This distribution aligns with the overall distribution of SNPs in the genome, thus the candidate SNPS are not enriched for their impact on gene expression or transcription factor binding. Our replication and validation analyses includes 2 protein coding variants, VEP shows that only, rs2066845 (SNP12) in NOD2/CARD15, is predicted to be damaging and disease causing.

The Blood eQTL browser determines if candidate SNPs have a significant role in cis gene expression of the candidate gene. Of the 171 genes included in our literature search results, 52% have at least one significant cis-eQTL at a probe-level false discovery rate (FDR) < 0.05. On a genome-wide level, approximately 44% of genes have blood cis-eQTLs (FDR $P<0.05$). However, despite over half of the candidate genes having blood cis-eQTLS, only 13% of the candidate SNPs reported in these articles are blood cis-eQTLs. Thus, while blood eQTLs have been identified in these genes, they were not genotyped and analyzed in these candidate gene studies. Furthermore, almost half of the eQTLs in the candidate gene studies are correlated with expression that is not the candidate gene but rather a nearby gene. For example, rs7975232 (VDR) is an eQTL for SLC48A1 while the CTLA4 SNPs are actually eQTLs for CD28. The remaining eQTLs were correlated with expression of the candidate gene of interest, but in most cases, were also significant eQTLs for several other nearby genes (Table S6).

Discussion

Our study aimed to replicate or validate all previous genetic association studies that investigated the non-HLA genetic effects on allogeneic BMT survival. Since previous studies selected SNPs in candidate genes, we conducted both single SNP and gene-based analyses to determine the aggregated SNP associations within can-

didate genes while still accounting for dependence between signals due to LD.

The only association with P<0.05 in our replication and validation analyses using DISCOVeRY-BMT was the donor SNP rs1800795 in IL-6 with OS. As reported,18 the rationale for studying this SNP was based on the immunological function of IL-6 and two prior findings showing that it was associated with GvHD93, and response to chronic Hepatitis C virus therapy.94 We found no evidence of association at P<0.05 between donor SNP rs1800795 with death due to either GvHD or infection in the DISCOVeRY-BMT cohort (data not shown). Furthermore, rs1800795 is located in the intronic region of IL-6, has no effect on IL-6 expression or levels,95 but rather is an eQTL for two other nearby genes.95,96

In addition to exploring this IL-6 association further we felt the validation of the CCR5 associations of H1/H1 genotype with outcome required additional efforts, as these associations were found in the largest study we attempted to validate, samples were also from CIBMTR (earlier years than our study population) and unlike many of the other studies survival effects only started to appear approximately two years post-transplant. Analyses outlined in Table 2 were performed without censor at 1 year for overall survival (median survival time 13.7 months, range <1 month-125.6 months) and progression free survival (median time 11.1 months, range <1 month-125.6 months). There were no genotype associations with either outcome at P<0.10.

Another frequently studied gene, CTLA4, highlights the heterogeneity specific to studies of genetic variation in transplant and perhaps helps explain why we did not replicate or validate associations. rs5742909 in CTLA4 was tested for association with various survival outcomes after transplant in 6 independent studies of HLA matched-related donor-recipient pairs. In donors, the variant was found to be associated with DRM in one small study (N=120), this was the only study

that tested donor genotype with DRM. Likewise, 1 out of 9 papers testing the association of rs231775 with survival outcomes measured the association of PFS with recipient rs231774 in 164 recipients (P=0.025). Despite the frequency with which these two CTLA4 variants were studied, for both SNP-outcome combinations DISCOVeRY-BMT is the only validation attempt. These SNPs are like those of many candidate gene hypotheses, in that they have not been tested in the same genome for the same outcome in similar populations, and if they have the N is small (Table S1).

Our inability to replicate or validate previous candidate gene associations could also be due to differences in inclusion criteria with respect to disease, donor relation, or to differences in our endpoint of 1-year survival versus longer-term survival. The previous genetic associations were hypothesized to be independent of underlying hematologic disease, therefore we would expect to replicate or validate these associations in a homogeneous patient population such as DISCOVeRY-BMT. When possible we aligned our study population to the original candidate gene study (i.e. restricted to AML patients only). While DISCOVeRY-BMT focused on early 1-year survival, which may have different genetic contributions than later survival, many of the survival curves in the significant candidate gene articles show separation by genotype well before 1-year post-transplant, thus the significant published variants do not appear to correlated with only longer-term survival.

