Dissecting the Genetic Architecture of Fetal Hemoglobin Expression

by

Aaron Cheng

Submitted in Partial Fulfillment of the Requirements for the M.D. Degree

with Honors in a Special Field at Harvard Medical School

February 2020

Abstract

Inducing production of fetal hemoglobin (HbF) is a promising therapeutic approach to ameliorate disease severity in β-thalassemia and sickle cell disease. While studies have characterized individual genetic factors affecting fetal hemoglobin levels and begun to elucidate some underlying mechanisms, a complete understanding of how these elements interact to influence overall fetal hemoglobin expression levels has yet to be achieved. We hypothesize that varying range of fetal hemoglobin expression in the human population is the result of complex genetic architecture involving the interaction between multiple common and rare genetic variants. To interrogate the underlying genetic architecture of this complex and clinically-relevant trait, we have performed large genome-wide association study (GWAS) from two distinct study populations ascertained in different ways: a Thai population and a Swedish population. We genotyped all samples and implemented standard quality-control measures. From the samples and genotypes that passed quality control, we performed an association study for HbF levels using a linear mixed model instantiated through the BOLT-LMM tool. Our initial results have replicated known loci above genome-wide significance levels, including *BCL11A*, *HBS1L-MYB*, and *HBB*. Moreover, several novel loci and rare variants, including unique structural variants, appear to be present in our study. We are integrating whole genome sequencing on a subset of samples and in general population controls to better define these loci using imputation approaches, and we will account for the aggregate contribution of rare variants with large effects, including the structural variants we have identified. This work has tremendous promise to improve our understanding of how HbF levels can vary in populations, characterize underlying mechanisms by which this clinically important factor

is regulated, and more generally elucidate how a range of allelic variants can collectively contribute to the genetic architecture of a complex trait.

Table of Contents

Acknowledgments

I am extremely grateful to my advisor, Dr. Vijay Sankaran, for providing the opportunity to work with and learn from him, and for providing academic and life advising during the course of my college and medical school experiences. I also appreciate the support of the Howard Hughes Medical Institute Medical Student Research Fellowship for providing funding for my project, a year's worth of incredible scientific experiences, and a community of like-minded individuals pursuing a career in research and medicine. I am deeply indebted to my network of advisors – Dr. David G. Nathan, Dr. Bernard Chang, Dr. Anthony D'Amico, among others – for providing the support and mentorship I needed to clarify my goals and values.

Thank you to Jeffrey Verboon for offering his expertise and guidance on my projects, for reading my thesis, and for providing valuable suggestions and edits to my manuscripts and posters. Thank you to the other members of the Sankaran Lab, members of Vertex Pharmaceuticals, and members of the Steve McCarroll lab for valuable discussions.

Thank you to my friends, with whom I have commiserated and celebrated numerous events during medical school, and with whom I have shared so many of my most memorable moments inside and outside of school.

Finally, I want to extend my deepest gratitude to my parents, without whom I would not have made it this far. Your unending encouragement, emotional support, compassion, and unrelenting sacrifices are the pillars upon which I have built my dreams.

Chapter 1: Introduction

Fetal hemoglobin (HbF) has been an exciting and promising therapeutic target since the discovery that increased HbF expression in patients with β-thalassemia and sickle cell disease (SCD) leads to a milder clinical phenotype. While much work has been done to characterize the genetic factors affecting HbF levels and underlying mechanisms, an integrated understanding of how these elements interact with each other to influence overall HbF expression levels has yet to be achieved. We hypothesize that HbF is the result of complex genetic architecture involving the interaction between multiple common and rare genetic variants. To this end, we have designed a series of experiments and analyses to identify loci that contribute to HbF expression. In this study, I perform common- and rare-variant genetic analysis incorporating blood samples from several study populations. To frame these analyses, I will briefly describe the landscape of research supporting HbF as a potential therapeutic target for hemoglobinopathies and introduce the state of complex trait genetics. This discussion will clarify the importance of dissecting the genetic architecture of HbF to further inform future studies surrounding its use in the clinical setting.

The burden of hemoglobin disorders

Sickle cell disease and β-thalassemia are two prevalent and important diseases of β-hemoglobin. Sickle cell disease is one of the most common monogenic diseases in the world, with an annual incidence of 2,600 in North America alone[1], and is characterized by deformation of red blood cells under stress, which can precipitate events such as hemolysis, vaso-occlusive (pain) crises, and other dangerous clinical consequences[2]. β-

thalassemia is closely related and results from genetic variants which cause a quantitative defect in the β-globin protein; the subsequent imbalance in α- and β-globin causes symptomatic anemia that often requires lifelong blood transfusions[3].

Sickle cell disease and thalassemia are common in developing countries. However, affordable management of these diseases has yet to be established, particularly for those with poor access to medical care; as a result, diseases of hemoglobin are a major cause of morbidity and mortality worldwide – particularly in Africa, the Mediterranean region, and regions in South and East Asia. In fact, a 2001 estimate demonstrated that in Sri Lanka alone, the cost of properly treating all people with thalassemia would account for 10% of the country's healthcare expenditures[4]. Although the economic and health burden of thalassemia and sickle cell disease continues to be understudied, the burden can reasonably be estimated to be similarly large in other developing nations where diseases of hemoglobin are common[5]. Fortunately, growing access to genetic tools and databases has allowed for fruitful study of these disorders. As a result, the past several decades have seen rapid improvement in our understanding of mechanisms of SCD and thalassemia, yielding several promising avenues of gene-targeted treatment.

Preclinical evidence of the therapeutic role for fetal hemoglobin induction

Adult human hemoglobin is formed as a tetramer composed of two α-globin polypeptide chains and two β-like globin polypeptides. Prior to roughly a year of age in humans, β-globin is not yet expressed; fetal hemoglobin instead consists of two α chains and two γ chains (**Figure 1**). The persistence of fetal hemoglobin into adulthood has been shown to

be a key modifier of the major β-hemoglobin disorders – sickle cell disease and β-thalassemia – where it is able to ameliorate symptoms through replacement of the mutated adult β-hemoglobin[6]. Although HbF was found to be highly heritable[7], little was known about its precise genetic modifiers. In late 2007 and 2008, two genome-wide association studies (GWAS) in non-anemic individuals identified three loci associated with variation in fetal hemoglobin levels[8,9]. These loci were also shown to be important in ameliorating the severity of symptoms in patients with sickle cell disease and β-thalassemia[9,10]. Among these was a locus on chromosome 2 within the *BCL11A* gene, which had been well-studied for its role in B lymphopoiesis and neurodevelopment, yet a role in hemoglobin switching had not been appreciated. As a result, initial functional studies revealed a key role for BCL11A in silencing of HbF[11]. In addition, BCL11A was shown to be a critical regulator of fetal hemoglobin switching in humans and mice[12,13]. Recent studies of rare individuals haplo-insufficient for BCL11A have provided additional insights into its critical *in vivo* role in silencing HbF in humans[14,15]. These findings have led to a considerable effort to target BCL11A to achieve HbF induction in patients with the β-hemoglobin disorders.

Considerable efforts have been undertaken over the past decade to elucidate the factors that regulate BCL11A. Studies of BCL11A regulation have revealed its interactions with transcription factors GATA1, SOX6, and ZFPM1/FOG1[11,16]. Furthermore, there appear to be long-range interactions between BCL11A and regions throughout the β-globin locus, which alters the conformation of the locus and its proximity to enhancer regions[17]. Recent research also indicates that BCL11A is regulated at the level of translation by RNA-binding protein LIN28B, illuminating a new mechanism of BCL11A

control[18]. BCL11A has emerged as a particularly interesting target for the development of gene therapies. These efforts include delivery of short-hairpin RNAs (shRNAs) targeting BCL11A and efforts to target an erythroid enhancer of BCL11A using genome editing approaches[19]. However, though numerous factors have been elucidated in the regulation of BCL11A, the relative contribution of genetic variation within each factor to the ultimate expression level of HbF has yet to be clarified. Developing a deeper understanding of the interplay of common and rare variants that affect HbF expression could lead to the identification of more therapeutic targets for gene therapies.

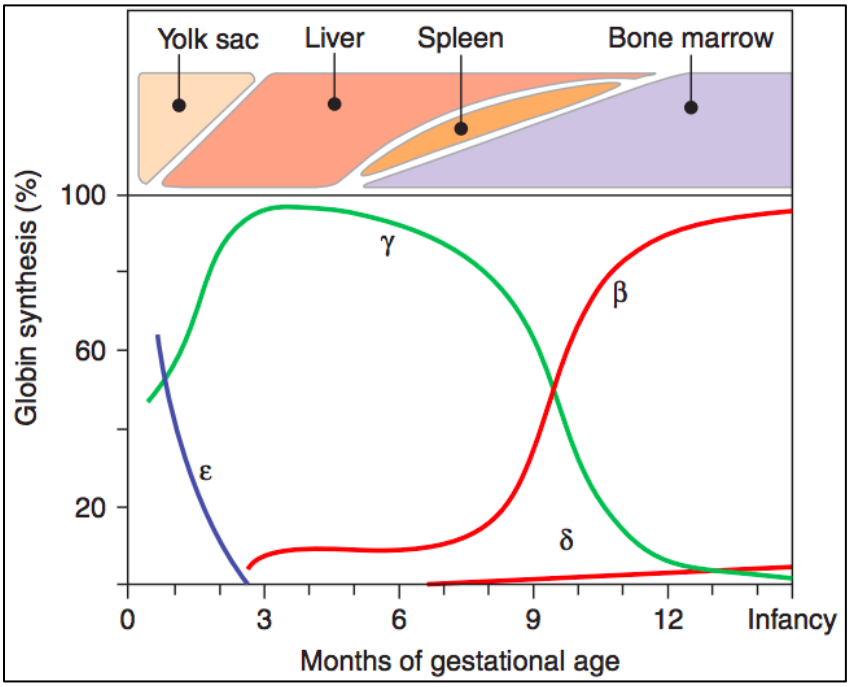

Figure 1. The switch from fetal to adult hemoglobin. The top panel depicts sites of β-like globin production. This figure is adapted from *Sankaran and Orkin (2013)*[20].

Importantly, many of the studies that characterized fetal hemoglobin as a therapeutic target were largely informed by cases of extremely rare variants causing hereditary persistence of fetal hemoglobin (HPFH)[21]. While these studies have been enormously informative to elucidate the underlying mechanisms of fetal hemoglobin

regulation, the complex interplay of common and rare genetic variants – and the effect sizes of genetic contributors on fetal hemoglobin expression – has not yet been clearly elucidated. There remains a great need to apply population-based genetic studies to this trait in order to better understand fetal hemoglobin in the context of its genetic background.

Recent advancements in genetic analysis

There has been a rapid adoption of techniques for genetic analyses in the field of biology. With decreasing costs of genotyping and sequencing technologies, a sudden explosion of freely available genetic information has prompted the development of new tools to interpret large-scale data. Emerging efforts to combine rare and common genetic studies have begun to elucidate a broader understanding of biological systems such as hematopoiesis. The following section provides a brief framework of human genetic studies and how they have the potential to unlock a more holistic understanding of human health and disease.

Human genetic studies can broadly be divided into common (allele frequency > 1%) and rare (allele frequency < 1%) variant association studies, each employing different approaches to work up their variants of interest. Common variant association studies (CVAS) usually take the form of genome-wide association studies (GWAS), in which individuals are genotyped using arrays that capture mostly higher-frequency variants. Statistical analyses can then be used to determine whether each variant is associated with a continuous or binary phenotype of interest. CVAS focus on traits with polygenic architectures comprised of many variants with small individual effects and usually include

a large proportion of healthy individuals in the study population. However, current limitations include high multiple testing burden from evaluating millions of variants, its inability to capture a substantial portion of heritability, and the difficulty of functionally characterizing association signals[22].

