Cold Spring Harbor Perspectives in Medicine
www.perspectivesinmedicine.org

The Search for Genetic Modifiers of Disease Severity in the β-Hemoglobinopathies

Guillaume Lettre

Montreal Heart Institute and Université Montréal, Montréal, Québec H1T 1C8, Canada

Correspondence: guillaume.lettre@mhi-humangenetics.org

Sickle cell disease (SCD) and β-thalassemia, two monogenic diseases caused by mutations in the *β-globin* gene, affect millions of individuals worldwide. These hemoglobin disorders are characterized by extreme clinical heterogeneity, complicating patient management and treatment. A better understanding of this patient-to-patient clinical variability would dramatically improve care and might also guide the development of novel therapies. Studies of the natural history of these β-hemoglobinopathies have identified fetal hemoglobin levels and concomitant α-thalassemia as important modifiers of disease severity. Several small-scale studies have attempted to identify additional genetic modifiers of SCD and β-thalassemia, without much success. Fortunately, improved knowledge of the human genome and the development of new genomic tools, such as genome-wide genotyping arrays and next-generation DNA sequencers, offer new opportunities to use genetics to better understand the causes of the many complications observed in β-hemoglobinopathy patients. Here I discuss the most important factors to consider when planning an experiment to find associations between β-hemoglobinopathy-related complications and DNA sequence variants, with a focus on how to successfully perform a genome-wide association study. I also review the literature and explain why most published findings in the field of SCD modifier genetics are likely to be false-positive reports, with the goal to draw lessons allowing investigators to design better genetic experiments.

Two of the promises of the Human Genome Project were to give the medical community the resources to better grasp interindividual variation in disease and treatment and to uncover previously unknown biology (Lander et al. 2001; Venter et al. 2001). In a recent review article to celebrate the 10-year anniversary of our genome's sequence, the genomicist Eric Lander highlighted many of the outstanding accomplishments made toward achieving these promises: from evolutionary conservation and non-coding RNAs to landscape of epigenetic marks and maps of genetic variation (Lander 2011). As anticipated, the field of human genetics has prospered on the foundation set by the Human Genome Project, in particular through the discovery and characterization of genetic variation by flagship projects such as the HapMap Project (Altshuler et al. 2005, 2010; Frazer et al. 2007) and the 1000 Genomes Project (1000 Genomes Project Consortium 2010). This has led to novel insights into the selective and recombination forces that have shaped our genome, but most and foremost this has had a

Editors: David Weatherall, Alan N. Schechter, and David G. Nathan
Additional Perspectives on Hemoglobin and Its Diseases available at www.perspectivesinmedicine.org

 doi: 10.1101/cshperspect.a015032
Cite this article as *Cold Spring Harb Perspect Med* 2012;2:a015032

dramatic impact on our ability to identify genetic risk factors for complex human diseases. Progress in our theoretical understanding of the genetic variation present in our genome—both in terms of its frequency spectrum and correlated nature—was paved by advances in the development of high-throughput DNA genotyping and sequencing technologies. Together, these new genetic knowledge and technologic capabilities led to the discovery of more than 1000 common DNA sequence variants (so-called single-nucleotide polymorphisms, SNPs) associated with common human diseases and other complex traits by genome-wide association studies (GWAS) (Manolio et al. 2009).

Although most of the attention following the release of the human genome sequence has focused on the genetics of common diseases—which is understandable after decades of unfruitful research—our DNA sequence also helped us make significant progress in identifying genes responsible for rare genetic diseases. More recently, we have seen many studies on various human syndromes that used whole-exome DNA resequencing to find causal genes, often after many years of unsuccessful positional cloning and linkage study efforts (Teer and Mullikin 2010). These successes, together with the convergence of tools initially developed to study complex human diseases, have made us better appreciate that many *simple* Mendelian genetic diseases are, in fact, all but simple. This is particularly true for the β-hemoglobinopathies sickle cell disease (SCD) and β-thalassemia, which are the most common monogenic diseases in the world (Weatherall and Clegg 2001; Weatherall 2010) and are characterized by a very heterogeneous clinical course between patients with the same mutations, ranging from nearly asymptomatic to life-threatening conditions (Steinberg et al. 2009). The interpatient clinical variability in the β-hemoglobinopathies has sparked a vivid interest in identifying genetic modifiers of severity for these diseases. The rationale for this quest is that such genetic modifiers, if they exist, might lead to more precise (or personalized) prognostic tests, but also guide the development of more specific and effective therapies.

Fetal hemoglobin (HbF) levels and concomitant α-thalassemia are the two best-characterized modifiers of severity in SCD and β-thalassemia. Because these two topics are reviewed in other work (Nienhuis and Nathan 2012; Schechter and Elion 2012), they are not extensively discussed here. Rather, I focus on the main practical questions that one needs to address before undertaking an experiment to find new genetic modifiers for SCD or β-thalassemia. What are the main considerations in terms of phenotype definition, genotype acquisition, data analysis, and result validation for such genetic experiments? Finally, with a focus on SCD, I review the existing literature on genetic modifiers and assess the robustness of previously identified genetic associations, highlight the lessons learned from these earlier studies, and discuss how we might build on these results to design future experiments.