The large sample size of the DISCOVeRY-BMT provides adequate statistical power to attempt replication and validation of previously published candidate gene analyses71, however we did not reproduce these findings, similar to two other recent studies attempting to replicate previous candidate gene associations with GvHD after BMT.73,97 Other reports have also concluded that a substantial amount of the published candidate gene literature has presented false positive associations.98

Confirming genetic association studies is vital to identify true positive genetic variants that may contribute to complex phenotypes. False associations lead to wasted time, energy and money in pursuit of confirmatory studies and could harm patients by delaying clinical discovery or by applying clinical studies too quickly without replication. Annotation of the previously reported SNP associations using publically available data show that few variants are functional; only one SNP is predicted to be damaging or deleterious, a small proportion of SNPs are correlated with gene expression, and an even smaller number are cis-eQTLs for the target gene of interest. Thus, while we did not replicate or validate these associations, the SNPs selected are not linked to functional annotation nor are they clearly related to the candidate genes. This underscores a fundamental problem with candidate gene studies which are hostage to the state of scientific knowledge at the time. Adequately powered testing of genetic associations with transplant outcomes remains critical to discovery and replication of genetic associations with the ultimate goal of improving patient outcomes.

CHAPTER 3: gwasurvivr

Introduction

Genome-wide association studies (GWAS) are population-level experiments that investigate genetic variation in individuals to observe single nucleotide polymorphism (SNPs) associations with a phenotype. Genetic variants tested for association are genotyped on an array and imputed from a reference panel of sequenced genomes, e.g. 1000 Genomes Project (Consortium 2015) or Haplotype Reference Consortium (HRC).

Imputed SNPs can be tested for association with binary outcomes (cases/controls) dnd quantitative outcomes (e.g., height) using a range of available software packages, including SNPTEST (Marchini et al. 2007) or PLINK (Purcell et al. 2007). However, existing software options for performing survival analyses, genipe (Lemieux Perreault et al. 2016), SurvivalGWAS_SV (Syed, Jorgensen, and Morris 2017), and GWASTools (S. M. Gogarten et al. 2012) either require user interaction with raw output, were not initially designed for survival and/or have long run times. For these reasons, we developed an R/Bioconductor package, gwasurvivr (A. A. Rizvi et al. 2018), for genome wide survival analyses of imputed data in multiple file formats with flexible analysis and output options.

Building an R package

To build an R package

Data Structure

Gwasurvivr can analyze data in IMPUTE2 format (B. N. Howie, Donnelly, and Marchini 2009), in VCF files derived from Michigan (Das et al. 2016) or Sanger imputation servers (S. McCarthy et al. 2016), and directly genotyped PLINK format (Purcell et al. 2007). Data from each are prepared in gwasurvivr by leveraging existing Bioconductor packages GWASTools (S. M. Gogarten et al. 2012) or VariantAnnotation (Obenchain et al. 2014) depending on the imputation file format.

IMPUTE2 Format: IMPUTE2 (Howie, et al., 2009) format is a standard genotype (.gen) file which store genotype probabilities (GP). We utilized GWASTools in R to compress files into genomic data structure (GDS) format (Gogarten, et al., 2012). This allows for efficient, iterative access to subsets of the data, while simultaneously converting GP into dosages (DS) for use in survival analyses.

VCF Format: VCF files generated from these Michigan or Sanger servers include a DS field and server-specific meta-fields (INFO score [Sanger] or r^2 [Michigan], as well as reference panel allele frequencies) that are iteratively read in by VariantAnnotation (Obenchain, et al., 2014).

PLINK Format: Plink bed files contain genotype information encoded in binary format. Fam and bim files include the information of phenotype and marker location, respectively (Purcell, et al., 2007).

gwasurvivr implements a Cox proportional hazards regression model (Cox, 1992) to test each SNP with an outcome with options for including covariates and/or SNP-covariate interactions. To decrease the number of iterations needed for convergence when optimizing the parameter estimates in the Cox model we modified the R package survival (Therneau and Grambsch, 2000). Covariates in the model are first fit without the SNP, and those parameter estimates are used

as initial points for analyses with each SNP. If no additional covariates are added to the model, the parameter estimation optimization begins with null initial value. (Supplementary Figure 1).