Rare variant association studies (RVAS) require alternative analytical methods, since single-variant analysis are underpowered to detect associations if the individual mutation is too rare in the study population. To counteract this, burden tests have been developed, which collapse many variants within a gene or region into a single risk score. This approach thus performs a per-gene or per-region association study as opposed to per-variant association tests in GWAS[23,24].

Importantly, GWAS and RVAS also generally employ different technologies for the identification of genetic variants. GWAS typically employ single nucleotide polymorphism (SNP) arrays to directly genotype up to a few million common variants. Millions of additional variants can then be inferred via imputation, which is the process of using linkage patterns in a more densely sequenced reference panel to predict unobserved genotypes in the study dataset. However, these methods are ineffective for identifying extremely rare variants – especially those in low linkage with other variants – and impossible for novel genetic variants[25]. Therefore, RVAS typically use targeted sequencing, whole-exome sequencing (WES), or whole-genome sequencing (WGS), which allow for unbiased variant calling to identify rare or novel variants that would not have been included on genotyping arrays or that are not confidently imputed[26]. In addition, RVAS study populations are on average smaller than in CVAS and are more enriched for disease cases due to the increased costs associated with these technologies. In addition

to cost, RVAS are limited in that they can miss noncoding associations due to exclusion (WES) or low sequencing depth (WGS), and they require assumptions about the underlying genetic model when aggregating variants[27].

Summary of planned investigations

The goal of this study is to approach HbF expression through the lens of population genetics, using some of the analytic tools mentioned above. Thailand has implemented a screening program given the high prevalence of thalassemia carriers. We have gained access to blood samples from a population of over 86,000 from Thailand. I aim to use genotyping data from this population to identify common genetic variants that may explain variant in HbF expression. Furthermore, we are privileged to have access to a randomly-ascertained cohort of blood samples from blood donors in Sweden, which will be useful to supplement our GWAS and provide further statistical power to our study of HbF. This study will evaluate the relative contributions of both common and rare variants in the final expression level of HbF, and may contribute substantially to our overall understanding of HbF regulation.

As the distribution of HbF in our Thai population suggests that there are rare, large-effect variants that may explain the upper tail of the HbF distribution, we will select a subset of Thai individuals to whole-genome sequence. Subsequent analysis will allow us to examine the presence of variants in the β-globin locus. Gene-based burden testing will allow us to increase power in detecting rare variants found in genes that may be associated with elevated HbF.

These investigations will form a suitable foundation for understanding the complex genetics that influence HbF expression. Eventually, integration of common and rare variant analyses may lead to better predictive tools for clinical outcomes, as well as a method to identify suitable therapeutic targets for patients with hemoglobinopathies.

Chapter 2: Methods

Collection and processing of Thai GWAS data

Collection and genotyping of patient samples

Peripheral blood samples were collected from 1,443 individuals selected from a population of over 86,000 patients seen at the Siriraj Hospital in Bangkok. Specifically, 448 samples were selected from those individuals with HbF levels > 2%, and 995 individuals were selected which had HbF levels measured < 2%. A subset of samples underwent multiplex ligation-dependent probe amplification (MLPA) to test for presence of deletions within the β-globin locus. These samples were subsequently genotyped with the Illumina MEGA array, which calls > 1.7 million variants genome wide. Single nucleotide polymorphisms (SNPs) were called using the Illumina auto call genotype-calling software, then mapped to the GRCh37 human reference genome.

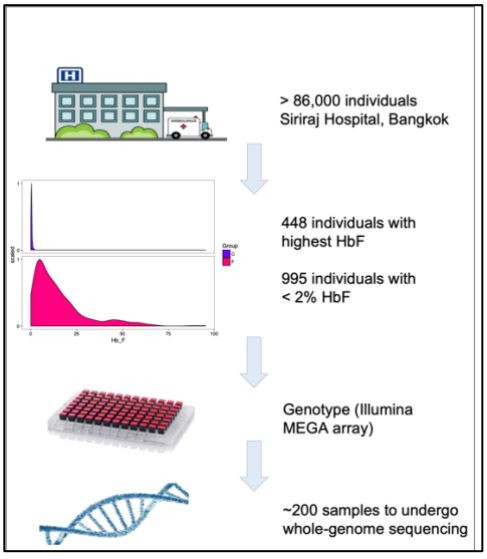

Figure 1. Schematic of Thai GWAS. Of 86,000 healthy individuals with known HbF levels, 1,443 individuals were selected for our study. They were genotyped, and a subset of them were further whole-genome sequenced.

HPFH deletion calling and validation

As only a subset of samples were tested for HBB-spanning deletions, we used the software package pennCNV[28] to both validate MLPA calls and de novo determine whether or not deletions existed in the untested samples. Deletions were called in two ways: de novo and validation-based wherein the known HPFH deletions endemic to the Thai population were validated.

Variant calling and annotation

Following sequencing, reads were aligned to the hg19 reference genome with BWA[29], and GATK v3.2 was applied. Base quality score recalibration, indel realignment, and duplicate removal were applied according to GATK Best Practices recommendations[30,31]. Variant call files (VCF) were processed using Tabix v1.3[32] and Bcftools v1.2[33]. Following variant calling, variants were annotated for functional effect using Variant Effect Predictor v77[34]. Variants were also annotated for allele frequencies from gnomAD v2.0.2[35]. For gnomAD, minor allele frequencies (MAF) from each of the ancestries within gnomAD were added, and MAF filtering was based on the highest MAF from each of the gnomAD populations.

Sample quality control

We used PLINK[36] to compute the proportion of missing genotype calls, sample heterozygosity, and relatedness through identity-by-descent. For each pair of samples with identity-by-descent calculated to be > 0.125, we selected one sample to exclude from

further analysis. We further excluded samples with greater than 2% missingness, as well as those samples with a heterozygosity score > 3 standard deviations from the mean.

Gender determination

We used PLINK to infer sex based on homozygosity across the sex chromosomes. For those samples which did not have confidently called sexes, we used reported data from anonymized patient records to assign the sample sex. We excluded samples with did not have either a PLINK-called sex or a reported sex.

Principal component analysis

We used PLINK to intersect the 1000 genomes[37] and our samples by variant. We filtered for variants that were common to both datasets and defined by the same alternate and reference alleles at identical genome coordinates. We then merged our genotyped data with the 1000 genomes, and pruned the resulting PLINK .bed file. Afterward, we calculated eigenvectors and plotted the first 2 principal components. Visual assessment discerned no obvious outliers; therefore, we did not exclude any samples based on principal component analysis.

SNP quality control

Using PLINK, we calculated minor allele frequency, missingness, and a P-value for Hardy-Weinberg equilibrium. We set a threshold of 5% SNP missingness, and $1x10^{-10}$ for Hardy-Weinberg P-value. We also excluded all but one of each set of duplicate SNPs.

Imputation against Genome Asia Pilot

We decided to use the Genome Asia Pilot[38] as an initial reference panel for imputation. Though we initially considered using the 1000 genomes reference panel (which does include two East Asian populations), Asia has a high degree of population diversity. The Thai population in particularly is poorly covered in the 1000 Genomes Asian populations and has been shown to be better covered by the Genome Asia Pilot reference panel[39]. We used the Michigan Imputation Server[40] to perform our imputation, and Eagle 2.4[41] for haplotype phasing. Imputation results were filtered for r^2 score of 0.3, 0.5, and 0.8.

Genome-wide association studies

Associations were performed using BOLT-LMM[42], an algorithm for mixed model association testing. Covariates included age, age squared, sex, presence of known β-globin deletions, and the top 10 principal components. HbF was normalized using box-cox normalization. Manhattan plots were produced using R and qqman[43], as well as Locuszoom[44].

Whole-genome sequencing

198 samples were selected to undergo whole-genome sequencing for further analyses. Samples were selected by balancing across three priorities: 1) representation of known globin locus deletions, 2) samples with potential novel globin deletions and/or high HbF (>25%), and 3) control samples to build an imputation reference panel for the full array genotype dataset.

Collection and processing of Swedish GWAS data

Collection and genotyping of patient samples

Via collaborators in Sweden, the Sankaran lab has attained access to a cohort of Swedish individuals with available genotyping data and measured HbF levels. The cohort is composed of 4,018 blood donors who were randomly ascertained independently of any phenotype. These donors include participants from a population-based study of immunoglobin levels[45]. Of these donors, 50 individuals were found to have HbF exceeding 2%. Genotyping was performed using the Human-Omni-1 Quad and the InfiniumOmniExpress-24 beadchips, which capture > 1.1 million SNPs per sample.

Pre-GWAS quality control

We used PLINK[36] to compute the proportion of missing genotype calls, sample heterozygosity, and relatedness through identity-by-descent. For each pair of samples with identity-by-descent calculated to be > 0.125, we selected one sample to exclude from further analysis. We further excluded samples with greater than 2% missingness, as well as those samples with a heterozygosity score > 3 standard deviations from the mean. These thresholds are identical to those imposed in the sample quality-control pipeline for our Thai cohort.

Imputation

Genotypes were submitted to the Michigan Imputation Server for imputation[40]. Eagle 2.4[41] was used for haplotype phasing. The 1000 genomes Imputation results were filtered for

r_2 score of 0.3, 0.5, and 0.8. SNPs with minor allele frequencies < 1% were removed from analysis.

Genome-wide association studies

Associations were performed using BOLT-LMM[42], an algorithm for mixed model association testing. Sex was included as a covariate. Phenotypes were normalized using box-cox normalization. Manhattan plots were produced using R and qqman[43]; higher-resolution plots were produced with LocusZoom[44].

GWAS meta-analysis

GWAS summary statistics for Thai and Swedish cohorts were further annotated with sample size of each study (1394 for the Thai study, and 3187 for the Swedish study). Meta-analysis was performed using METAL, a tool commonly employed for combining multiple GWAS using p-values and taking sample size and direction of effect into account[46]. In meta-analyzing the Thai and Swedish cohorts, consistent quality control thresholds, imputation R_2 thresholds, and minor allele frequency cutoffs were applied to each set of GWAS summary statistics prior to utilizing METAL.

Imputation analysis

Analysis of imputation accuracy

To assess the accuracy of imputed genotypes, we treated our genotyping arrays as our truth set. From our quality-controlled genotyping arrays, we randomly removed 10% (or 11,416) of SNPs on chromosome 2 (leaving 102,739 of 114,155 variants). The remaining

90% SNPs were then sent to the Michigan Imputation Server for imputation against several reference panels.

Rare variant gene-based burden tests

Quality control

We used Bcftools to ascertain sample missingness, sample depth of coverage, single nucleotide variants (SNVs) per sample, indels per sample, singletons per sample, and transition/transversion ratios per sample. We used PLINK to determine sample heterozygosity and sample relatedness. We removed outliers of each analysis, defined as those with parameters values that exceeded 3 standard deviations from the mean. Furthermore, we filtered our SNVs to include only those that lie within 1000 base pairs of an exon, for file size management. We then removed variants for which the total depth (represented in a variant call file as the "DP" field) did not exceed 10 in at least 90% of our samples.

Burden analysis

We annotated our VCF with the Variant Effect Predictor (VEP)[34], after which we removed all variants with minor allele frequency of > 1% in any database. We further filtered for only those rare variants whose consequence is a nonsynonymous variant. Next, we performed a Box-Cox normalization of the % HbF phenotype of our samples, controlling for β-globin, sex, age, and top 3 principal components[47]. We additively aggregated all rare variants by gene. We then performed a linear regression over all genes.