PHENOTYPE CONSIDERATIONS

Phenotype Definitions

Until a few years ago, the main preoccupation of human geneticists was to collect sufficient genetic information in their patient samples to begin the search for genetic risk factors. Nowadays, the emergence of genome-wide genotyping platforms and next-generation DNA sequencers has streamlined what used to be an incredible task: One can now sequence a complete human genome at high coverage in 2 wk for less than $5000. This problem solved, scientists are now redirecting their attention to the importance of the quality and harmonization of the phenotypes that are being studied. Although many phenotypes are straightforward to measure (e.g., HbF levels, white blood cell counts), others are more difficult to ascertain (e.g., pulmonary hypertension), and phenotype definitions are not always consistent, even between clinicians at the same clinic. As one example in the field of complex trait genetics, phenotype heterogeneity has been proposed as one possible reason why it has proven so difficult to understand the genetics of psychiatric diseases, despite a very large genetic contribution (Rasetti and

Weinberger 2011). Generally, simpler phenotypes that are reproducible and easy to collect (e.g., height) or endo-phenotypes, which are often quantitative traits related to disease etiology (e.g., cholesterol levels for atherosclerosis), have been studied more successfully than complex clinical end points using genetic strategies. There are also important practical considerations, especially in the context of the β-hemoglobinopathies, that influence the large-scale clinical availability of certain phenotypes when attempting to find genetic associations. Transcranial Doppler (TCD) ultrasonography and magnetic resonance imaging (MRI) results provide robust prognostic tools for cerebral vascular accidents (strokes) in children with SCD (Adams et al. 1992; DeBaun et al. 1995). They would thus appear as ideal phenotypes to identify modifiers of stroke risk in SCD. But these exquisite imaging measurements are often impossible to collect in populations in which the prevalence of SCD is high (e.g., Africa), limiting these studies to a small number of patients, which often leads to false-positive reports (Lohmueller et al. 2003) (see below).

The choice of phenotypes and ascertainment scheme, for instance, selection of a case-control sample as defined by the presence or absence of a particular clinical SCD complication, will also determine how analyses of associations of phenotypes with genotype information must be performed. Both dichotomous and continuous traits can be analyzed using simple statistics (e.g., χ^2 or Fisher's exact test, linear or logistic regression), but these analyses need to control for relevant confounding factors or matching variables such as age, sex, or medications. They also need to take into account family structure if related individuals are recruited. For continuous traits, the investigator could focus on patients with the most extreme phenotypes. For example, analyzing the bottom and top fifth percentile of the phenotypic distribution would reduce genotyping or sequencing costs by 90%; this strategy was successfully used by one of the two groups that found the association between the *BCL11A* locus and HbF levels (Menzel et al. 2007). This is a promising strategy as long as there is evidence that the genetic architecture of the trait is the same in the extremes as in the rest of the distribution. In other words, this approach works if the same genes/alleles control phenotypic variation across the phenotypic distribution, which is usually unknown, although recent work on adult height suggests that it is not always the case (Chan et al. 2011). Finally, although retrospective case-control samples are often easier to assemble, one cannot sufficiently emphasize the virtues of prospective studies. They are more expensive and time-consuming but eliminate many caveats and offer unique advantages to identify and test predictive risk factors. Longitudinal data are also amenable to powerful statistical methodologies, which can become important when the phenotypes are rare. This is certainly one of the main reasons why the large prospective Cooperative Study of Sickle Cell Disease (CSSCD), which was initiated in 1977 in the United States to determine the natural history of SCD from birth to death, has had such a tremendous impact on how we understand and treat SCD (Farber et al. 1985; Gaston et al. 1987).

Heritability

Most importantly, for genetic association studies to be successful, the phenotypes need to have a genetic component. Heritability is the key concept here, but it is often misunderstood (Visscher et al. 2008). It is defined, in the case of narrow-sense heritability (h^2), as the ratio of the phenotypic variation explained by additive genetic factors over the total phenotypic variation observed in the population. A low heritability estimate indicates that a large fraction of the phenotypic variation (or disease risk on a liability scale) is due to environmental factors (plus other non-additive genetic effects) and that, consequently, the phenotype is harder to track genetically. Heritability is usually estimated by comparing phenotype concordance among relatives, although new genomic-based methods have been developed that use only distantly related individuals (Visscher et al. 2006; Yang et al. 2011).

There is surprisingly little information available on the heritability of the phenotypes and complications relevant to the severity of the

β-hemoglobinopathies. The estimate for the heritability of HbF levels is high ($h^2 \sim 0.6$–0.9), but this was obtained in healthy individuals of European descent (Garner et al. 2000; Pilia et al. 2006). In Jamaica, a study with nine identical pairs of twins with sickle cell anemia showed good concordance for many hematological traits, whereas the clinical complications were often discordant (Weatherall et al. 2005). With a focus on stroke, two groups also found evidence of familial aggregation in SCD patients (Driscoll et al. 2003; Kwiatkowski et al. 2003). Finally, in our analysis of genome-wide genotypes in the CSSCD, we identified 104 pairs of full sibs in which we analyzed phenotypic correlation (Table 1). Correlation coefficients were high for HbF levels, pain crisis rate, and stroke; moderate for acute chest syndrome, leg ulcer, osteonecrosis, and priapism; and low for survival. As the size and clinical quality of the β-hemoglobinopathy cohorts increase, estimating heritability of the complications before trying to find their genetic modifiers is certainly an avenue of research that should be pursued.