Survival Analysis

Survival analyses are run using genetic data in either VCF or IMPUTE2 (Howie, et al., 2009) formats and a phenotype file, which contains survival time, survival status and additional covariates, both files are indexed by sample ID. In addition to genomic data, the VCF files contain both sample IDs and imputation quality metrics (INFO score or r^2), while IMPUTE2 (Howie, et al., 2009) come in separate files (.gen, .sample, and .info). Gwasurvivr functions for IMPUTE2 (impute2CoxSurv or gdsCoxSurv) and VCF (michiganCoxSurv or sangerCoxSurv) include arguments for the survival model (event of interest, time to event, and covariates) and arguments for quality control that filter on minor allele frequency (MAF) or imputation quality (michiganCoxSurv and sangerCoxSurv only). INFO score filtering using impute2CoxSurv can be performed by accessing the .info file from IMPUTE2 results and subsequently providing the list of SNPs to 'exclude.snps' argument to gwasurvivr. Users can also provide a list of sample IDs for gwasurvivr to internally subset the data. gwasurvivr outputs two files: (1) .snps_removed file, listing all SNPs that failed QC parameters and (2) .coxph file with the results from the analyses, including parameter estimates, p-values, MAF, the number of events and total sample N for each SNP. gwasurvivr also allows the number of cores used during computation on Windows and Linux to be specified. Users can keep compressed GDS files after the initial run by setting keepGDS argument to TRUE when analyzing IMPUTE2 data (Howie, et al., 2009). On successive runs, gdsCoxSurv can then be used instead of impute2CoxSurv to avoid

compressing the data on each GWAS run.

Simulations and Benchmarking

Computational runtimes for gwasurvivr were benchmarked against existing software comparing varying sample sizes and SNP numbers, with 4, 8 or 12 covariates and for a single chromosome with 15,000-25,000 individuals. In addition, we evaluated time for gwasurvivr for a GWAS (~6 million SNPS) for 3000, 6000 and 9000 samples. All benchmarking experiments were performed using IMPUTE2 format (comparison packages do not take VCF from either imputation servers).

Descriptions of simulated genotype and phenotype data are in the Supplementary Data.

Results

gwasurvivr was faster than genipe (Lemieux Perreault, et al., 2016), SurvivalGWAS_SV (Syed, et al., 2017), and GWASTools (Gogarten, et al., 2012) for 100,000 SNPs at N=100, and 5000, with the exception of SurvivalGWAS_SV at N=1000 (Figure 1A). Similarly, increasing the number of covariates for gwasurvivr has minimal effects on runtime versus other software (Figure 1B). Gwasurvivr computes for large sample sizes, however, compression time increases with increasing sample size, and likely will be limited by available RAM on a machine or cluster (Figure 1C). The keepGDS argument helps address this and results in reduced run times (Figures 1C and 1D), i.e. < 3 hours for a GWAS of N=9,000. A ~6 million SNP GWAS can be run in < 10 hours for 9000 samples when using separately scheduled jobs on a supercomputer (Figure 1D). However, gwasurvivr overcomes memory limitations often attributed to R by processing subsets of the entire data,

and thus it is possible to conduct genome-wide survival analyses on a typical laptop computer.

gwasurvivr is a fast, efficient, and flexible program well suited for multi-core processors and easily run in a computing cluster environment.

`gwasurvivr` is an R package that can be used to conduct survival analysis (Cox proportional hazards model) on imputed GWAS data from either IMPUTE2 (Howie, et al., 2009) or VCF files from the Michigan and/or Sanger imputation servers. `gwasurvivr` can also be used on directly typed data in plink format (`.bed`, `.bim` and `.fam` files).

Herein, we detail our implementation of the Cox model, generation of the simulated data and survival benchmarking and graphically report the correlation of `gwasurvivr` beta coefficient estimates, minor allele frequencies (MAF) and p-values with those produced from SurvivalGWAS_SV, genipe, and GWASTools.

To reproduce the data and create Figure 1 and Supplementary Figures 2-4, the data is available on the gwasurvivr manuscript repository. GitHub Large File Storage (LFS).

To clone the whole repository:

```
git lfs clone https://github.com/suchestoncampbelllab/gwasurvivr_manuscript.git
```

Implementation of Survival Model in gwasurvivr

Modifying coxph

We decrease computation time by decreasing the number of Newton-Raphson iterations used to optimize the partial likelihood function in the Cox proportional hazard models. To do this, a survival model was fit using only non-genetic covariates (i.e. the SNP is not included and only covariates are fit); `survival::coxph`

(Therneau and Grambsch, 2000) is modified such that gwasurvivr manually creates the objects found in the helper function (`survival::coxph.fit`) that fits the Cox model.