Structural variant analysis

Structural variant analysis of whole-genome sequenced samples

We developed an ensemble method of 3 existing tools – Lumpy, MANTA, and Delly – in order to confidently assign structural variants. Briefly, we followed established protocols to run these packages, and then jointly assessed the evidence that a structural variant was called by at least two of these packages and of the same type (i.e. inversion, duplicate, deletion). In particular, flexible endpoints within a 50-bp window were allowed, as each algorithm was found to categorize these slightly differently when unable to identify precise breakpoints. Finally, for the globin locus, HPFH deletions were manually validated and found to be accurate using this ensemble method.

For the gene conversion event, we utilized a tool called Parasol – a paralogue-aware SNP and indel caller – in collaboration with the McCarroll lab.

Chapter 3: Genome-wide association study of fetal hemoglobin expression in a Thai and Swedish population

Abstract

Prior genome-wide association studies have been performed in populations with hemoglobinopathies, revealing several loci that significantly associate with fetal hemoglobin expression. We performed a genome-wide association study in a healthy Thai population in collaboration with the Siriraj Hospital in Bangkok, Thailand, which offers a study population that has not been adequately represented in prior genome-wide association studies of this phenotype. Our results replicated three loci that have been previously identified – namely HBS1L-Myb, BCL11A, and HBB. We furthermore performed a follow-up association study in a healthy Swedish population, which yielded further support for our results, and suggests an additional genome-wide significant locus.

Contributions

Aaron Cheng[1,2], Jeffrey M. Verboon[1,2], Bob Handsaker[2], Steve McCarroll[2], Ellinor Johnsson[3], Bjorn Nilsson[3], Vijay G. Sankaran[1,2]

[1]Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School

[2]Broad Institute of Harvard and MIT

[3]Lund University, Lund, Sweden

ANC performed genome-wide association studies. ANC and JMV performed quality control of Thai and Swedish cohorts. ANC and JMV prepared the manuscript. BH and SM led structural variant analysis. EJ and BN provided Swedish cohort. VGS supervised all aspects of this study.

Note: The authors thank members of the Sankaran laboratory for valuable discussions. Work in our laboratory was supported by the New York Stem Cell Foundation and National Institutes of Health Grant R01 DK103794. VGS is a New York Stem Cell Foundation—Robertson Investigator. ANC received support from the Howard Hughes Medical Institute Medical Fellows Program.

Introduction

Prior genome-wide association studies performed in healthy individuals and patients with sickle cell disease and beta thalassemia have elucidated loci associated with HbF expression[9,10,48–52]. In particular, these GWAS have identified variation within chromosome 11p (*HBB* locus), chromosome 2p (*BCL11A*), and chromosome 6q (*HBS1L-MYB*) as important regulatory regions affecting fetal globin expression and disease severity in sickle cell disease and thalassemia. However, these studies principally investigated genetic variation in African, European, and admixed American populations. We performed a GWAS in 1,443 healthy individuals in Thailand selected from a cohort of over 86,000 individuals to validate and identify new genetic loci associated with HbF levels, and followed up with a GWAS of blood samples collected from healthy Swedish blood donors.

Results

Collection and analysis of Thai and Swedish cohort samples

We collaborated with the Siriraj Hospital in Bangkok, Thailand and screened over 86,000 individuals. Of these individuals, 1,443 samples were selected for use in our GWAS. Specifically, 448 samples were selected from those with the highest HbF levels, while 995 were selected from individuals with HbF levels in the normal range, defined as < 2% HbF as measured by high performance liquid chromatography (HPLC) (**Figure 1**).

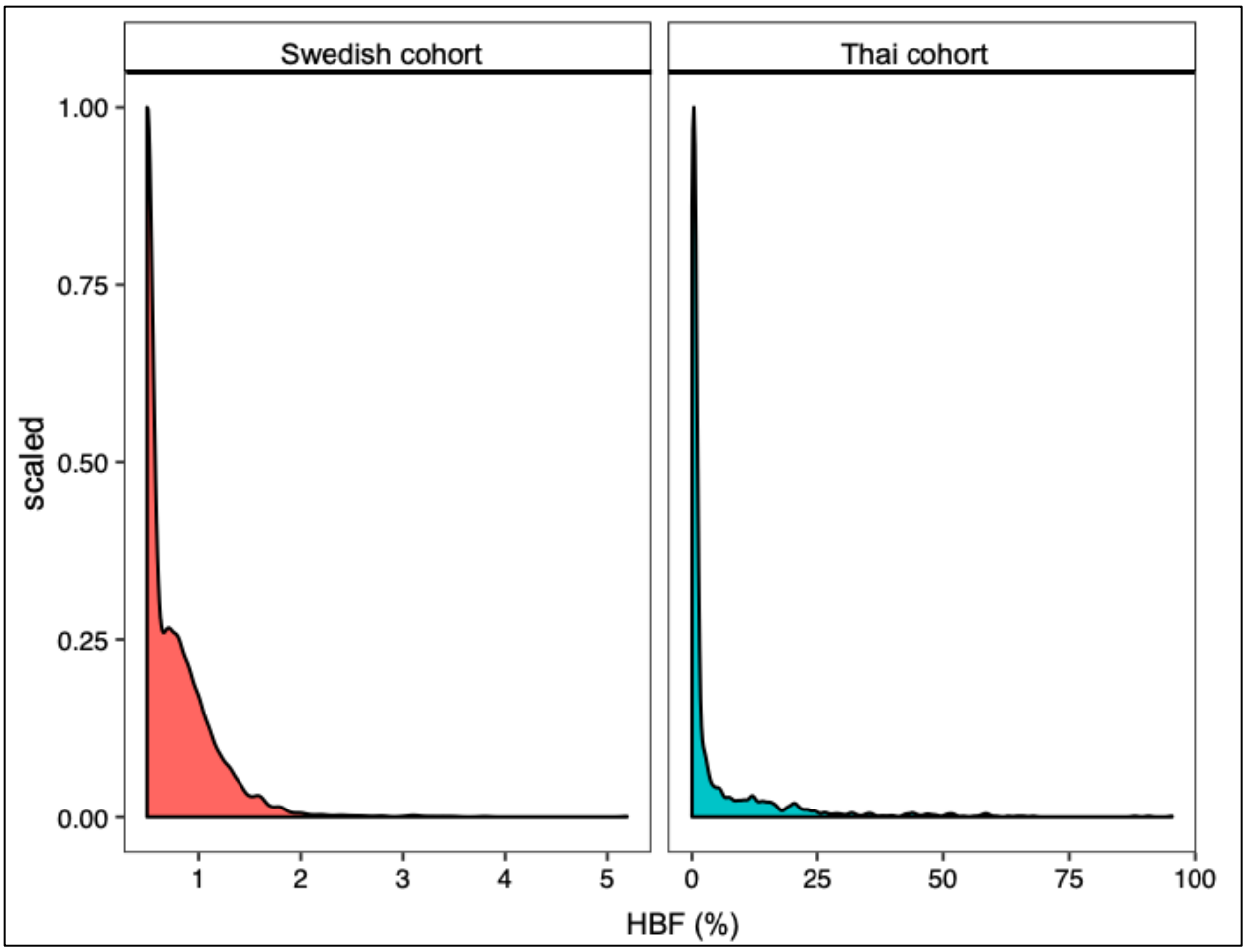

Figure 2. Density plot showing the distribution of HbF levels in Thai and Swedish cohorts. Horizontal axis indicates HbF% measured by HPLC.

A subset of samples underwent multiplex ligation-dependent probe amplification (MLPA) to test for presence of deletions within the β-globin locus (**Figure 2**). Samples were genotyped with the Illumina MEGA array, which calls > 1.7 million variants genome wide. Single nucleotide polymorphisms (SNPs) were called using the Illumina auto call genotype-calling software, then mapped to the GRCh37 human reference genome. MLPA called deletions within the β-globin locus were validated with pennCNV, and untested samples were screened for these deletions as well.

We additionally gained access to a cohort of 4,018 Swedish blood donors who were randomly ascertained independently of any phenotype. These donors included

participants from a study of immunoglobin levels[45]. 50 individuals were found to have HbF levels exceeding 2%, a distribution that is consistent with previously characterized populations which are not under selective pressure (**Figure 1**).

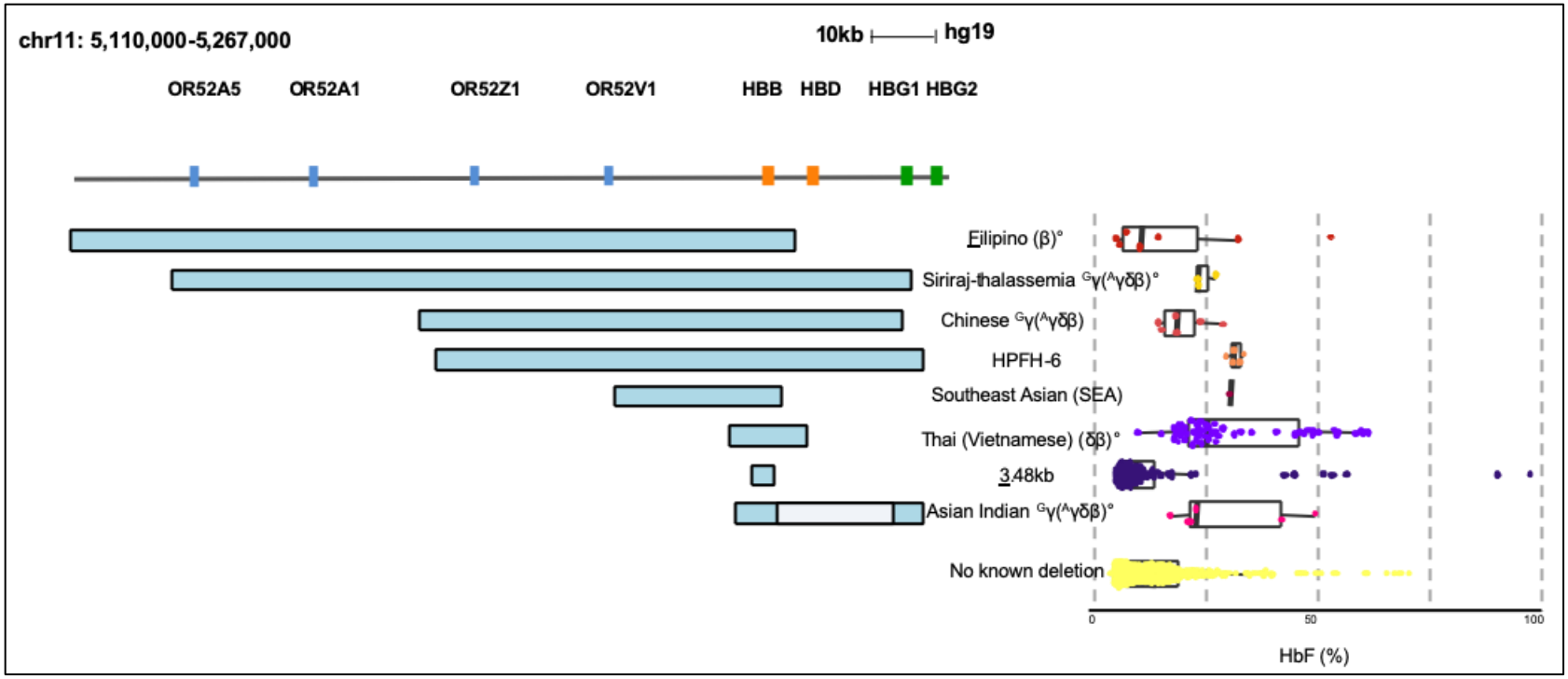

Figure 2. Deletions in Thai cohort. Samples suspected to have possible deletions in β-globin underwent multiplex PCR to identify deletion status for 8 previously characterized deletions in the β-globin locus. Scatterplot and box-and-whisker plot on right indicate HbF distribution in individuals based on their deletion status.