Table 1. Correlation of eight phenotypes in 104 pairs of full sibs from the Cooperative Study of Sickle Cell Disease (CSSCD)

Phenotype	Number of pairs with phenotype available (affected siblings/unaffected siblings)	Correlation coefficient
Acute chest syndrome rate	104	0.166
Fetal hemoglobin levels	84	0.579
Leg ulcer	84 (26/142)	0.183
Osteonecrosis	104 (37/171)	0.265
Pain crisis rate	104	0.683
Priapism (male only)	25 (9/41)	0.218
Stroke	104 (16/192)	0.361
Survival	104 (7/201)	0.024

These estimates must be considered carefully because the sample size is small and they do not take into account important confounding factors, such as socioeconomic status.

GENOTYPE CONSIDERATIONS

GWAS

Before the development of genome-wide genotyping arrays, genetic searches for complex diseases and traits used a combination of linkage- and candidate-gene-based approaches. For reasons that fall outside the scope of this article, linkage is inappropriate to find common genetic variants of small effect on phenotypic variation (Risch and Merikangas 1996), and candidate-gene studies are limited to the known biology. Except for the identification of the association between HbF levels and genetic variation at the *HBS1L–MYB* intergenic region, first by linkage (Craig et al. 1996; Garner et al. 1998) and subsequently by association study (Thein et al. 2007), these strategies have not proven very useful to find genetic modifiers of severity in the β-hemoglobinopathies.

Genetic association analysis consists of testing whether a specific allele, for example, the A allele at an A/G biallelic SNP, is found more often in cases than controls or whether it is correlated with a continuous trait (Hirschhorn and Daly 2005). Because alleles at different markers may be correlated in the genome, a phenomenon called linkage disequilibrium (LD) (Slatkin 2008), it is possible to test association with a large number of variants by genotyping only a subset of them. Because of LD, GWAS test associations systematically across the whole genome, but for markers that are common in the population of interest (usually SNPs with a minor allele frequency [MAF] >5%; for rarer markers, DNA sequencing is required because of weak LD) (see below).

The first successful GWAS was performed in 2005 and identified the complement factor H (*CFH*) gene as a risk factor for age-related macular degeneration among a cohort of 96 cases and 50 controls, and genotypes at ~100,000 SNPs (Klein et al. 2005). Since then, GWAS have increased in size both in terms of the number of participants genotyped and markers tested. For example, in a recent study on body mass index, about 2.5 million SNPs were analyzed in about 250,000 individuals (Speliotes et al.

2010). For the β-hemoglobinopathies, GWAS have so far been limited to the discovery of *BCL11A* and its role on HbF levels (Menzel et al. 2007; Uda et al. 2008), SCD pain crises (Lettre et al. 2008), and β-thalassemia transfusion dependency (Uda et al. 2008; Nuinoon et al. 2009). In the next sections, the key concepts required for a successful GWAS are reviewed.

Statistical Power

For genetic association analysis, the statistical power is defined as the probability of rejecting the null hypothesis of no association when the null hypothesis is indeed false. In other words, if there is a true association between genotype and phenotype, what is the probability of finding it given your study design? Power depends on sample size, marker allele frequency, linkage disequilibrium (LD) strength between the genotyped markers and the causal variants, effect size of the genetic variants on the phenotype, and the number of independent markers tested. This latter factor, also referred to as the "multiple hypothesis burden," is particularly important in GWAS because hundreds of thousands or even millions of SNPs are usually tested. To limit the number of false-positive associations (type I errors), we use very stringent significance criteria to declare significance, but this has the disadvantage of reducing power. For GWAS, a generally accepted criterion to claim genome-wide significance is a p-value $< 5 \times 10^{-8}$ (Altshuler et al. 2005). Power is a critical concept to estimate before embarking on any GWAS project but also essential to interpret published GWAS results correctly. Fortunately, user-friendly tools exist to compute the power of association studies (Purcell et al. 2003; Gauderman and Morrison 2006).

Population: Coverage, Stratification, and Admixture

Genetic variation differs between populations because of mutations, genetic drift, migration, and natural selection. As a consequence, not all genetic variants are present in all populations, or the allele frequency of a given variant may differ across populations. For example, populations of African ancestry are genetically more ancient and carry more genetic variation (both rare and common) than populations of European or Asian descent (1000 Genomes Project Consortium 2010). Because the content of commercial genome-wide genotyping arrays is fixed, they cover more exhaustively the common genetic variations in non-African populations (Bhangale et al. 2005). This can have a major impact on genetic discovery efforts, especially for diseases that are more prevalent in Africans or individuals of African ancestry such as SCD. Novel genotyping arrays that were designed using empirical data from the 1000 Genomes Project, as well as imputation methods (Li et al. 2009) that can probabilistically infer genotypes at ungenotyped markers using genotypes from reference sets such as the HapMap or 1000 Genomes Projects, can, however, mitigate the bias introduced by genotyping the more limited markers present in commercial genotyping arrays.

Allele frequency differences between populations can also result in false-positive genetic association results when the prevalence of a disease (or the distribution of a quantitative trait) differs across these populations. The classic example of this important confounder, called population stratification (Price et al. 2010), is the spurious association between height and the DNA polymorphisms responsible for lactase persistence in populations of European descent because both height and lactose tolerance follow a North–West/South–East cline within Europe (Campbell et al. 2005). Using genome-wide genotypes or a limited set of ancestry informative markers, several methods now exist to detect and correct for the effect of population stratification on association results (Price et al. 2006; Kang et al. 2010). It is also worth mentioning that family-based association tests can be protected from the effect of population stratification, providing additional value to this study design.