These variables are then passed to `survival::coxph.fit`.

Benchmarking with survival package

To assess if providing initial estimates from covariates versus using the survival function as implemented in the survival package improves computational time, we tested a dataset of 500 individuals at 7255 SNPs with 1, 2, or 3 covariates. These data are a subset of the simulated data described in detail below.

The helper function `gwasurvivr:::coxParam`, adjusted for this Supplementary documentation is labeled `gcoxph`. In `gcoxph_model.R` we fit the model without the SNP and the parameter estimates are then used as initial points for all subsequent models and applied over all SNPs in the dataset. If there were no covariates, the initial estimates would be null. The function `coxph_model.R` implements a `survival` model (survival package, Therneau and Grambsch, 2000) without using the optimization starting point obtained from including covariates in the model.

To test the package runtime over a pre-specified number of iterations and including 1, 2, or 3 covariates the `microbenchmark` package in R was used. The code for Supplementary Figure 1) is available.

By leveraging an initalization point from the analyses with covariates `gwasurvivr` (gcoxph) is several seconds faster than the survival analyses function as implemented in `survival` (coxph, Therneau and Grambsch, 2000) in R (**Supplementary Figure 1**). While this is a small test dataset, in practice this would be an appreciable difference when testing across several thousands of samples and millions of SNPS. In the `gwasurvivr` package, we opted to use

`parallel::parApply` instead of `base::apply` as shown above to compute across multiple cores.

Computational Experiments

We used the University at Buffalo Computational Center for Research (UB CCR) academic cluster for our benchmarking analyses. Each analysis was run exclusively on node CPU-L5520 with the same system specifications, controlling the computational resources for each run. The UB CCR uses Simple Linux Utility for Resource Management (SLURM) scheduling for jobs. SLURM scripts to run the analyses were generated using shell scripts below. Benchmarking was performed using identical CPU constraints, 1 node (2.27 GHz Clock Rate) and 8 cores with 24 GB of RAM, on the University at Buffalo Center for Computational Research supercomputer. With the exception of the larger sample size tests, these were run using the same node but 12 CPUs. genipe (Lemieux Perreault, et al., 2016), SurvivalGWAS_SV (Syed, et al., 2017), and GWASTools (Gogarten, et al., 2012) were performed as specified by the authors on available online documentation. We performed the following benchmarking runtime experiments either against existing software or against time with varying N and SNP numbers that were performed:

Simulation 1. Compare gwasurvivr against genipe, GWASTools and SurvivalGWAS_SV - varying sample sizes (n=100, n=1000, n=5000) and 100,000 SNPs (m=100000) and 3 non-genetic covariates

Simulation 2. Comparison of gwasurvivr, genipe, GWASTools and SurvivalGWAS_SV with N=5,000 and 100,000 SNPs (m=100,000) with 4 covariates (age, drug treatment, sex and 1 PC), 8 covariates (age, drug treatment, sex and 5 PCs) and 12 covariates (age, drug treatment, sex and 9 PCs)

Simulation 3. Increasingly larger sample sizes (N=15K, 20K and 25K) tested

on Chromosome 22

Simulation 4. Full autosomal GWAS with varying sample sizes (N=3K, 6K and 9K)

Simulating Genotypes and Phenotypes

Genotypes

HAPGENv2 (Su, Marchini, and Donnelly 2011) was used to generate simulated genetic datasets from 1000 Genomes Project CEU data (NCBI Build 36) for all benchmarking experiments. To replicate simulations the 1000 Genomes Project CEU data should be downloaded in its entirety (only a subset is available on our GitHub repo). The code for all HAPGENv2 simulations are available on our GitHub.

Phenotypes

For each sample size tested, survival events (alive/dead) were simulated as two separate datasets. For the dead dataset, time to event and covariates were simulated using a normal distribution. For the alive dataset, time was simulated by randomly sampling weighted probabilities for times to simulate few samples being censored, covariates were simulated from a normal distribution. Principal components (PCs) were simulated using random normal distributions with decreasing variance for each additional PC. Furthermore, the `.sample` file from IMPUTE2 includes 4 columns (ID_1, ID_2, missing, and sex) which link individuals with their respective genotypes. For SurvivalGWAS_SV and GWASTools, the simulated phenotypes were appended to column 5 onward in the `.sample` file.