Pre-GWAS sample and SNP quality control of Thai samples

Genotype data of the Thai cohort was processed by filtering SNPs based on missingness (proportion of individuals missing genotype information at that site) and Hardy-Weinberg equilibrium p-value. A kinship coefficient was calculated for each pair of individuals, and one of each pair whose identity-by-descent (IBD) coefficient exceeded 0.125 was excluded from analysis. Finally, population substructure was estimated through principal component analysis projected onto the 1000 Genomes Project; there were no outliers, so no further individuals were removed based on principal component analysis (**Figure 3**). We only included autosomal content (chromosomes 1-22) for further analysis.

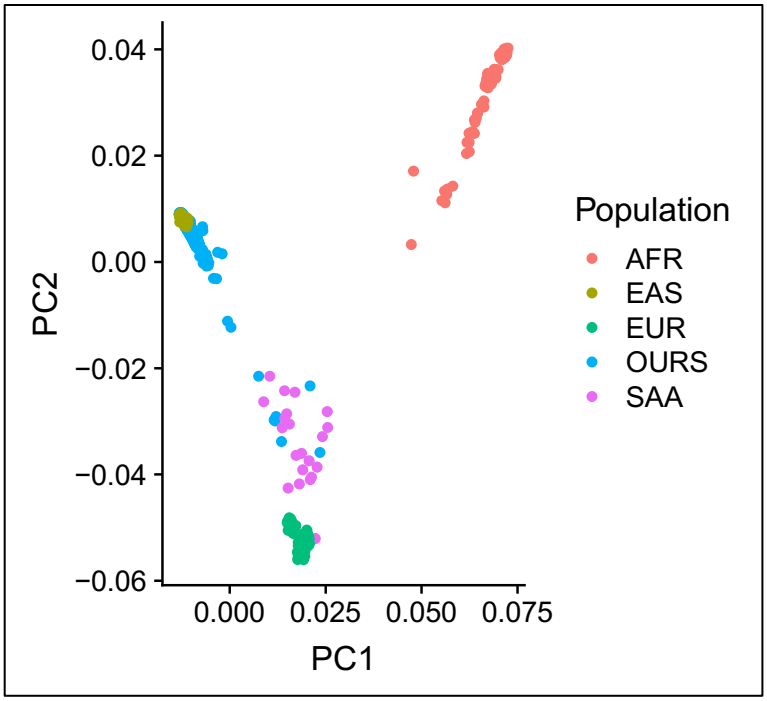

Figure 3. Principal component analysis of Thai cohort against 1000 Genomes. Individuals in Thai cohort are blue. AFR: African, EAS: East Asian, EUR: European, SAA: South Asian, OURS: our samples

Imputation was subsequently conducted using the Michigan Imputation Server. Specifically, Eagle v2.3 was used to phase input haplotypes, and Minimac4 was used for imputation. We used the Genome Asia Pilot reference panel for imputation; this reference

panel has been shown to enable genetic discoveries in specific Asian populations that have historically been underrepresented in genetic studies[53] (see **chapter 4** for a deeper discussion on population representation in genetic studies). Imputed SNPs were subsequently filtered for $R_2 > 0.8$ to prune low-confidence calls.

Genome-wide association study of Thai cohort

BOLT-LMM was used to perform mixed model association testing, and accounted for age, sex, top 10 principal components, and presence of known β-globin locus deletions as determined by MLPA in the covariates. The resulting association statistics were filtered for SNPs with minor allele frequency (MAF) > 0.01, then plotted (**Figure 4**). Peaks exceeded the genome-wide significance threshold ($p < 5x10^{-8}$) at three loci: chromosome 2, chromosome 6, and chromosome 11. The genomic inflation factor (λ_{GC}) was 9.979515e-01, indicating no evidence of overdispersion.

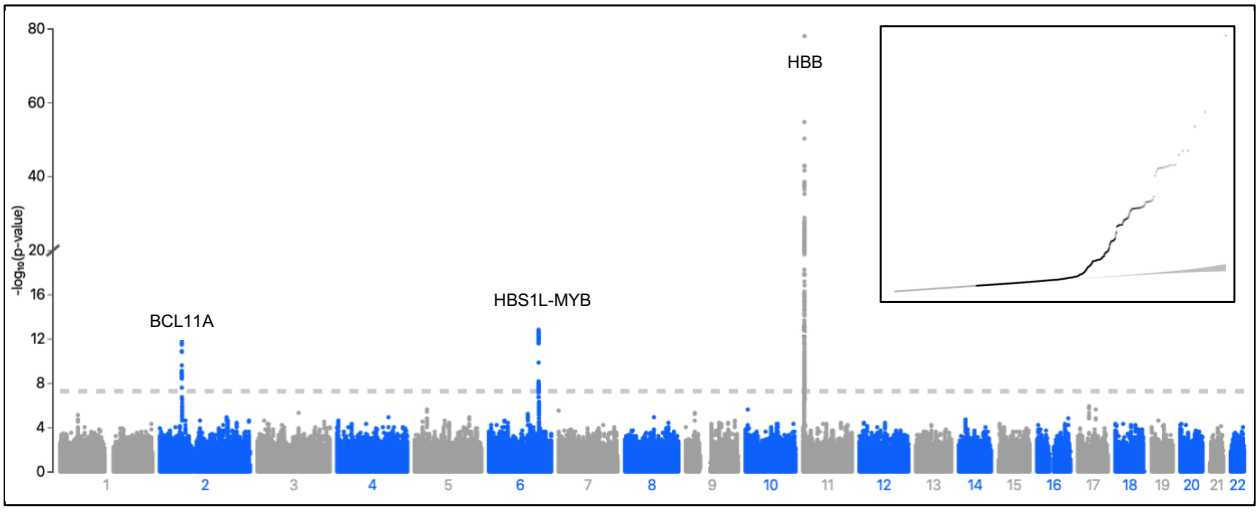

Figure 4. Manhattan plot of Thai GWAS. Dashed line: $p = 5x10^{-8}$. Peaks exceeding this line indicate genome-wide significance of association. Q-Q plot shown in upper right, with grey line indicating expected chi square distribution.

The sentinel SNPs at each peak are described in **Table 1**. All three sentinel SNPs lie within genes that have previously been identified in GWAS of distinct study populations to associate with HbF expression.

Table 1. Sentinel SNPs in Thai GWAS. BP: base position, Ref: reference allele, Alt: alternate allele, Ref allele freq: reference allele frequency, S.E.: standard error

Chromosome and locus	BP	Ref	Alt	Ref allele freq	Beta	S.E.	p-value
2 (*BCL11A*)	60713235	A	G	0.760981	-0.43867	0.0785655	2.4×10^{-8}
6 (*HBS1L-MYB*)	135450755	T	C	0.415366	-0.393298	0.0715301	3.8×10^{-8}
11 (*HBB*)	5525654	C	T	0.408265	0.388278	0.0711725	4.9×10^{-8}

Genome-wide association study in Swedish cohort

After applying similar pre-imputation quality control steps to the Swedish cohort (as described in the **Methods** section), we imputed the genotypes to the 1000 genomes using the Michigan Imputation Server. Eagle v2.3 was used to phase input haplotypes, and Minimac4 was used for imputation. Afterward, imputed results were filtered for those with $R_2 > 0.8$. We used BOLT-LMM to perform association testing on the remaining SNPs, with only sex as a covariate.

The resulting association study identified four peaks which exceeded genome-wide significance ($p = 5 \times 10^{-8}$) (**Figure 5**). The genomic inflation factor (λ_{GC}) was 1.0, indicating no evidence of overdispersion. Three of the four peaks are consistent with

previously validated loci and correspond to variants within *BCL11A*, *HBS1L-MYB*, and *HBB* (**Table 2**). One additional peak was identified at chromosome 15 with a sentinel SNP at position 98,190,432 ($-\log_{10}p = 7.481$) (**Figure 6**).

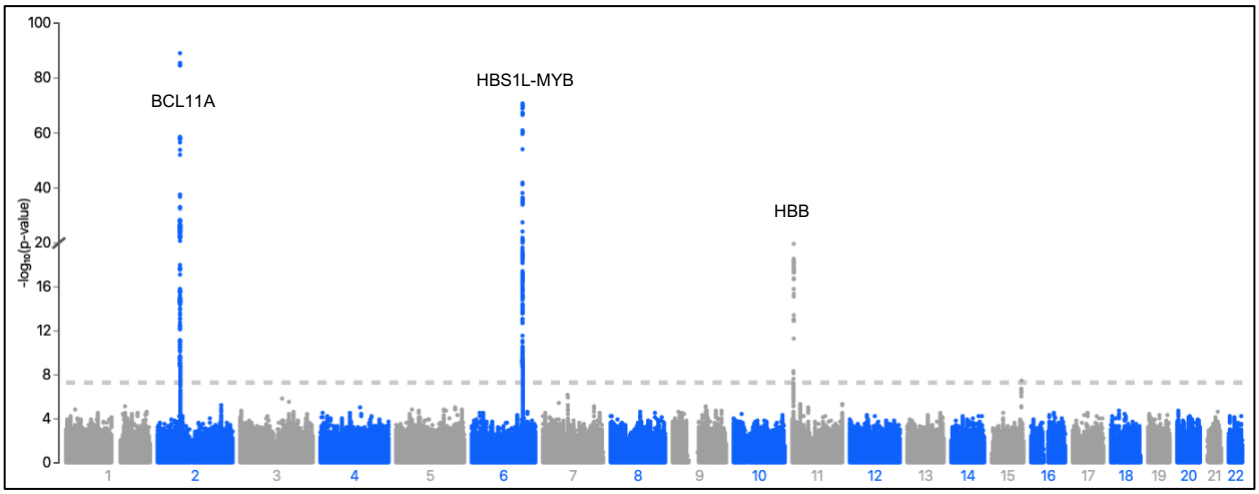

Figure 5. Manhattan plot of Swedish GWAS. Dashed line: $p = 5x10^{-8}$; peaks exceeding this line indicate genome-wide significance of association.

Table 2. Sentinel SNPs in Swedish GWAS. BP: base position, Ref: reference allele, Alt: alternate allele, Ref allele freq: reference allele frequency, S.E.: standard error

Chromosome and locus	BP	Ref	Alt	Ref allele freq	Beta	S.E.	p-value
2 (*BCL11A*)	60718043	T	G	0.848386	-0.258141	0.0128474	$8.5x10^{-90}$
6 (*HBS1L-MYB*)	135418916	A	G	0.267436	0.190469	0.0106551	$1.8x10^{-71}$
11 (*HBB*)	5271063	C	T	0.295538	0.0958744	0.0102938	$1.2x10^{-20}$
15 (*ARRDC4*)	98190432	T	C	0.161512	0.0715621	0.0129481	$3.3x10^{-8}$