Whereas population stratification applies to subpopulations from the same continent (e.g., European Americans), admixture describes individuals whose ancestors are from different continents, such as African Americans or Hispanic Americans (Seldin et al. 2011). The

chromosomes in admixed populations are mosaics with chromosomal segments from each ancestral populations: This can be useful for gene mapping but also presents several analytical challenges that must be appropriately accounted for when performing GWAS (Price et al. 2009). Correction for admixture is particularly important in the search for genetic modifiers of the β-hemoglobinopathies because many studied populations are, in fact, admixed.

Replication

As mentioned above, the challenge in GWAS is to identify the few authentic genetic associations among millions of markers tested. We can increase the stringency of the significance threshold to account for the number of tests performed, but this has the disadvantage of decreasing power. To compensate for the power loss, we can analyze more clinical or population samples by GWAS, but this increases costs. Most successful GWAS have used staged designs: All markers are first tested in an initial "discovery" population, and then the most promising markers (based on statistical evidence or biological candidacy) are analyzed in replication cohorts. In fact, the replication of genetic associations in independent samples—ideally, a replication study would test the same phenotype, allele, direction of effect, and genetic model with a different genotyping platform—is now the gold standard and is essentially mandatory before publication because it has been shown to be key to avoid false-positive reports (Lohmueller et al. 2003; Chanock et al. 2007). Given the importance of replicating GWAS findings, any proposal to identify β-hemoglobinopathy modifiers should include a plan to replicate association results in independent samples. It is critical to consider accessibility to replication cohorts (independent samples, same phenotypes) early in the design of any GWAS experiments.

Next-Generation DNA Sequencing

The GWAS framework is extremely efficient to capture the role of common genetic variation (MAF $>$ 5%) in phenotypic variation. Rare (MAF $<$ 0.1%) and low-frequency (MAF 0.1% − 5%) variants are usually poorly correlated with common variants surveyed by GWAS and have therefore not been tested comprehensively for their role in disease risk. With the development of affordable next-generation DNA sequencing technologies and data release from the 1000 Genomes Project (1000 Genomes Project Consortium 2010), we now have the tools to test the role of rare and low-frequency variants in complex human phenotypes. Genetic results for HbF, implicating rare familial mutations in the *β-globin* locus and *KLF1* (Borg et al. 2010) and low-frequency variants in *MYB* (Galarneau et al. 2010), strongly hint that rare genetic variation might also modify disease severity in the β-hemoglobinopathies. With many ongoing whole-exome sequencing projects, we should have a better idea of the role of rare and low-frequency genetic variants in SCD and β-thalassemia complications in the coming years.

REVIEW OF PUBLISHED GENETIC MODIFIERS IN SCD

Several studies have been published describing genetic associations between DNA polymorphisms in strong candidate genes and many complications specific to SCD (for review, see Fertrin and Costa 2010). Table 2 provides the details of most of the published genetic associations with acute chest syndrome, leg ulcer, osteonecrosis, pain crisis, and priapism. Based on these data, we can make the following observations: (1) Most sample sizes are small. (2) Reported odds ratios (OR) are high in comparison to what we now regard as realistic effect sizes in genetic association studies (usually OR $<$ 1.2). (3) For many studies, the significance threshold was not adjusted for the number of tests performed. (4) Few results have been replicated in independent cohorts (Table 2). These are all characteristics usually linked to false-positive reports (Lohmueller et al. 2003), which is reflected by the low discovery power that any of these studies had to find genetic associations even under an ideal scenario (OR = 1.5; last column in Table 2).

Table 2. Published genetic associations with sickle cell disease–related complications

Complication	References	Sample size (cases/controls)	Gene (variant)	Reported odds ratio	Reported *p*-value	Appropriate significance threshold[a]	Replication study (reference)	Power (%)[b]
Acute chest syndrome	Sharan et al. 2004	18/27	*NOS3* (rs2070744)	6.5	0.005	0.05	No	10
Leg ulcer	Nolan et al. 2006	243/516	*KL* (rs516306)	1.8	0.0076	0.0002 (0.05/215 tests)	No	26
			MAP3K7 (rs157702)	1.6	0.008		No	
			SMAD7 (rs736839)	2.0	0.0004		No	
	Ofosu et al. 1987	9/29	*HLA* (B35/Cw4)	17.0	<0.005	0.05	No	10
Osteonecrosis	Baldwin et al. 2005	442/455	*KL* (rs211239)	2.6	0.001	0.0008 (0.05/66 tests)	Yes, but association did not replicate (Ulug et al. 2009).	47
			BMP6 (rs267196)	1.9	0.001		Yes, but association did not replicate; see text (Ulug et al. 2009).	
			ANXA2 (rs7170421)	3.4	<0.001		Yes, but association did not replicate (Ulug et al. 2009).	
Pain crisis	Mendonca et al. 2010	48/39	MBL2 (exon 1 0/A and −221 X/Y)	3.2	0.02	0.05	No	29
	Al-Subaie et al. 2009	127/130	*GP1BA* (rs6065)	1.8	0.004	0.02 (0.05/3 tests)	No	45
			ITGA2B (rs5911)	2.0	0.0008		No	
			ITGA2 (rs1801106)	1.9	0.002		No	
Priapism	Nolan et al. 2005	148/529	*KL* (rs2249358)	2.6	Not reported	0.001 (0.05/44 tests)	Yes, but association did not replicate (Elliott et al. 2007).	40
	Elliott et al. 2007	83/116	*TGFBR3* (rs7526590)	2.7	0.00058	0.001 (0.05/49 tests)	No	5
			AQP1 (rs10244884)	2.1	0.00068		No	
			ITGAV (rs3768780)	2.0	0.0009		No	

This is not an exhaustive list. Meeting abstracts were not considered because of the difficulty to assess the quality of the results.