The following genotypes and phenotypes were simulated:

Simulations 1 and 2. Subset of chromosome 18 for 100,000 SNPs 1) varying N and 3 covariates done in triplicate and 2) with 4, 8 and 12 covariates

- genotype code
- phenotype code
- PCs phenotype code

Simulation 3. chromosome 22 (~117,000 SNPs) for larger sample sizes (N=15000-25000)

- genotype code

Simulation 4. Full GWAS for N=9000 (the smaller subsets were just parsed from the data during analyses)

- genotype code
- phenotype code
- simulate sample ids code

Benchmarking with other software capable of GWAS coxph survival analysis

We benchmarked `gwasurvivr` with GWAS survival analysis software, genipe, SurvivalGWAS_SV and GWASTools using simulated phenotype and genotype data. Genetic data were formatted as output from IMPUTE2 software (.GEN). Genipe, SurvivalGWAS_SV, and GWASTools do not directly take VCF data output from Sanger or Michigan imputation servers. SurvivalGWAS_SV does accept VCF files as an input but uncompressed and not explicity the same format that Sanger and Michigan imputation servers output, rendering additional steps to be taken. The benchmarking with IMPUTE2 was done for (1) varying sample sizes and (2) varying additional non-genetic covariates. Both are described here.

gwasurvivr

The following scripts were used to run gwasurvivr using `impute2CoxSurv`. These R scripts are run using a shell script (SLURM script) that pass the system variables into R (facilitated by the R package `batch`).

N=100, 1000 and 5000 with M=100K SNPs + 3 non-genetic covariates in triplicate:

- `run_gwasurvivr.R`

- `create_gwasurvivr_scripts.sh`

N=5,000 and M=100K with 4, 8 and 12 covariates:

- `run_gwasurvivr_covs.R`

- `gwasurvivr_cov4.sh`

- `gwasurvivr_cov8.sh`

- `gwasurvivr_cov12.sh`

genipe

For genipe, the shell scripts was used to generate SLURM scripts for genipe and each sample and SNP set. We used specific settings for OPENBLAS that are suggested on genipe's website to ensure that computational efficiency was maximized.

varying sample sizes + 3 non-genetic covariates:

- `create_genipe_scripts.sh`

additional covariates:

- `genipe_cov4.sh`

- `genipe_cov8.sh`

- `genipe_cov12.sh`

SurvivalGWAS_SV

To maximize the performance of SurvivalGWAS_SV, these jobs were run using "array" jobs as recommended by the authors. An example batch script, provided in the SurvivalGWAS_SV documentation, was converted from PBS to SLURM. 24GB of ram was not needed on all runs, however was used to ensure each run remained uniform. The jobs were split into array sets of 1000 SNPs for m=100,000, totaling 100 batched jobs in a single array. We define rate-limiting array as the array index that had the longest runtime. In the main manuscript, we report SurvivalGWAS_SV runtimes as the rate-limiting array runtime. This is an important caveat and bears consideration when using SurvivalGWAS_SV. Depending on availability on the computing cluster, the analyses could be completed as quickly as the longest individual array job (which is shown in Figure 1), or potentially the entire runtime could be equal to the summation runtime of all of the array indices if these cannot be run simultaneously (or if there are failures with any of the array indices). The shell script below was used to generate SLURM scripts for SurvivalGWAS_SV for each sample and SNP set.

N=100, 1000 and 5000 with M=100K SNPs + 3 non-genetic covariates in triplicate:

- `create_sv_scripts.sh`

N=5,000 and M=100K with 4, 8 and 12 covariates:

- `sv_cov4.sh`
- `sv_cov8.sh`
- `sv_cov12.sh`

GWASTools

For GWASTools, the files are converted to GDS format and survival is run using `GWASTools::assocCoxPH` within `gwastools_survival.R`. The R script was passed to the SLURM scripts using the script `create_gwastools_scripts.sh`. GWASTools does not run in parallel across multiple cores on a single computing processor internally, however experienced users could code this themselves.

N=100, 1000 and 5000 with M=100K SNPs + 3 non-genetic covariates in triplicate:

- `gwastools_survival.R`
- `create_gwastools_scripts.sh`

N=5,000 and M=100K with 4, 8 and 12 covariates:

- `gwastools_survival_covs.R`
- `gwastools_cov4.sh`
- `gwastools_cov8.sh`
- `gwastools_cov12.sh`

Runtime large N chromosomes to test size limitations

We tested chr22 with different sample sizes of N=15,000; N=20,000; N=25,000 using `gwasurvivr::impute2CoxSurv`. The code for all of the runs can be found here. The R script called from the shell scripts to run these analyses is labeled `run_bigNs.R`.

