REVIEW ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

Accounting for Linkage Disequilibrium in Association Analysis of Diverse Populations

Bashira A. Charles,* Daniel Shriner, and Charles N. Rotimi

Center for Research on Genomics and Global Health, National Human Genome Research Institute, Bethesda, Maryland, United States of America

Received 5 July 2013; Revised 14 November 2013; accepted revised manuscript 3 December 2013.
Published online 26 January 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21788

ABSTRACT: The National Human Genome Research Institute's catalog of published genome-wide association studies (GWAS) lists over 10,000 genetic variants collectively associated with over 800 human diseases or traits. Most of these GWAS have been conducted in European-ancestry populations. Findings gleaned from these studies have led to identification of disease-associated loci and biologic pathways involved in disease etiology. In multiple instances, these genomic findings have led to the development of novel medical therapies or evidence for prescribing a given drug as the appropriate treatment for a given individual beyond phenotypic appearances or socially defined constructs of race or ethnicity. Such findings have implications for populations throughout the globe and GWAS are increasingly being conducted in more diverse populations. A major challenge for investigators seeking to follow up genomic findings between diverse populations is discordant patterns of linkage disequilibrium (LD). We provide an overview of common measures of LD and opportunities for their use in novel methods designed to address challenges associated with following up GWAS conducted in European-ancestry populations in African-ancestry populations or, more generally, between populations with discordant LD patterns. We detail the strengths and weaknesses associated with different approaches. We also describe application of these strategies in follow-up studies of populations with concordant LD patterns (replication) or discordant LD patterns (transferability) as well as fine-mapping studies. We review application of these methods to a variety of traits and diseases.
Genet Epidemiol 38:265–273, 2014. © 2014 Wiley Periodicals, Inc.

KEY WORDS: genome-wide association study; African ancestry; fine-mapping

Introduction

Linkage disequilibrium (LD) refers to the tendency of alleles at linked loci to be inherited together [Pritchard and Przeworski, 2001]. In evolutionary terms, African-ancestry populations are the oldest of modern human populations. Compared to European- or Asian-ancestry populations, African-ancestry populations have experienced more generations for LD to decay [Tishkoff et al., 1996]. Furthermore, bottlenecks associated with Out-of-Africa migrations have led to reduced haplotypic diversity in European- and Asian-ancestry populations relative to African-ancestry populations (Fig. 1) [Campbell and Tishkoff, 2008]. Consequently, as a function of demography and time, African-ancestry individuals tend to have shorter segments of linked alleles whereas European- or Asian-ancestry individuals tend to have longer segments of linked alleles. The relative lengths of these segments create distinct patterns that can be leveraged in genomic studies.

Genome-wide association studies (GWAS) use tag variants to locate regions in the genome in which phenotype-associated variants reside. This strategy is based on the common disease-common variant hypothesis, which assumes that tag variants associated with a phenotype are in LD with causal variants [Cantor et al., 2010]. European- and Asian-ancestry populations with strong LD are ideal for conducting discovery analyses whereas African-ancestry populations with weak LD are better suited for fine-mapping [Gabriel et al., 2002]. African-ancestry populations are also well suited for discovery of novel loci, due to differences in LD structure and/or allele frequencies. Several studies provide examples in which associated variants were discovered in African-ancestry populations [Bensen et al., 2012; Dumitrescuet al., 2011; Freedman et al., 2010]. The application of LD to derive information regarding genetic epidemiology is not surprising given the fact that LD provides insight into the history of populations, including their mating patterns, geographic subpopulation structure, natural selection, gene conversion, mutation, and changes in allele frequencies over time [Weir, 2008].

GWAS are a valuable tool for establishing correlations between genetic variants and phenotypes, with human disease- or trait-associated single nucleotide polymorphisms (SNPs) exceeding 10,000 [Hindorff et al., 2009]. GWAS have led to identification of disease-associated loci and biologic pathways involved in disease development [Haines et al., 2005; Klein et al., 2005]. Some of these findings have led to the development of novel medical therapies [Ashtari et al., 2011; Hayes et al., 2007; Los et al., 2007] or evidence for

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Dr. Bashira Charles, Center for Research on Genomics and Global Health, National Human Genome Research Institute, Building 12A, Room 4047, 12 South Drive, MSC 5635, Bethesda, MD 20892, United States of America. E-mail: rotimic@mail.nih.gov