[a]The appropriate significance threshold corresponds to a Bonferroni correction for the number of genes/tests performed in each study.

[b]Discovery power for each study was calculated under the following assumptions: odds ratio = 1.5, effect allele frequency = 20%, sample size = number of cases and controls available in the study, α level = appropriate significance threshold, prevalence as determined from the Cooperative Study of Sickle Cell Disease (CSSCD) phase I clinical data set.

The reported association between SNPs at the *BMP6* locus and osteonecrosis risk is a good example of the inconsistency that exists in the literature. The association was first identified in 442 cases and 455 controls from the CSSCD (Baldwin et al. 2005) with a replication attempt in another SCD cohort that included 39 osteonecrosis cases and 205 controls (Ulug et al. 2009). Although the investigators of the replication study claimed that they could replicate the *BMP6*–osteonecrosis association, careful analysis of the articles shows that the two studies have marginal associations with different *BMP6* SNPs (rs267196 and rs3812163) that are 124 kb away from each other and not in LD ($r^2 \sim 0$). By definition, because a different SNP was tested in the replication study, the original association between *BMP6* and osteonecrosis has not been replicated yet. Thus, none of the associations presented in Table 2 appear robust.

Stroke

Because of its devastating consequences and apparent heritability (Table 1) (Driscoll et al. 2003; Kwiatkowski et al. 2003), many investigators have attempted to identify genetic risk factors for stroke in SCD (Hoppe et al. 2004, 2007; Sebastiani et al. 2005; Voetsch et al. 2008). In a recent replication study, Flanagan and colleagues attempted to confirm these previous stroke association results in an independent SCD cohort of 130 stroke cases and 103 controls (Flanagan et al. 2011). This study had 44% power to replicate an association with stroke (using the same assumptions reported in Table 2). The researchers could replicate the known protective effect of α-thalassemia on stroke risk, and also reported nominally significant association results for SNPs in four genes: *ADYC9*, *ANXA2*, *TEK*, and *TGFBR3* (Flanagan et al. 2011). These results may seem encouraging, but once again they must be interpreted with caution. SNPs in *ADYC9*, *ANXA2*, *TEK*, and *TGFBR3* were initially related to stroke risk in SCD through a Bayesian network (Sebastiani et al. 2005). Bayesian networks integrate many variables, in this case, laboratory measures such as HbF levels but also genotypes, to predict disease outcome. Although they are theoretically interesting as methods to analyze large data sets with many variable dependencies, the biological and/or clinical interpretation of the different nodes in the network is difficult. In this specific example, although the Bayesian network was validated in an independent cohort of SCD patients with or without stroke, the marginal effect of the SNPs in the network on stroke risk remains unclear. In fact, we tested by logistic regression in the CSSCD the association between stroke and genotypes at *ADCY9*-rs2238432, *ANXA2*-rs11853426, *TEK*-rs489347, and *TGFBR3*-rs28 4875, and observed no significant results ($p > 0.15$) (KS Lo and G Lettre, unpubl.).

Two Convincing Associations

Besides HbF levels and α-thalassemia, there are two other convincing genetic modifiers of complications in SCD. Importantly, these associations are more compelling because they were first very robustly confirmed with phenotypes in non-SCD populations. Polymorphisms in the promoter of the *UGT1A1* gene are associated with serum bilirubin levels in many studies and different populations, including patients with SCD; the same polymorphisms are also associated with the risk of formation of gallstones in SCD patients (Passon et al. 2001; Fertrin et al. 2003; Chaar et al. 2005, 2006; Haverfield et al. 2005; Carpenter et al. 2008; Milton et al. 2012). Renal failure is one of the major complications of SCD and is a strong predictor of mortality in SCD (Platt et al. 1994). Admixture mapping and association studies have identified the *MYH9*–*APOL1* locus as an important genetic risk factor for kidney disease in populations of African ancestry (Kao et al. 2008; Kopp et al. 2008; Genovese et al. 2010). More recently, risk alleles at the *MYH9*–*APOL1* locus were also associated in SCD patients with proteinuria and decreased glomerular filtration rate, both hallmarks of SCD nephropathy (Ashley-Koch et al. 2011). The *UGT1A1* and *MYH9*–*APOL1* results are reminders of the importance of large sample size, access to replication cohorts, and clearly defined or well-harmonized phenotypes to find true-positive genetic associations.

CONCLUSIONS

The identification of genetic modifiers of disease severity for the β-hemoglobinopathies promises to improve patient management and treatment. As highlighted in this article, many of the previous associations between SCD complications and genetic variation are false-positive reports. This result is not specific to SCD because most of the association studies performed before the GWAS era were underpowered. Fortunately, robust associations for many complex human phenotypes, including some relevant to hemoglobin disorders (e.g., HbF levels, kidney disease), now exist and allow us to identify the key factors that lead to successful genetic searches: clear phenotype definition and heritability estimate, large sample size and knowledge of statistical power, control for the multiple hypothesis burden and other confounders such as population stratification and admixture, and access to independent samples for replication. Because the importance of these factors are well appreciated and with large collaborations of geneticists and clinicians working on the β-hemoglobinopathies already in place, we are finally in a position to ask convincingly important questions, test clinically meaningful hypotheses, and, hopefully, gain insight into the pathophysiology of SCD and β-thalassemia. This is clearly an exciting time for β-hemoglobinopathy patients, their families, and physicians.