Runtime GWAS with different sample sizes

We performed three GWAS (chr1-chr22) with different sample sizes (n=3000; n=6000; n=9000) using `gwasurvivr::impute2CoxSurv`. The code to simulate the

GWAS is available on our repository. The R script used to run these analyses is `run_fullgwas.R`. The shell script run these scripts on SLURM can be found here.

Time Plots

Figure 1

To generate Figure 1 times from the computation runtime were pulled from SLURM log files and collected using the perl scripts, which can be found in each of the log folders on our manuscript GitHub repository, compiled and Figure 1 was generated using the R code shown here.

Diagnostic Plots

Supplementary Figures 2, 3 and 4 below show the correlation of the coefficient estimates, minor allele frequency and p-values, respectively between gwasurvivr and all other software assessed. The correlations show excellent agreement. The R code used to generate supplemental figures 2-4 can be found here.

Coefficient Estimates

Minor Allele Frequency (MAF)

P-value Estimates

Full GWAS Runtimes

gwasurvivr calculations

Minor Allele Frequency (MAF)

For a given SNP with alleles A and B, where n_{AB} and n_{BB} are the number of individuals with AB and BB genotype respectively, and N is the sample size, the expected allele frequency of allele B ($freq_B$) be can be calculated as:

$$freq_B = \frac{n_{AB} + 2n_{BB}}{2N}$$

For individual i, the allele dosage of SNP j (D_{ij}) with alleles A and B, where allele B is the effect allele and p_{AB} and p_{BB} are the posterior genotype probabilities as computed by the imputation, is calculated as:

$$D_{ij} = p_{AB_{ij}} + 2 \cdot p_{BB_{ij}}$$

For SNP j The estimated allele frequency of an effect allele B (θ_{B_j}) can therefore be calculated as:

$$\theta_{B_j} = \frac{\sum_{i=1}^{N} D_{ij}}{2N}$$

This was coded in R as follows:

```
# calculate MAF
# genotypes variable is a matrix of dosages,
## where each column is a sample and each row is a SNP
exp_freq_A1 <- round(matrixStats::rowMeans2(genotypes)*0.5,4)
MAF <- ifelse(exp_freq_A1 > 0.5,
                  1-exp_freq_A1,
                  exp_freq_A1)
```

Imputation quality metric

Michigan Imputation Server

For the Michigan imputation server, imputation is performed using the minimac3 algorithm (Das et al., 2016). minimac3 computes and outputs an imputation quality metric known as R^2. R^2 is the estimated value of the squared correlation betwen imputed genotypes and true, unobserved genotypes (Das et al, 2016). The R^2 value is extracted directly from the Michigan imputation output VCF in `gwasurvivr::michiganCoxSurv`

Sanger Imputation Server

For the Sanger imputation server, we grab the `INFO` field directly from the VCF file in `gwasurvivr::sangerCoxSurv`. The INFO field is the IMPUTE2 (Howie, et al., 2009) score as calculated by the `bcftools + impute-info` plugin from posterior genotype probabilities (McCarthy et al., 2016).

IMPUTE2 Imputation

The INFO score for IMPUTE2 (Howie, et al., 2009) results are not calculated in `gwasurvivr` internally, instead we use the INFO scores that are provided in a separate file after performing imputation (`.info` file). Users select SNPs from the `.info` file to remove based on preferred criterion (ie INFO $< .8$) these are then used in the argument `exclude.snps` in `impute2CoxSurv` to filter out the SNPs prior to analysis.

CHAPTER 4: Application and Pipeline

CHAPTER 5: Acute Lymphobastic Leukemia (ALL) GWAS

CHAPTER 6: Conclusion and Future Work

REFERENCES

Botstein, David, and Neil Risch. 2003. "Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease." *Nat Genet* 33 Suppl (March): 228–37. doi:10.1038/ng1090.

Clay-Gilmour, Alyssa I., Theresa Hahn, Leah M. Preus, Kenan Onel, Andrew Skol, Eric Hungate, Qianqian Zhu, et al. 2017. "Genetic Association with B-Cell Acute Lymphoblastic Leukemia in Allogeneic Transplant Patients Differs by Age and Sex." *Blood Advances* 1 (20): 1717–28.

Colhoun, Helen M, Paul M McKeigue, and George Davey Smith. 2003. "Problems of Reporting Genetic Associations with Complex Outcomes." *The Lancet* 361 (9360). Elsevier: 865–72.

