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, B.S., M.S., M.Sc.

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

Dedicate it.

ACKNOWLEDGEMENTS

0. Lara
1. Ezgi
2. Barbara Foster
3. Martin Morgan
4. Seb
5. Moray
6. Friends
7. Bernie

This work was supported by the NIH/NHLBI R01HL102278 and NIH/NCI R03CA188733.

Vita

2008	Williamsville North High School
2012	SUNY Fredonia
2015	University of Luxembourg
2015	SUNY at Buffalo, Roswell Park Cancer Institute, Graduate Division
2016-present	Graduate Research Associate, Department of Pharmaceutics, The Ohio State University.

Publications

1.
2.

Fields of Study

Major Field: Pharmaceutical Sciences

CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1: Introduction

Genetic Association Studies

Genetic association studies are

Candidate Gene Association Studies

Candidate gene association studies (CGAS)

Genome Wide Association Studies

Genome wide association studies (GWAS)

DISCOVeRY-BMT

Hematopoetic Stem Cell Transplantation

Hematopoietic 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). (Rizvi et al. 2018) 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. 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 MUD-HSCT recipients die within one-year post-transplant due to either their disease or transplant-related causes.2 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.2 Reducing TRM without increasing risk of disease death and vice versa continue to represent a substantial clinical challenge.

Genotyping and Imputation

All samples were genotyped using the Illumina Human OmniExpress BeadChip and the Illumina HumanExome BeadChip (University of Southern California Genomics Facility). 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.74 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. Stringent quality control was performed on both samples and SNPs within this population. Population outliers were removed using EIGENSTRAT75 (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. IMPUTE2 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 <0.00576,77. The recipient-donor mismatch 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 Models

Prior to genetic analyses, clinical covariates for inclusion in genome-wide survival models were selected using bidirectional stepwise Cox proportional hazard models of OS, PFS, TRM and DRM using R statistical software.78 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.79 PFS was defined as the time to disease progression or death. Deaths from TRM and DRM were treated as competing risks and analyzed accordingly.80 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. For each SNP, heterogeneity of effect size estimates between cohorts 1 and 2 was assessed us-

ing p-values from significance tests of heterogeneity (phet) and I2.81 Variants with phet<0.05 and I2>50 were meta-analyzed with a random effects models using meta in R.82

Cox Proportional Hazards Model

Power Calculations

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.

Literature Review

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).1-70 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 unrelated donors with various degrees of HLA-matching and patients treated across multiple decades, from the 1980s through early 2000s.

Reproducing results of previous studies

Results from genetic association studies should be reproduced in independent samples in order to confirm findings.83 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.84 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.84 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.84 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 re-

lated 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.

Our study

Gene-Based Association Testing

VErsatile Gene-based Association Study 2 (VEGAS2) software was used for gene-based association testing.85 VEGAS2 uses 106 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,86 Blood expression quantitative trait loci (eQTL) Browser,87 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 86,89 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.89 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 SIFT90, Mutation Taster91 and PolyPhen-2.92

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 2: gwasurvivr

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 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, et al., 2017), and GWASTools (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, for genome wide survival analyses of imputed data in multiple file formats with flexible analysis and output options.

Data Structure

Gwasurvivr can analyze data in IMPUTE2 format (Howie, et al., 2009), in VCF files derived from Michigan (Das, et al., 2016) or Sanger imputation servers (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 (Gogarten, et al., 2012) or VariantAnnotation (Oben-

chain, 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.

CHAPTER 3: Benchmarking survival model software

`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, et al., 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 imputa-

tion 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: Acute Lymphobastic Leukemia (ALL) GWAS

CHAPTER 5: Conclusion

REFERENCES

Rizvi, Abbas A, Ezgi Karaesmen, Martin Morgan, Leah Preus, Junke Wang, Michael Sovic, and Lara Sucheston-Campbell. 2018. "Gwasurvivr: An R Package for Genome Wide Survival Analysis." *bioRxiv*, May. Cold Spring Harbor Laboratory. doi:10.1101/326033.