ACKNOWLEDGMENTS

I thank Alexander P. Reiner, Aikaterini Kritikou, and Geneviève Galarneau for comments on earlier versions of this article, and Samuel Lessard for identifying related individuals in the CSSCD. Dr. Martin H. Steinberg kindly provided access to the CSSCD GWAS data set (funded by the NHLBI STAMPEED). Work on SCD in my laboratory is supported by an Innovations in Clinical Research Award grant from the Doris Duke Charitable Foundation, as well as by the Canada Research Chair program, the Canadian Institute of Health Research, and the Fonds de Recherche Santé Québec.

REFERENCES

* *Reference is also in this collection.*

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.

Adams R, McKie V, Nichols F, Carl E, Zhang DL, McKie K, Figueroa R, Litaker M, Thompson W, Hess D. 1992. The use of transcranial ultrasonography to predict stroke in sickle cell disease. *N Engl J Med* **326:** 605–610.

Al-Subaie AM, Fawaz NA, Mahdi N, Al-Absi IK, Al-Ola K, Ameen G, Almawi WY. 2009. Human platelet alloantigens (HPA) 1, HPA2, HPA3, HPA4, and HPA5 polymorphisms in sickle cell anemia patients with vaso-occlusive crisis. *Eur J Haematol* **83:** 579–585.

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467:** 52–58.

Ashley-Koch AE, Okocha EC, Garrett ME, Soldano K, De Castro LM, Jonassaint JC, Orringer EP, Eckman JR, Telen MJ. 2011. MYH9 and APOL1 are both associated with sickle cell disease nephropathy. *Br J Haematol* **155:** 386–394.

Baldwin C, Nolan VG, Wyszynski DF, Ma QL, Sebastiani P, Embury SH, Bisbee A, Farrell J, Farrer L, Steinberg MH. 2005. Association of klotho, bone morphogenic protein 6, and annexin A2 polymorphisms with sickle cell osteonecrosis. *Blood* **106:** 372–375.

Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. 2005. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum Mol Genet* **14:** 59–69.

Borg J, Papadopoulos P, Georgitsi M, Gutierrez L, Grech G, Fanis P, Phylactides M, Verkerk AJ, van der Spek PJ, Scerri CA, et al. 2010. Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. *Nat Genet* **42:** 801–805.

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN. 2005. Demonstrating stratification in a European American population. *Nat Genet* **37:** 868–872.

Carpenter SL, Lieff S, Howard TA, Eggleston B, Ware RE. 2008. UGT1A1 promoter polymorphisms and the development of hyperbilirubinemia and gallbladder disease in children with sickle cell anemia. *Am J Hematol* **83:** 800–803.

Chaar V, Keclard L, Diara JP, Leturdu C, Elion J, Krishnamoorthy R, Clayton J, Romana M. 2005. Association of UGT1A1 polymorphism with prevalence and age at onset of cholelithiasis in sickle cell anemia. *Haematologica* **90:** 188–199.

Chaar V, Keclard L, Etienne-Julan M, Diara JP, Elion J, Krishnamoorthy R, Romana M. 2006. UGT1A1 polymorphism outweighs the modest effect of deletional (−3.7 kb) α-thalassemia on cholelithogenesis in sickle cell anemia. *Am J Hematol* **81:** 377–379.

Chan Y, Holmen OL, Dauber A, Vatten L, Havulinna AS, Skorpen F, Kvaloy K, Silander K, Nguyen TT, Willer C, et al. 2011. Common variants show predicted polygenic effects on height in the tails of the distribution, except in extremely short individuals. *PLoS Genet* **7:** e1002439.

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, et al. 2007. Replicating genotype–phenotype associations. *Nature* **447:** 655–660.

Craig JE, Rochette J, Fisher CA, Weatherall DJ, Marc S, Lathrop GM, Demenais F, Thein S. 1996. Dissecting the loci controlling fetal haemoglobin production on chromosomes 11p and 6q by the regressive approach. *Nat Genet* **12:** 58–64.

DeBaun MR, Glauser TA, Siegel M, Borders J, Lee B. 1995. Noninvasive central nervous system imaging in sickle cell anemia. A preliminary study comparing transcranial Doppler with magnetic resonance angiography. *J Pediatr Hematol Oncol* **17:** 29–33.

Driscoll MC, Hurlet A, Styles L, McKie V, Files B, Olivieri N, Pegelow C, Berman B, Drachtman R, Patel K, et al. 2003. Stroke risk in siblings with sickle cell anemia. *Blood* **101:** 2401–2404.

Elliott L, Ashley-Koch AE, De Castro L, Jonassaint J, Price J, Ataga KI, Levesque MC, Brice Weinberg J, Eckman JR, Orringer EP, et al. 2007. Genetic polymorphisms associated with priapism in sickle cell disease. *Br J Haematol* **137:** 262–267.

Farber MD, Koshy M, Kinney TR. 1985. Cooperative Study of Sickle Cell Disease: Demographic and socioeconomic characteristics of patients and families with sickle cell disease. *J Chronic Dis* **38:** 495–505.

Fertrin KY, Costa FF. 2010. Genomic polymorphisms in sickle cell disease: Implications for clinical diversity and treatment. *Exp Rev Hematol* **3:** 443–458.