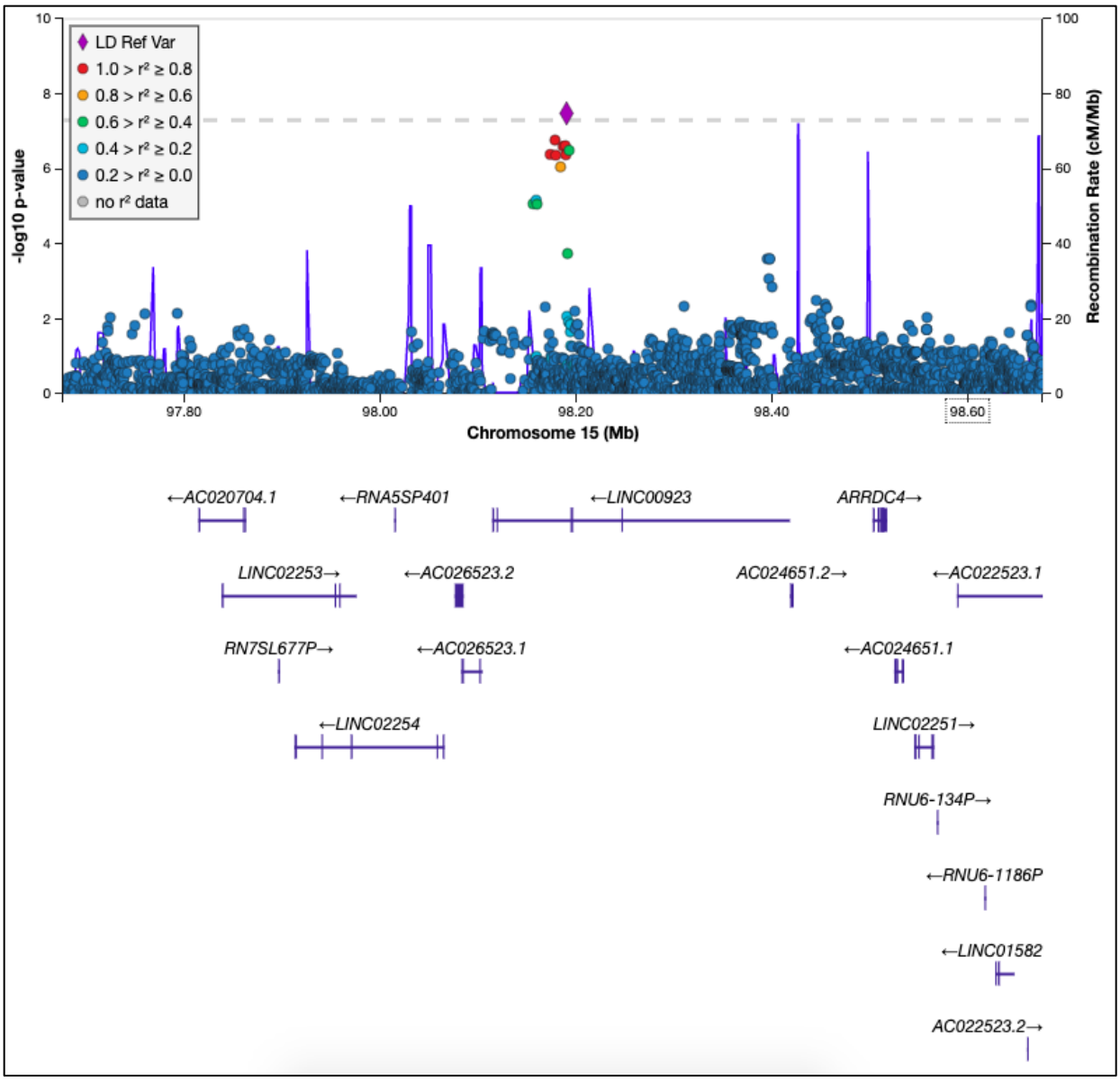

Figure 6. *Close-up of chromosome 15 locus. Coordinates are in hg19. LD information uses all subpopulations of 1000 genomes.*

Finally, we performed a GWAS meta-analysis using METAL, a tool commonly used to combine summary statistics of multiple common variant studies. Meta-analyses are a useful tool to increase sample size by incorporating samples across numerous genetic studies; this increases statistical power and often allows the detection of associations with

smaller effect sizes[54]. This approach is particularly useful in investigating complex traits that are likely to be influenced by multiple variants of small effect size. Furthermore, meta-analyses can be useful in evaluating the consistency or heterogeneity of results across multiple datasets[55].

We took care to apply identical quality control thresholds to the Thai and Swedish populations; the only differences in the two studies was the reference panel used during imputation, and the covariates used during association. The meta-analysis replicated the three loci at chromosomes 2, 6, and 11 as previously identified in the Thai and Swedish GWAS; the chromosome 15 locus fell under the genome-wide significance line but remained suggestive of a significant association ($p = 6.5 \times 10^{-7}$) (**Figure 7**).

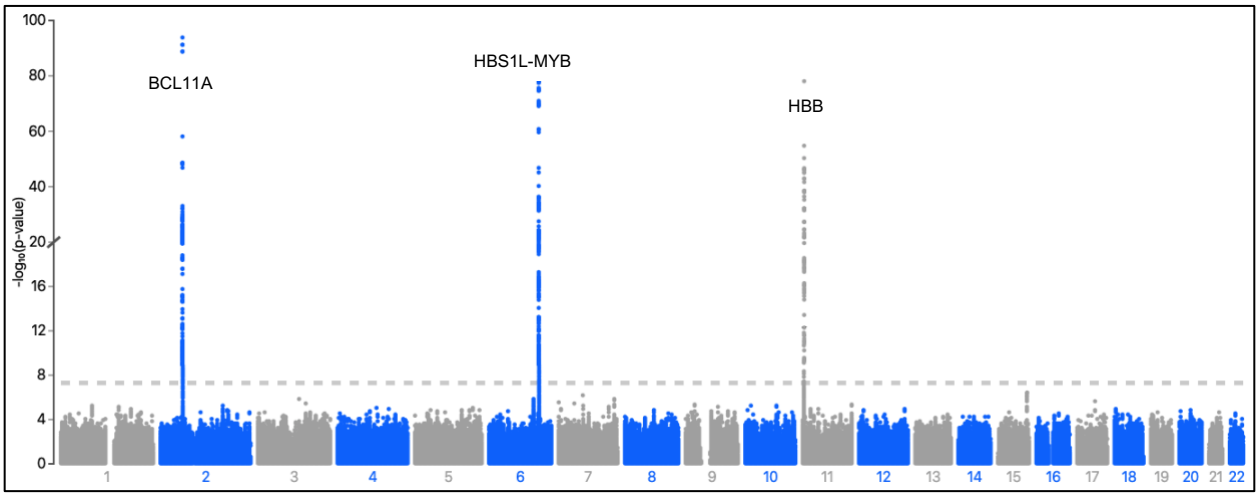

Figure 7. Meta-analysis of Thai and Swedish cohorts. Dashed line indicates genome-wide significance at $p = 5 \times 10^{-8}$.

Discussion

This collection of genome-wide association studies constitutes one of the largest genetic studies of HbF to date. The GWAS of a healthy Thai population is particularly valuable as it incorporates a population that has been classically underrepresented in population-based genetic studies. The combination of the Thai and Swedish cohort allows for a study that benefits from the novelty of understudied populations and the statistical power conferred by including a large number of individuals.

In this study, we collaborated closely with the Siriraj Hospital in Bangkok to identify a cohort of 1,458 adults for a GWAS. These individuals were carefully chosen to include people found to have abnormally high levels of HbF in adulthood (defined as those with HbF of > 2%) and a 2:1 ratio of controls to cases. After these individuals were identified, quality control was carefully performed to include only SNPs and samples that passed stringent quality thresholds. These steps yielded a high-quality GWAS with minimal genomic control inflation, suggesting that we properly accounted for confounding from population structure.

Each sample within our Thai cohort underwent analysis for 8 known deletions within the β-globin locus. Because these deletions are known to influence HbF levels, we included each deletion as a binary covariate in our model, along with standard covariates (age, age squared, sex, and top 10 principal components). Our common variant analysis study replicated genome-wide significant associations at three loci – BCL11A, HBS1L-MYB, and HBB – and did not identify any additional signals. It is possible that our study was underpowered to detect weaker but significant signals, or that the reference panel used for imputation could be further improved to better represent our study sample. It is

also possible that conditional analysis at the three significant loci may reveal independent signals.

We followed up our Thai study with a GWAS using over 3,800 Swedish blood samples, which have had HbF levels measured. The absence of other phenotypic information limited the number of covariates we could include in our analysis; the association only included sex as an additional covariate. The resulting GWAS again replicated similar findings, but also demonstrated a small but genome-wide significant peak at chromosome 15, suggesting a new locus that may be relevant in HbF expression. Exploration of this locus revealed that the cluster of significant SNPs lie within a region encoding a long intergenic non-coding RNA (lncRNA).

Finally, we used METAL to conduct a meta-analysis of the two GWAS. We ensured that the SNPs and samples in each analysis were subject to the same quality-control thresholds, and that post-imputation results were filtered using the same R_2 and MAF, to ensure consistency across studies. In our meta-analysis, the three aforementioned peaks – at BCL11A, HBS1L-MYB, and HBB – were replicated, though the chromosome 15 peak no longer reached genome-wide significance. It is possible that this peak confers a population-specific effect on HbF that is unseen in the Thai population; combining studies may thus suppress a GWAS peak that otherwise would reach significance in only a specific population. It is also possible that the Thai cohort is too small to adequately capture new signals, or that improvements in covariates may reveal additional loci that have not yet been seen.

In summary, our GWAS have replicated three loci previously known to associate significantly with HbF levels. Additionally, a new signal at chromosome 15 has been

identified in a healthy Swedish population, suggesting a possible new factor involved in HbF regulation.

Future directions

To more completely assess the GWAS results in this study, we will perform conditional analyses to identify independently associated variants[56]. Furthermore, we aim to incorporate summary statistics from other HbF GWAS to better power our meta-analyses, as variants with low effect sizes require much larger sample sizes to reveal. Finally, we aim to identify causal variants using statistical fine-mapping methods in order to prioritize SNPs for further study.

Chapter 4: Assessing the influence of reference panels on imputation and genome-wide association studies

Abstract

Imputation refers to the process of utilizing linkage patterns in a more densely sequenced reference panel to predict unobserved genotypes in the study dataset. In performing the genome-wide association studies described previously, we had to make careful decisions about the reference panels we used to perform our genetic imputations. Because Asians have historically been poorly represented in whole-genome sequencing databases, imputation of our Thai cohort against a reference panel with poor Asian representation risks missing associations that can only be elucidated by using population-specific linkage information. In this chapter, I discuss the process of selecting and evaluating reference panels and measuring their effects on GWAS results.

Contributions

Aaron Cheng[1,2], Jeffrey M. Verboon[1,2], Vijay G. Sankaran[1,2]

[1]Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School

[2]Broad Institute of Harvard and MIT

ANC performed genome-wide association studies, imputation, and imputation quality assessments. ANC prepared the chapter. VGS supervised all aspects of this study.

Note: The authors thank members of the Sankaran laboratory for valuable discussions. Work in our laboratory was supported by the New York Stem Cell Foundation and National Institutes of Health Grant R01 DK103794. VGS is a New York Stem Cell Foundation—Robertson Investigator. ANC received support from the Howard Hughes Medical Institute Medical Fellows Program.

Introduction

Imputation is a statistical technique which uses haplotype patterns in a reference panel to predict unobserved genotypes in a study dataset. Recent technological advances have allowed for denser imputation reference panels and more accurate imputation algorithms. Larger reference panels such as the UK10K Project, Haplotype Reference Consortium (HRC), and TOPMed feature substantially more variants than older panels, and include far more rare variants[57–59]. Imputation is particularly useful for association studies because typical GWAS contain incomplete genetic information. As an example, the genotyping arrays used in the prior chapter contained information at just over 1 million variants. For the vast majority of SNPs, observations exist for the reference panel (which are typically whole-genome or whole-exome sequenced and thus contain far more genetic information). Using the pattern of linkage disequilibrium (LD), one can impute missing genotypes, and use these imputed genotypes in association tests to improve statistical power in a cost-efficient manner.

Until now, most GWAS have been performed in populations that are well represented by reference panels. However, understudied populations – such as Asia and Africa – pose an important problem in GWAS[39]. Several studies have demonstrated that inclusion of specific populations in reference panels can increase the imputation accuracy[60–62]. In particular, Asian populations are characterized by high levels of population-specific variation which is incompletely captured by preexisting reference panels such as the 1000 genomes. Similar problems exist for other ethnicities as well. The Genome Asia Pilot study has also revealed that allele frequencies of variants differ between populations; by using population-specific allele frequencies that better represent

the East Asian population structure, a greater number of variants were correctly identified as pathogenic when using Asian allele frequencies for cutoffs instead of the 1000 genomes and other preexisting databases[53]. Furthermore, more inclusive panels are particularly useful in imputing variants with lower minor allele frequencies. In some cases, ancestry-specific associations have revealed new genetic associations that were previously unknown[63,64].