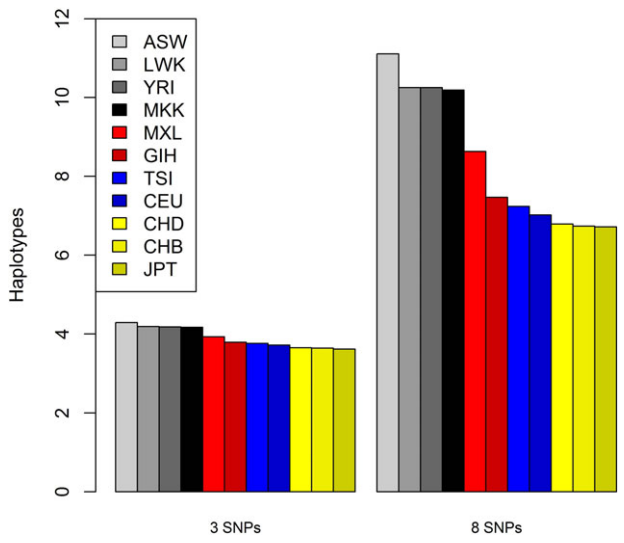

Figure 1. Global haplotype diversity. The maximum number of haplotypes for a window of three consecutive SNPs is eight and for eight SNPs is 256. All counted haplotypes have an estimated frequency ≥1% in the HapMap Phase III data. Population descriptors: ASW = African ancestry in Southwest USA; LWK = Luhya in Webuye, Kenya; YRI = Yoruba in Ibadan, Nigeria; MKK = Maasai in Kinyawa, Kenya; MXL = Mexican ancestry in Los Angeles, California; GIH = Gujarati Indians in Houston, Texas; TSI = Tuscan in Italy; CEU = Utah residents with Northern and Western European ancestry from the CEPH collection; CHD = Chinese in metropolitan Denver, Colorado; CHB = Han Chinese in Beijing, China; JPT = Japanese in Tokyo, Japan. As expected, African ancestry populations have more haplotype diversity.

prescribing a particular drug as the appropriate treatment for a given patient beyond phenotypic appearances or socially defined constructs of race or ethnicity [Rotimi and Jorde, 2010]. Such findings will have an impact on the treatment of disease in populations around the world. Although most GWAS have been conducted in European-ancestry populations [Need and Goldstein, 2009], more diverse populations are increasingly being included in GWAS [Adeyemo et al., 2009; Charles et al., 2011; Norris et al., 2009; Rich et al., 2009; Tin et al., 2011].

Coverage

The tag SNPs on a genome-wide association array have varying amounts of coverage based on the ancestral origins of the population upon which the array's development was based and the population upon which it is being used. LD is central in identifying the tag variants representing the genomes of various populations. Early SNP arrays were largely based on SNP discovery in individuals of European ancestry with the recognition that most of the tag variants would be portable to Asian-ancestry populations but that African-ancestry populations would require supplementation with additional markers [Conrad et al., 2006; Need and Goldstein, 2006]. Coverage of common variation (minor allele frequency ≥5%) in the HapMap Phase II data by these early GWAS panels at $r^2 \geq 0.08$ ranged from 75–86% for CEU samples to 28–49% for YRI samples [Barrett and Cardon, 2006]. Newer arrays including data from all three phases of HapMap as well as variants identified by the 1000 Genomes Project achieve coverage of SNPs with a minor allele frequency of ≥1% of 93%, 92%, and 76% for the CEU, CHB+JPT, and YRI populations, respectively [Delano et al., 2010]. Population-specific arrays such as the Axiom Genome-Wide PanAFR Array Plate Set achieve coverage of ∼90% for common variants and ∼85% for low-frequency variants (minor allele frequency ≥2%) for the African-ancestry HapMap YRI, LWK, MKK, and ASW samples [Affymetrix, 2011].

Tag variants on SNP arrays are chosen to reduce redundancy as measured by r^2. Genotypes for markers not on an array can be estimated using imputation, which is essentially the inverse of tagging. Imputation fills in missing genotypes by relying on a reference set of haplotypes; the weaker LD is, the less confidently missing genotypes can be imputed [Marchini and Howie, 2010]. Consequently, imputation performs worse in African-ancestry populations compared to European- or Asian-ancestry populations [Huang et al., 2005].

Commonly used measures of LD have been thoroughly reviewed in prior publications [Slatkin, 2008; Weir 2008]; however it is important that this content be briefly reviewed as groundwork for our discussion regarding uses of LD to interrogate the genome between diverse populations and its application to GWAS, GWAS follow-up, fine-mapping, and rare variant analysis. A brief review of measures of LD is provided in the Supporting Information Appendix A.

Challenges With Using LD to Interrogate the Genome

A replicate sample is defined as a random sample from the same population as the discovery sample. This definition implies that LD structure is concordant between samples. If a follow-up sample has discordant LD structure compared to the discovery sample, the definition of replication is not met and the term *transferability* is used [Shriner et al., 2009]. When transferability fails, the variant originally associated with the trait in the discovery population may be population specific [Shriner et al., 2009].