Fertrin KY, Melo MB, Assis AM, Saad ST, Costa FF. 2003. UDP-glucuronosyltransferase 1 gene promoter polymorphism is associated with increased serum bilirubin levels and cholecystectomy in patients with sickle cell anemia. *Clin Genet* **64:** 160–162.

Flanagan JM, Frohlich DM, Howard TA, Schultz WH, Driscoll C, Nagasubramanian R, Mortier NA, Kimble AC, Aygun B, Adams RJ, et al. 2011. Genetic predictors for stroke in children with sickle cell anemia. *Blood* **117:** 6681–6684.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851–861.

Galarneau G, Palmer CD, Sankaran VG, Orkin SH, Hirschhorn JN, Lettre G. 2010. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat Genet* **42:** 1049–1051.

Garner C, Mitchell J, Hatzis T, Reittie J, Farrall M, Thein SL. 1998. Haplotype mapping of a major quantitative-trait locus for fetal hemoglobin production, on chromosome 6q23. *Am J Hum Genet* **62:** 1468–1474.

Garner C, Tatu T, Reittie JE, Littlewood T, Darley J, Cervino S, Farrall M, Kelly P, Spector TD, Thein SL. 2000. Genetic influences on F cells and other hematologic variables: A twin heritability study. *Blood* **95:** 342–346.

Gaston M, Smith J, Gallagher D, Flournoy-Gill Z, West S, Bellevue R, Farber M, Grover R, Koshy M, Ritchey AK, et al. 1987. Recruitment in the Cooperative Study of Sickle Cell Disease (CSSCD). *Control Clin Trials* **8:** 131S–140S.

Gauderman WJ, Morrison JM. 2006. QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies. http://hydra.usc.edu/gxe.

Genovese G, Friedman DJ, Ross MD, Lecordier L, Uzureau P, Freedman BI, Bowden DW, Langefeld CD, Oleksyk TK, Uscinski Knob AL, et al. 2010. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329:** 841–845.

Haverfield EV, McKenzie CA, Forrester T, Bouzekri N, Harding R, Serjeant G, Walker T, Peto TE, Ward R, Weatherall DJ. 2005. UGT1A1 variation and gallstone formation in sickle cell disease. *Blood* **105:** 968–972.

Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6:** 95–108.

Hoppe C, Klitz W, Cheng S, Apple R, Steiner L, Robles L, Girard T, Vichinsky E, Styles L. 2004. Gene interactions and stroke risk in children with sickle cell anemia. *Blood* **103:** 2391–2396.

Hoppe C, Klitz W, D'Harlingue K, Cheng S, Grow M, Steiner L, Noble J, Adams R, Styles L. 2007. Confirmation of an association between the TNF(−308) promoter polymorphism and stroke risk in children with sickle cell anemia. *Stroke* **38:** 2241–2246.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42:** 348–354.

Kao WH, Klag MJ, Meoni LA, Reich D, Berthier-Schaad Y, Li M, Coresh J, Patterson N, Tandon A, Powe NR, et al. 2008. MYH9 is associated with nondiabetic end-stage renal disease in African Americans. *Nat Genet* **40:** 1185–1192.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308:** 385–389.

Kopp JB, Smith MW, Nelson GW, Johnson RC, Freedman BI, Bowden DW, Oleksyk T, McKenzie LM, Kajiyama H, Ahuja TS, et al. 2008. MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. *Nat Genet* **40:** 1175–1184.

Kwiatkowski JL, Hunter JV, Smith-Whitley K, Katz ML, Shults J, Ohene-Frempong K. 2003. Transcranial Doppler ultrasonography in siblings with sickle cell disease. *Br J Haematol* **121:** 932–937.

Lander ES. 2011. Initial impact of the sequencing of the human genome. *Nature* **470:** 187–197.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lettre G, Sankaran VG, Bezerra MA, Araujo AS, Uda M, Sanna S, Cao A, Schlessinger D, Costa FF, Hirschhorn JN, et al. 2008. DNA polymorphisms at the *BCL11A*, *HBS1L-MYB*, and *β-globin* loci associate with fetal hemoglobin

levels and pain crises in sickle cell disease. *Proc Natl Acad Sci* **105:** 11869–11874.

Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype imputation. *Annu Rev Genomics Hum Genet* **10:** 387–406.

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. 2003. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33:** 177–182.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461:** 747–753.

Mendonca TF, Oliveira MC, Vasconcelos LR, Pereira LM, Moura P, Bezerra MA, Santos MN, Araujo AS, Cavalcanti MS. 2010. Association of variant alleles of *MBL2* gene with vasoocclusive crisis in children with sickle cell anemia. *Blood Cells Mol Dis* **44:** 224–228.

Menzel S, Garner C, Gut I, Matsuda F, Yamaguchi M, Heath S, Foglio M, Zelenika D, Boland A, Rooks H, et al. 2007. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet* **39:** 1197–1199.

Milton JN, Sebastiani P, Solovieff N, Hartley SW, Bhatnagar P, Arking DE, Dworkis DA, Casella JF, Barron-Casella E, Bean CJ, et al. 2012. A genome-wide association study of total bilirubin and cholelithiasis risk in sickle cell anemia. *PLoS ONE* **7:** e34741.

* Nienhuis AW, Nathan DG. 2012. Pathophysiology and clinical manifestations of the β-thalassemias. *Cold Spring Harb Perspect Med* doi: 10.1101/cshperspect.a011726.