In performing our GWAS, we had to make careful choices about the reference panels we used during genotype imputation. In this chapter, I discuss our rationale for selecting reference panels, as well as several methods for evaluating imputation quality.

Results

The GenomeAsia Pilot dataset

The GenomeAsia consortium has released a valuable dataset consisting of 1,739 individuals from 219 population groups across Asia; these samples have been specifically chosen to emphasize population groups that are underrepresented in previous genetic studies[53]. In deciding which reference panels to use for our GWAS, we devised a method of comparing imputation accuracy with different reference panels.

Using chromosome 2, we randomly removed 10% of SNPs (after pre-GWAS quality control as described in Methods) on the genotyping array using PLINK. The resulting 90% of SNPs were then submitted to the Michigan Imputation Server where Eagle v2.4 and Minimac 4 were used for haplotype inference and imputation, respectively. We performed this analysis using the 1000 genomes as well as the GAsP dataset as reference panels.

Afterward, we compared the resulting imputed genotypes with the SNPs that were removed from the earlier step. For each sample, accuracy of imputation was calculated with the following formula: $\frac{\text{\# SNPs with correct dosage}}{\text{\# total removed SNPs}}$. The resulting analysis demonstrated that for 99.85% (or 1392 out of 1394) of samples, imputation accuracy was higher using the GAsP reference panel when compared to the 1000 genomes (**Figure 1A,B**).

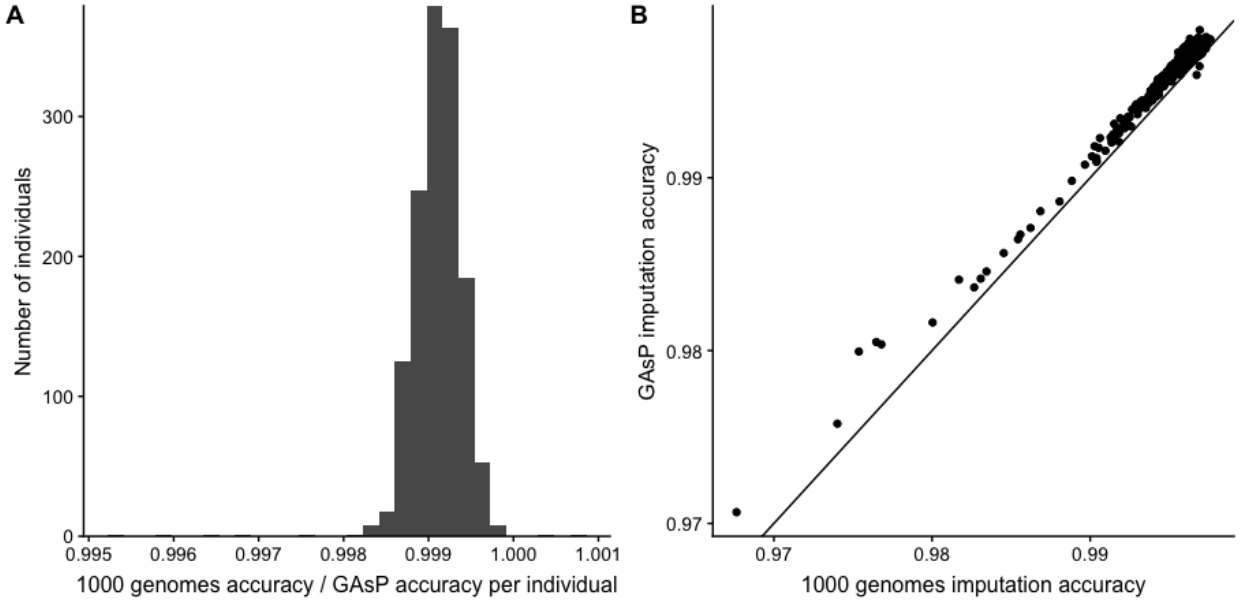

Figure 1. Schematic of imputation accuracy. **A)** Histogram of individuals versus 1000 genomes:GAsP imputation accuracy ratio. <1 indicates greater accuracy in GAsP reference panel. **B)** Scatterplot of 1000 genomes imputation accuracy versus GAsP imputation accuracy. Each dot corresponds to an individual sample. Points above solid line indicate greater accuracy after imputation to GAsP reference panel compared to 1000 genomes reference panel.

Merging reference panels

Prior to the release of the GAsP, we employed a different strategy for evaluating imputation quality. We selected 198 individuals in our Thai cohort to undergo whole-genome sequencing in order to build a panel of controls, verify β-globin deletions, and identify new variants that may explain elevated HbF levels in individuals without a known deletion in the β-globin locus.

Recognizing that Asia is the most populous continent with considerable population substructure, we decided to combine our whole-genome sequenced samples with the 1000 genomes reference panel in order to make a more representative reference panel for the remainder of our genotyped samples. Indeed, this strategy has been utilized in studies of samples with high degrees of relatedness; this allows low-coverage sequencing data to serve as a high-density reference for individuals with similar haplotypes that do not have any sequence data[65,66]. Furthermore, merging reference panels has been shown to improve the accuracy of imputation, particularly when reference panels are created from multiple different populations[67].

We used a tool called IMPUTE2, which allows merging multiple reference panels in addition to performing genotype imputation[68]. Briefly, this method identifies variants unique to reference panel 1, imputes these genotypes in reference panel 2, then performs the same procedure for reference panel 2, thereby creating a large reference panel composed of the sum of samples in both panels, which is equally dense across all samples (**Figure 2**).

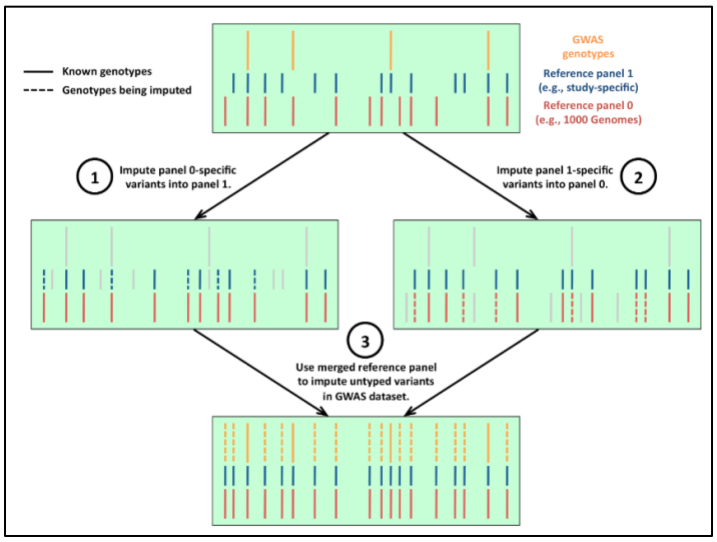

Figure 2. Schematic of merging reference panels using IMPUTE2, taken from https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#merging_panels. SNPs from one reference panel are imputed onto the other reference panel; this is then repeated in the reverse direction to create a combined reference panel that is composed of more SNPs and samples than either one panel alone.

We then performed genome-wide association studies instantiated through BOLT-LMM, using the same covariates and settings for each reference panel, to compare the outcomes of each association. Each association study replicated 3 genome-wide significant peaks at identical loci (at chromosome 2, 6, and 11) (**Table 1**). At two of the loci, the p-value of sentinel SNPs was lower using the merged reference panel compared to using either the 1000 genomes or the 198 whole-genome sequenced Thai samples in isolation.

Table 1. Comparison of p-values at sentinel SNPs of GWAS performed with different reference panels. The three reference panels used were the 1000 genomes, our study's 198 whole-genome sequenced individuals, and the combination of the two sets of genomes.

Chromosome	1000 genomes minimum p-value	WGS minimum p-value	Merged panel minimum p-value
2	9.2×10^{-13}	2.90×10^{-13}	3.1×10^{-12}
6	3.0×10^{-14}	3.10×10^{-15}	3.5×10^{-16}
11	6.80×10^{-69}	6.70×10^{-83}	1.1×10^{-83}

Discussion

In performing our GWAS with Thai samples, we needed to make careful decisions about the reference panels we selected for our analysis. Initially, prior to the 2019 release of the GAsP reference panel, we used SHAPEIT[69] for haplotype inference and IMPUTE2 for imputation. We evaluated the impact of three reference panels on our GWAS: 1) the 1000 genomes, 2) a subset of 198 Thai samples which were whole-genome sequenced, and 3) a reference panel composed of both the 1000 genomes and our whole-genome sequenced Thai samples. For the three loci that are genome-wide significant, we discovered that the reference panel composed of our Thai samples resulted in greater statistical significance at the sentinel SNPs compared to the 1000 genomes. Furthermore, the combined reference panel yielded greater statistical significance at two of the three loci compared to either reference panel alone. It is important to note that the lower p-values at these sentinel SNPs should not be interpreted as improved imputation accuracy; however, as these loci are known to be important modulators of HbF in prior studies, the improved statistical significance at the sentinel SNPs is suggestive that a combined

reference panel may be more appropriate in detecting significant associations compared to either reference panel alone.

In 2019, the Genome Asia Pilot reference panel was released. The reference panel includes a much higher density of underrepresented Asian populations compared to previous panels. As a result, we decided to use the GAsP for our Thai cohort. We used the Michigan Imputation Server in order to perform haplotype inference and imputation, and compared the imputation quality using the 1000 genomes and GAsP as reference panels. For this measurement, we removed 10% of our sample genotypes (which we treated as the truth set), and used the remaining SNPs for imputation. We then compared the resulting imputed genotypes against the truth set and found that for all individuals except two, using the GAsP as a reference panel yielded greater accuracy in imputed results compared to using the 1000 genomes. These results suggest that selection of a reference panel with higher representation of individuals that are ethnically similar to the study samples allows for greater accuracy in imputed genotypes.

Future directions

Analyses of imputed results are ongoing. In comparing the 1000 genomes and GAsP as imputation reference panels, it will be important to investigate whether imputation quality is affected by minor allele frequency. Furthermore, we are in the process of merging the GAsP with 1000 genomes and with our whole-genome sequenced samples to produce a larger and denser reference panel. It will be interesting to determine whether using this new reference panel will yield even higher accuracy, and whether this will be important in the GWAS results.

We have also considered the strategy of using our subset of whole-genome sequenced individuals as the “truth set” of genotypes when calculating imputation accuracy. However, we had difficulties in performing the conversion from hg19 to hg38 (a necessary step as our genotyped samples were called in hg19 coordinates, while our WGS samples were called in hg38 samples) given a large rate of SNP dropout and mis-mapping. We will continue to investigate this method as a potential method of measuring imputation quality.

Chapter 5: Rare variant analysis and identification of novel deletion in Thai cohort

Abstract

The distribution of HbF in our Thai cohort (described in Chapter 1) suggested that there were rare large-effect variants that could be treated as covariates in our GWAS. To further characterize these variants, we selected a set of 198 individuals to undergo whole-genome sequencing, as described in Chapter 4. These sequenced samples permitted us to perform a rare variant burden test, in order to identify genes that may relate to HbF expression. Our burden test identified HBB as the top candidate. Careful examination of the HBB locus suggested that HBE is a suitable covariate to include in our GWAS, as HBE is known to modulate HbF levels. Furthermore, we identified one novel conversion event involving HBG1 and HBG2 which appears to influence HbF expression as well.