Consortium, 1000 Genomes Project. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571). Nature Publishing Group: 68.

Cox, David R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 34 (2). Wiley Online Library: 187–202.

Das, Sayantan, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, et al. 2016. "Next-Generation Genotype Imputation Service and Methods." *Nature Genetics* 48 (10). Nature Publishing Group: 1284.

D'Souza, Anita, Stephanie Lee, Xiaochun Zhu, and Marcelo Pasquini. 2017. "Current Use and Trends in Hematopoietic Cell Transplantation in the United States." *Biology of Blood and Marrow Transplantation* 23 (9): 1417–21. doi:10.1016/j.bbmt.2017.05.035.

Fine, Jason P, and Robert J Gray. 1999. "A Proportional Hazards Model for the Subdistribution of a Competing Risk." *Journal of the American Statistical Association* 94 (446). Taylor & Francis: 496–509.

Gogarten, Stephanie M, Tushar Bhangale, Matthew P Conomos, Cecelia A Laurie, Caitlin P McHugh, Ian Painter, Xiuwen Zheng, et al. 2012. "GWASTools: An

R/Bioconductor Package for Quality Control and Analysis of Genome-Wide Association Studies." *Bioinformatics* 28 (24). Oxford University Press: 3329–31.

Green, Richard E, Benjamin P Lewis, R Tyler Hillman, Marco Blanchette, Liana F Lareau, Aaron T Garnett, Donald C Rio, and Steven E Brenner. 2003. "Widespread Predicted Nonsense-Mediated mRNA Decay of Alternatively-Spliced Transcripts of Human Normal and Disease Genes." *Bioinformatics* 19 Suppl 1: i118–21.

Hahn, Theresa, Lara E Sucheston-Campbell, Leah Preus, Xiaochun Zhu, John A Hansen, Paul J Martin, Li Yan, et al. 2015. "Establishment of Definitions and Review Process for Consistent Adjudication of Cause-Specific Mortality After Allogeneic Unrelated-Donor Hematopoietic Cell Transplantation." *Biology of Blood and Marrow Transplantation* 21 (9): 1679–86. doi:10.1016/j.bbmt.2015.05.019.

Henig, Israel, and Tsila Zuckerman. 2014. "Hematopoietic Stem Cell Transplantation-50 Years of Evolution and Future Perspectives." *Rambam Maimonides Medical Journal* 5 (4): e0028. doi:10.5041/RMMJ.10162.

Howie, Bryan N, Peter Donnelly, and Jonathan Marchini. 2009. "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies." *PLoS Genetics* 5 (6): e1000529. doi:10.1371/journal.pgen.1000529.

Igl, Bernd-Wolfgang, Inke R Konig, and Andreas Ziegler. 2009. "What Do We Mean by 'Replication' and 'Validation' in Genome-Wide Association Studies?" *Human Heredity* 67 (1): 66–68. doi:10.1159/000164400.

International HapMap Consortium. 2005. "A Haplotype Map of the Human Genome." *Nature* 437 (7063): 1299–1320. doi:10.1038/nature04226.

Jorde, L B. 2000. "Linkage Disequilibrium and the Search for Complex Disease Genes." *Genome Res* 10 (10): 1435–44.

Karaesmen, Ezgi, Abbas A. Rizvi, Leah M. Preus, Philip L. McCarthy, Marcelo C. Pasquini, Kenan Onel, Xiaochun Zhu, et al. 2017. "Replication and Validation of Genetic Polymorphisms Associated with Survival After Allogeneic Blood or Marrow Transplant." *Blood* 130 (13). American Society of Hematology: 1585–96. doi:10.1182/blood-2017-05-784637.

Lander, E S, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature*

409 (6822): 860–921. doi:10.1038/35057062.

Lemieux Perreault, Louis-Philippe, Marc-André Legault, Géraldine Asselin, and Marie-Pierre Dube. 2016. "Genipe: An Automated Genome-Wide Imputation Pipeline with Automatic Reporting and Statistical Tools." *Bioinformatics* 32 (23). Oxford University Press: 3661–3.

Lewis, Cathryn M, and Jo Knight. 2012. "Introduction to Genetic Association Studies." *Cold Spring Harb Protoc* 2012 (3): 297–306. doi:10.1101/pdb.top068163.