Nolan VG, Baldwin C, Ma Q, Wyszynski DF, Amirault Y, Farrell JJ, Bisbee A, Embury SH, Farrer LA, Steinberg MH. 2005. Association of single nucleotide polymorphisms in klotho with priapism in sickle cell anaemia. *Br J Haematol* **128:** 266–272.

Nolan VG, Adewoye A, Baldwin C, Wang L, Ma Q, Wyszynski DF, Farrell JJ, Sebastiani P, Farrer LA, Steinberg MH. 2006. Sickle cell leg ulcers: Associations with haemolysis and SNPs in Klotho, TEK and genes of the TGF-β/BMP pathway. *Br J Haematol* **133:** 570–578.

Nuinoon M, Makarasara W, Mushiroda T, Setianingsih I, Wahidiyat PA, Sripichai O, Kumasaka N, Takahashi A, Svasti S, Munkongdee T, et al. 2009. A genome-wide association identified the common genetic variants influence disease severity in β^0-thalassemia/hemoglobin E. *Hum Genet* **127:** 303–314.

Ofosu MD, Castro O, Alarif L. 1987. Sickle cell leg ulcers are associated with HLA-B35 and Cw4. *Arch Dermatol* **123:** 482–484.

Passon RG, Howard TA, Zimmerman SA, Schultz WH, Ware RE. 2001. Influence of bilirubin uridine diphosphate-glucuronosyltransferase 1A promoter polymorphisms on serum bilirubin levels and cholelithiasis in children with sickle cell anemia. *J Pediatr Hematol Oncol* **23:** 448–451.

Pilia G, Chen WM, Scuteri A, Orru M, Albai G, Dei M, Lai S, Usala G, Lai M, Loi P, et al. 2006. Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2:** e132.

Platt OS, Brambilla DJ, Rosse WF, Milner PF, Castro O, Steinberg MH, Klug PP. 1994. Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N Engl J Med* **330:** 1639–1644.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal Components Analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38:** 904–909.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* **5:** e1000519.

Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11:** 459–463.

Purcell S, Cherny SS, Sham PC. 2003. Genetic power calculator: Design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19:** 149–150.

Rasetti R, Weinberger DR. 2011. Intermediate phenotypes in psychiatric disorders. *Curr Opin Genet Dev* **21:** 340–348.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* **273:** 1516–1517.

* Schechter A, Elion J. 2012. *Cold Spring Harb Perspect Med* (to be published).

Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH. 2005. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nat Genet* **37:** 435–440.

Seldin MF, Pasaniuc B, Price AL. 2011. New approaches to disease mapping in admixed populations. *Nat Rev Genet* **12:** 523–528.

Sharan K, Surrey S, Ballas S, Borowski M, Devoto M, Wang KF, Sandler E, Keller M. 2004. Association of T-786C eNOS gene polymorphism with increased susceptibility to acute chest syndrome in females with sickle cell disease. *Br J Haematol* **124:** 240–243.

Slatkin M. 2008. Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9:** 477–485.

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Magi R, et al. 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* **42:** 937–948.

Steinberg MH, Forget BG, Higgs DR, Weatherall D. 2009. *Disorders of hemoglobin: Genetics, pathophysiology, and clinical management*, 2nd ed. Cambridge University Press, Cambridge.

Teer JK, Mullikin JC. 2010. Exome sequencing: The sweet spot before whole genomes. *Human Mol Genet* **19:** R145–R151.

Thein SL, Menzel S, Peng X, Best S, Jiang J, Close J, Silver N, Gerovasilli A, Ping C, Yamaguchi M, et al. 2007. Intergenic variants of *HBS1L–MYB* are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc Natl Acad Sci* **104:** 11346–11351.

Uda M, Galanello R, Sanna S, Lettre G, Sankaran VG, Chen W, Usala G, Busonero F, Maschio A, Albai G, et al. 2008. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin

and amelioration of the phenotype of β-thalassemia. *Proc Natl Acad Sci* **105:** 1620–1625.

Ulug P, Vasavda N, Awogbade M, Cunningham J, Menzel S, Thein SL. 2009. Association of sickle avascular necrosis with bone morphogenic protein 6. *Ann Hematol* **88:** 803–805.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Visscher PM, Medland SE, Ferreira MA, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin NG. 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* **2:** e41.

Visscher PM, Hill WG, Wray NR. 2008. Heritability in the genomics era—Concepts and misconceptions. *Nat Rev Genet* **9:** 255–266.

Voetsch B, Jin RC, Bierl C, Deus-Silva L, Camargo EC, Annichino-Bizacchi JM, Handy DE, Loscalzo J. 2008. Role of promoter polymorphisms in the plasma glutathione peroxidase (GPx-3) gene as a risk factor for cerebral venous thrombosis. *Stroke* **39:** 303–307.

Weatherall DJ. 2010. The inherited diseases of hemoglobin are an emerging global health burden. *Blood* **115:** 4331–4336.

Weatherall DJ, Clegg JB. 2001. Inherited haemoglobin disorders: An increasing global health problem. *Bull World Health Organ* **79:** 704–712.

Weatherall MW, Higgs DR, Weiss H, Weatherall DJ, Serjeant GR. 2005. Phenotype/genotype relationships in sickle cell disease: A pilot twin study. *Clin Lab Haematol* **27:** 384–390.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* **88:** 76–82.

Cold Spring Harbor Perspectives in Medicine
www.perspectivesinmedicine.org