Contributions

Aaron N. Cheng[1,2], Jeffrey M. Verboon[1,2], Bob Handsaker[2], Steve McCarroll[2], Vijay G. Sankaran[1,2]

[1]Division of Hematology/Oncology, Boston Children's Hospital, Harvard Medical School and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School

[2]Broad Institute of Harvard and MIT

ANC performed rare variant burden tests. JMV performed structural variant analysis and characterization of novel structural variant. BH and SM provided structural variant analysis and tools to characterize the gene conversion event. VGS supervised all aspects of this study.

Note: The authors thank members of the Sankaran laboratory for valuable discussions. Work in our laboratory was supported by the New York Stem Cell Foundation and National Institutes of Health Grant R01 DK103794. VGS is a New York Stem Cell Foundation—Robertson Investigator. ANC received support from the Howard Hughes Medical Institute Medical Fellows Program.

Introduction

While genome-wide association studies are appropriate tools for investigating the contributions of common variants to a phenotype of interest, rare variant studies require a different approach. In this chapter, we introduce our implementation of a rare variant burden test as applied to our Thai cohort. A rare variant burden test typically involves higher-density sequencing (such as whole exome sequencing or whole genome sequencing) to call extremely rare variants (typically defined as < 1% minor allele frequency) with higher confidence. A "collapsing function" – which aggregates rare variants by gene or by region – is then applied to the genetic data. A per-gene (or per-region) association test can then be performed, as compared to per-variant association tests in GWAS. We implemented a rare variant burden test on our cohort of 198 sequenced Thai samples, revealing HBB to be significantly associated with HbF expression. Close inspection of the HBB locus identified two variants – HBE and a novel structural variant – which appear to influence HbF levels.

Results

Rare variant burden test in Thai cohort

As described in Chapter 4, we whole-genome sequenced a subset of 198 of our Thai samples. We then annotated all variants using Variant Effect Predictor (VEP)[34], after which we removed all variants with minor allele frequency of > 1% in any database. We further filtered for only those rare variants whose consequence is a nonsynonymous variant. Next, we performed a Box-Cox normalization of the % HbF phenotype of our samples, controlling for β-globin, sex, age, and top 3 principal components[47]. We

additively aggregated all rare variants by gene. We then performed a linear regression over all genes. The resulting association identified HBB as the only gene whose burden of rare variants was significantly associated with HbF expression after Bonferroni adjustment (**Table 1**).

Gene	**Beta**	**p-value (unadjusted)**
HBB	1.48	3.65×10^{-12}
HBG1	-0.81	1.61×10^{-4}
NBPF10	0.17	4.07×10^{-4}
GUF1	-3.21	4.20×10^{-4}
GLB1	-3.64	8.57×10^{-4}
SMARCA4	2.46	1.22×10^{-3}
ZNF585B	-1.02	1.85×10^{-3}

HBE and its influence of HbF expression

One of the top candidate SNPs in the β-globin locus is the Hemoglobin E variant, an abnormal HBB gene carrying a single missense mutation. We find that this variant confers increased fetal hemoglobin, both in the presence and absence of an additional HBB-spanning deletion (**Figure 1**). Of note, without measurements of the absolute measurements of hemoglobin in grams per deciliter (g/dL) it is unclear if these increases in fetal hemoglobin are due to increases in γ-globin production, impaired β-globin production, or a combination of both.

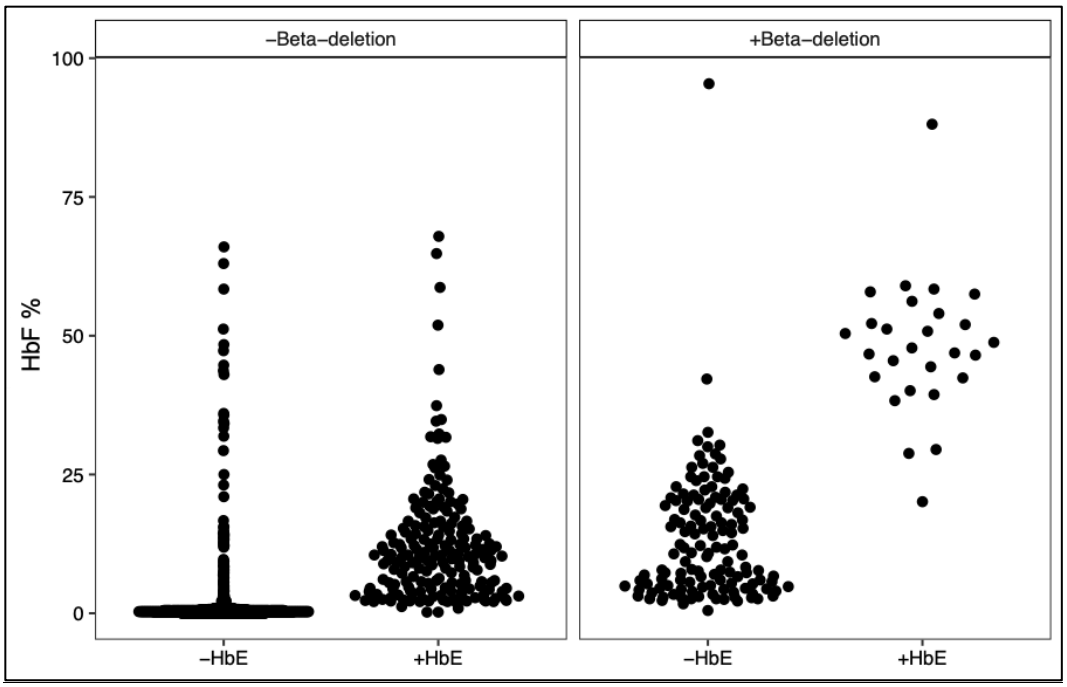

Figure 1. The presence of the HbE SNP influences HbF expression, in both samples with and without known β-globin deletions. Left panel shows samples without a β-globin deletion; right panel shows samples with β-globin deletion.

Whole-genome sequencing reveals a gene conversion event involving HBG1 and HBG2

In investigating the HBB locus, we identified an interesting pattern of decreasing coverage within the HBG1 and concurrent increase in HBG2 that occurred in 62 of 157 (41.9%) analyzed whole-genome sequenced Thai individuals **(Figure 2A**). In order to better understand this locus, we collaborated with the McCarroll lab which has developed a new tool called Parasol, a copy-number- and paralog- aware SNP and indel variant caller. Using this tool, we determined that samples with the pattern of coverage in HBG1 and HBG2 (**Figure 1**) actually demonstrate a gene conversion event wherein HBG1 is being converted to HBG2. Importantly, while this gene conversion is correlated with increased fetal hemoglobin, this relationship only occurs in genetic background of individuals who also have a beta globin deletion (**Figure 2B**).

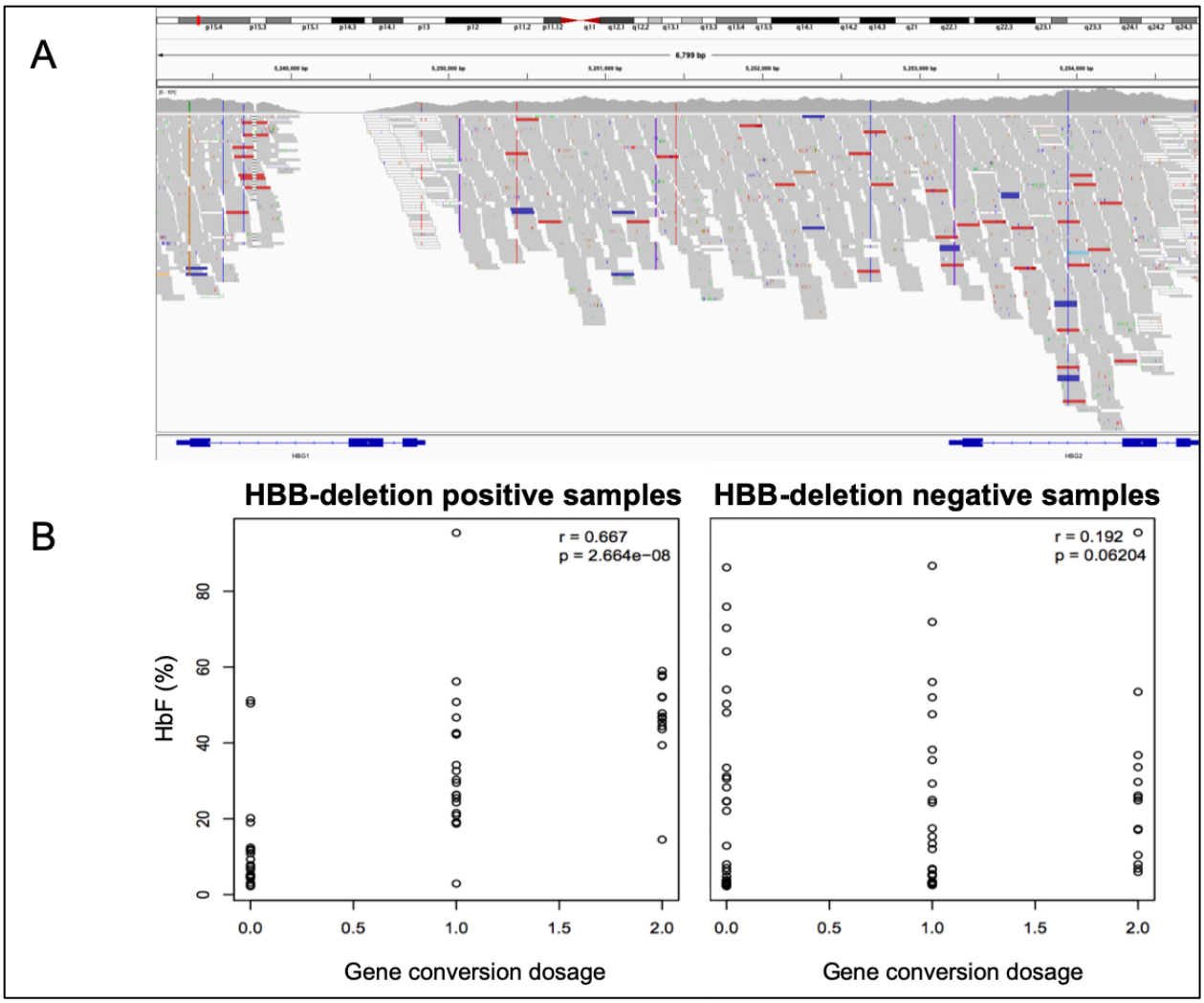

Figure 2. *Novel variant appears to be common within our cohort, and correlates with HbF expression. A) Sequencing reads demonstrate a variant involving HBG1 and HBG2. Image obtained from IGV[70] of a representative sample. B) Side-by-side comparison of gene conversion dosage vs. HbF % in samples with concurrent HBB deletion and samples without.*

Discussion and Future Directions

In this analysis, we performed a rare burden analysis on our whole-genome sequenced Thai individuals, using a standard additive aggregating function to count the number of rare variants present in each gene and subsequently performing a linear regression over all genes. Our results preliminarily demonstrate that HBB is the top candidate for a gene whose burden of rare variants is associated with HbF expression levels.

In carefully investigating the HbF locus, we identified a variant that causes Hemoglobin E disease which has been previously shown in other populations to correlate with HbF levels. We found a similar trend within our own Thai cohort, and found that HbF % increases with the HbE SNP regardless of the presence or absence of a co-occurring deletion within the beta globin gene. These results suggest that our GWAS could be further refined by treating the HbE variant as a covariate.

In collaboration with the McCarroll lab, we have also identified a novel gene conversion event wherein HBG1 is converted to HBG2. This variant is common within our WGS cohort and appears to confer higher HbF levels when a beta globin deletion is concurrent. Efforts are ongoing to further characterize this variant, interrogate other population databases for this conversion event, and ultimately recapitulate the conversion in hematopoietic stem cells to better assess its role in HbF regulation.