Marchini, Jonathan, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. 2007. "A New Multipoint Method for Genome-Wide Association Studies by Imputation of Genotypes." *Nature Genetics* 39 (7). Nature Publishing Group: 906.

McCarthy, Shane, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, et al. 2016. "A Reference Panel of 64,976 Haplotypes for Genotype Imputation." *Nature Genetics* 48 (10). Nature Publishing Group: 1279.

Mishra, Aniket, and Stuart Macgregor. 2015. "VEGAS2: Software for More Flexible Gene-Based Testing." *Twin Research and Humans Genetics* 18 (1): 86–91. doi:10.1017/thg.2014.79.

Obenchain, Valerie, Michael Lawrence, Vincent Carey, Stephanie Gogarten, Paul Shannon, and Martin Morgan. 2014. "VariantAnnotation: A Bioconductor Package for Exploration and Annotation of Genetic Variants." *Bioinformatics* 30 (14). Oxford University Press: 2076.

Pasquini, Marcelo, Zhiwei Wang, Mary M Horowitz, and Robert Peter Gale. 2013. "2013 Report from the Center for International Blood and Marrow Transplant Research (Cibmtr): Current Uses and Outcomes of Hematopoietic Cell Transplants for Blood and Bone Marrow Disorders." *Clinical Transplants*, 187–97.

Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38 (8): 904–9. doi:10.1038/ng1847.

Pritchard, J K, and M Przeworski. 2001. "Linkage Disequilibrium in Humans: Models and Data." *Am J Hum Genet* 69 (1): 1–14. doi:10.1086/321275.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for

Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81 (3). Elsevier: 559–75.

R Core Team. 2018. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reich, D E, M Cargill, S Bolk, J Ireland, P C Sabeti, D J Richter, T Lavery, et al. 2001. "Linkage Disequilibrium in the Human Genome." *Nature* 411 (6834): 199–204. doi:10.1038/35075590.

Rizvi, Abbas A, Ezgi Karaesmen, Martin Morgan, Leah Preus, Junke Wang, Michael Sovic, Theresa Hahn, and Lara E Sucheston-Campbell. 2018. "Gwasurvivr: An R Package for Genome Wide Survival Analysis." *Bioinformatics*, November. doi:10.1093/bioinformatics/bty920.

Schwarzer, Guido. 2007. "Meta: An R Package for Meta-Analysis." *R News* 7 (3): 40–45.

Shi, Zhen, and John Moult. 2011. "Structural and Functional Impact of Cancer-Related Missense Somatic Mutations." *J Mol Biol* 413 (2): 495–512. doi:10.1016/j.jmb.2011.06.046.

Su, Zhan, Jonathan Marchini, and Peter Donnelly. 2011. "HAPGEN2: Simulation of Multiple Disease Snps." *Bioinformatics* 27 (16). Oxford University Press: 2304–5.

Syed, Hamzah, Andrea L Jorgensen, and Andrew P Morris. 2017. "SurvivalGWAS_SV: Software for the Analysis of Genome-Wide Association Studies of Imputed Genotypes with 'Time-to-Event' Outcomes." *BMC Bioinformatics* 18 (1). BioMed Central: 265.

Therneau, Terry M, and Patricia M Grambsch. 2013. *Modeling Survival Data: Extending the Cox Model*. Springer Science & Business Media.

Timpson, Nicholas J, Celia M T Greenwood, Nicole Soranzo, Daniel J Lawson, and J Brent Richards. 2018. "Genetic Architecture: The Shape of the Genetic Contribution to Human Traits and Disease." *Nat Rev Genet* 19 (2): 110–24. doi:10.1038/nrg.2017.101.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.

Visscher, Peter M, Matthew A Brown, Mark I McCarthy, and Jian Yang.

2012. "Five Years of Gwas Discovery." *Am J Hum Genet* 90 (1): 7–24. doi:10.1016/j.ajhg.2011.11.029.

Willer, Cristen J, Yun Li, and Gonçalo R Abecasis. 2010. "METAL: Fast and Efficient Meta-Analysis of Genomewide Association Scans." *Bioinformatics* 26 (17): 2190–1. doi:10.1093/bioinformatics/btq340.

Yan, Li, Changxing Ma, Dan Wang, Qiang Hu, Maochun Qin, Jeffrey M Conroy, Lara E Sucheston, et al. 2012. "OSAT: A Tool for Sample-to-Batch Allocations in Genomics Experiments." *BMC Genomics* 13 (December): 689. doi:10.1186/1471-2164-13-689.