Chapter 6: Discussion and Future Directions

In the work contained within this thesis, we have performed common and rare variant analysis on fetal hemoglobin expression. Specifically, we performed genome-wide association studies in a Thai and Swedish cohort; these studies have replicated associations at loci known to modulate HbF levels (at the BCL11A, HBS1L-MYB, and HBB loci), and have identified one more genome-wide significant signal at chromosome 15 in the Swedish population which may reveal a new association. Our group will continue to follow up on these results with meta-analyses of multiple populations to improve statistical power, and perform variant-to-function analyses with statistical tools such as fine-mapping in order to more precisely identify top candidates of our new association.

As our Thai population contained individuals with hereditary persistence of fetal hemoglobin (HPFH), we continued our investigation of this cohort by performing a rare variant burden test, as we suspected there were large-effect rare variants that could be driving this distribution of HbF, and implicated HBB as a gene in which a high burden of rare variants could produce an important correlation with HbF. Further analysis is ongoing and will employ additional covariates and different aggregating functions to our model in order to better assess the contribution of rare variants to HbF expression in this cohort. We carefully investigated the HBB locus and nearby genes, and identified several other interesting findings. First, Hemoglobin E is present in many of our samples, and may be an appropriate covariate for our GWAS and rare burden tests in future analyses, as it appears to correlate with HbF levels and is known to modulate HbF levels in other studies. Second, we identified a conversion event within HBG1 and HBG2 which appears to affect

HbF expression levels. Future studies will aim to recapitulate this variant in hematopoietic stem cells to determine its effect on HbF expression.

In performing these studies, we found that selecting appropriate reference panels was crucial for high-quality association tests and imputation accuracy. Because we studied samples from Thailand, a population generally underrepresented in preexisting genetic databases, we experimented with several imputation tools and developed several metrics to determine quality of output. We found that merging whole genomes from a subset of our samples with the 1000 genomes reference panel appeared to improve the quality and significance of our genome-wide association study. Furthermore, we found that the recently-released GenomeAsia reference panel led to higher imputation accuracy when compared to the 1000 genomes. These results validate our intuition that larger reference panels that contain more representative samples improve GWAS results. Future studies will analyze the effects of merging the 1000 genomes with the GAsP reference panels on imputation accuracy and GWAS.

References

1. Meremikwu MM, Okomo U. Sickle cell disease. *BMJ Clin Evid*. 2011;2011(10091):311-323.
2. Wastnedge E, Waters D, Patel S, et al. The global burden of sickle cell disease in children under five years of age: A systematic review and meta-analysis. *J Glob Health*. 2018;8(2):21103.
3. Rund D, Rachmilewitz E. Medical progress: β-thalassemia. *N Engl J Med*. 2005;353(11):1135-1146.
4. De Silva S, Fisher CA, Premawardhena A, et al. Thalassaemia in Sri Lanka: Implications for the future health burden of Asian populations. *Lancet*. 2000;355(9206):786-791.
5. Weatherall DJ. The inherited diseases of hemoglobin are an emerging global health burden. *Blood*. 2010;115(22):4331-4336.
6. Sankaran VG, Weiss MJ. Anemia: Progress in molecular mechanisms and therapies. *Nat Med*. 2015;21(3):221-230.
7. Garner C, Tatu T, Reittie JE, et al. Genetic influences on F cells and other hematologic variables: A twin heritability study. *Blood*. 2000;95(1):342-346.
8. Menzel S, Garner C, Gut I, et al. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet*. 2007;39(10):1197-1199.
9. Uda M, Galanello R, Sanna S, et al. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of -thalassemia. *Proc Natl Acad Sci*. 2008.

10. Lettre G, Sankaran VG, Bezerra MAC, et al. DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc Natl Acad Sci*. 2008;105(33):11869.
11. Sankaran VG, Menne TF, Xu J, et al. Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. *Science (80-)*. 2008;322(5909):1839-1842.
12. Sankaran VG, Xu J, Ragoczy T, et al. Developmental and species-divergent globin switching are driven by BCL11A. *Nature*. 2009;460(7259):1093-1097.
13. Xu J, Peng C, Sankaran VG, et al. Correction of sickle cell disease in adult mice by interference with fetal hemoglobin silencing. *Science (80-)*. 2011;334(6058):993-996.
14. Dias C, Estruch SB, Graham SA, et al. BCL11A Haploinsufficiency Causes an Intellectual Disability Syndrome and Dysregulates Transcription. *Am J Hum Genet*. 2016;99(2):253-274.
15. Basak A, Hancarova M, Ulirsch JC, et al. BCL11A deletions result in fetal hemoglobin persistence and neurodevelopmental alterations. *J Clin Invest*. 2015;125(6):2363-2368.
16. Xu J, Bauer DE, Kerenyi MA, et al. Corepressor-dependent silencing of fetal hemoglobin expression by BCL11A. *Proc Natl Acad Sci U S A*. 2013;110(16):6518-6523.
17. Xu J, Sankaran VG, Ni M, et al. Transcriptional silencing of {gamma}-globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes Dev*. 2010;24(8):783-798.

18. Basak A, Munschauer M, Lareau CA, et al. Control of human hemoglobin switching by LIN28B-mediated regulation of BCL11A translation. *Nat Genet*. 2020.

19. Esrick EB, Bauer DE. Genetic therapies for sickle cell disease. *Semin Hematol*. 2018;55(2):76-86.

20. Sankaran VG, Orkin SH. The switch from fetal to adult hemoglobin. *Cold Spring Harb Perspect Med*. 2013;3(1):a011643.

21. Sankaran VG, Xu J, Byron R, et al. A functional element necessary for fetal hemoglobin silencing. *N Engl J Med*. 2011;365(9):807-814.

22. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet*. 2019;20(8):467-484.

23. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012;13(4):762-775.

24. Zuk O, Schaffner SF, Samocha K, et al. Searching for missing heritability: Designing rare variant association studies. *Proc Natl Acad Sci*. 2014;111(4):E455 LP-E464.

25. Van Hout C V, Tachmazidou I, Backman JD, et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *bioRxiv*. January 2019:572347.

26. Wainschtein P, Jain DP, Yengo L, et al. Recovery of trait heritability from whole genome sequence data. *bioRxiv*. January 2019:588020.

27. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet*. 2014;95(1):5-23.

28. Wang K, Li M, Hadley D, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*. 2007;17(11):1665-1674.
29. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
30. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinforma*. 2013;43(1):11.10.1-11.10.33.
31. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-498.
32. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*. 2011;27(5):718-719.
33. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987-2993.
34. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.
35. Karczewski KJ, Francioli LC, Tiao G, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. 2019:531210.
36. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets.

Gigascience. 2015;4(1).

37. Consortium T 1000 GP, Auton A, Abecasis GR, et al. A global reference for human genetic variation. *Nature*. 2015;526:68.

38. Wu D, Dou J, Chai X, et al. Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell*. 2019;179(3):736-749.e15.

39. Lu D, Xu S. Principal component analysis reveals the 1000 Genomes Project does not sufficiently cover the human genetic diversity in Asia. *Front Genet*. 2013;4:127.

40. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48:1284.

41. Loh P-R, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48:1443.

42. Loh PR, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet*. 2015;47(3):284-290.

43. D. Turner S. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J Open Source Softw*. 2018;3(25):731.

44. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*. 2010;26(18):2336-2337.

45. Jonsson S, Sveinbjornsson G, de Lapuente Portilla AL, et al. Identification of sequence variants influencing immunoglobulin levels. *Nat Genet*. 2017;49(8):1182-1191.

46. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of

genomewide association scans. *Bioinformatics*. 2010;26(17):2190-2191.

47. Box GEP, Cox DR. An Analysis of Transformations. *J R Stat Soc Ser B*. 1964;26(2):211-252.

48. Thein SL, Menzel S, Lathrop M, Garner C. Control of fetal hemoglobin: New insights emerging from genomics and clinical implications. *Hum Mol Genet*. 2009;18(R2):R216-23.

49. Creary LE, Ulug P, Menzel S, et al. Genetic variation on chromosome 6 influences F cell levels in healthy individuals of African descent and HbF levels in sickle cell patients. *PLoS One*. 2009;4(1):e4218.

50. Solovieff N, Milton JN, Hartley SW, et al. Fetal hemoglobin in sickle cell anemia: Genome-wide association studies suggest a regulatory region in the 5′ olfactory receptor gene cluster. *Blood*. 2010;115(9):1815-1822.

51. Nuinoon M, Makarasara W, Mushiroda T, et al. A genome-wide association identified the common genetic variants influence disease severity in β 0 - thalassemia/hemoglobin e. *Hum Genet*. 2010;127(3):303-314.

52. Bhatnagar P, Purvis S, Barron-Casella E, et al. Genome-wide association study identifies genetic variants influencing F-cell levels in sickle-cell patients. *J Hum Genet*. 2011;56(4):316-323.

53. Wall JD, Stawiski EW, Ratan A, et al. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*. 2019;576(7785):106-111.

54. Gögele M, Minelli C, Thakkinstian A, et al. Methods for Meta-Analyses of Genome-wide Association Studies: Critical Assessment of Empirical Evidence. *Am J Epidemiol*. 2012;175(8):739-749.

55. Zeggini E, Ioannidis JPA. Meta-analysis in genome-wide association studies. *Pharmacogenomics*. 2009;10(2):191-201.

56. Ulirsch JC, Lareau CA, Bao EL, et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat Genet*. 2019:https://doi.org/10.1038/s41588-019-0362-6.

57. Consortium TU, Walter K, Min JL, et al. The UK10K project identifies rare variants in health and disease. *Nature*. 2015;526:82.

58. the Haplotype Reference C, McCarthy S, Das S, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48:1279.

59. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv*. January 2019:563866.

60. Huang L, Li Y, Singleton AB, et al. Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet*. 2009;84(2):235-250.

61. Song Q, Xu W, Li W, et al. Accurate haplotype imputation with individualized ancestry-adjusted reference panels. *Genomics*. 2018;110(5):329-335.

62. Ahmad M, Sinha A, Ghosh S, et al. Inclusion of Population-specific Reference Panel from India to the 1000 Genomes Phase 3 Panel Improves Imputation Accuracy. *Sci Rep*. 2017;7(1):6733.

63. Wyss AB, Sofer T, Lee MK, et al. Multiethnic meta-analysis identifies ancestry-specific and cross-ancestry loci for pulmonary function. *Nat Commun*. 2018;9(1):2976.

64. Rappoport N, Toung J, Hadley D, et al. A genome-wide association study identifies only two ancestry specific variants associated with spontaneous preterm

birth. *Sci Rep*. 2018;8(1):226.

65. Ros-Freixedes R, Whalen A, Chen C-Y, et al. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. *bioRxiv*. January 2019:771576.

66. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res*. 2011;21(6):940-951.

67. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3 (Bethesda)*. 2011;1(6):457-470.

68. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*. 2012.

69. O'Connell J, Gurdasani D, Delaneau O, et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet*. 2014;10(4):e1004234.

70. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26.

Glossary: Abbreviations

CVAS: common variant association study

DNA: deoxyribonucleic acid

DP: total read depth

GAsP: GenomeAsia Pilot

GWAS: genome-wide association study

HbE: hemoglobin E

HbF: fetal hemoglobin

HPFH: hereditary persistence of fetal hemoglobin

HPLC: high performance liquid chromatography

HRC: Haplotype Reference Consortium

IBD: identity by descent

LD: linkage disequilibrium

lncRNA: long non-coding RNA

MAF: minor allele frequency

MLPA: multiplex ligation-dependent probe amplification

RNA: ribonucleic acid

RVAS: rare variant association study

SCD: sickle cell disease

shRNA: short-hairpin RNA

SNP: single nucleotide polymorphism

SNV: single nucleotide variant

VCF: variant call file

VEP: variant effect predictor

WES: whole-exome sequencing

WGS: whole-genome sequencing