Many follow-up studies have resulted in failure to find association for the exact SNP identified in the discovery population. The phrase exact replication refers to this strategy [Clarke et al., 2007; Colhoun et al., 2003; Ioannidis et al., 2001; Ioannidis, 2007]. Some follow-up attempts have led to identification of different loci being associated with the disease or trait while others have led to the identification of different diseases or traits being associated with the loci in different populations or differences in the direction of the effect of the associated loci [Clarke et al., 2007; Ober, 2005]. These instances may be due in part to discordant LD structure between populations [Clarke et al., 2007; Ong and Teo, 2010] and allelic, genetic, or trait heterogeneity [Clarke et al., 2007]. As investigators began to examine findings from GWAS from diverse populations, discordant LD structure between populations would often yield false negative findings, leaving

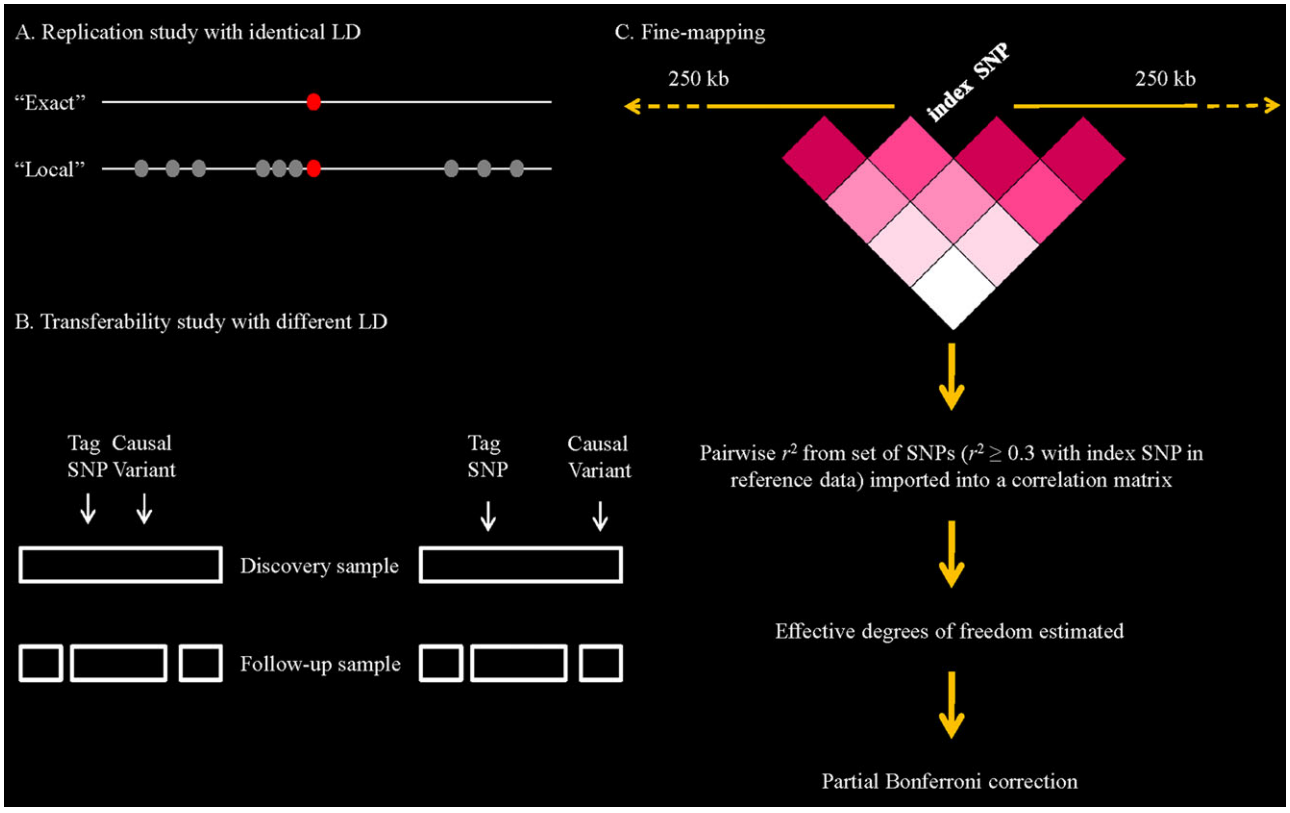

Figure 2. Transferability of genomic findings. Schematic diagram of different linkage disequilibrium patterns in discovery and follow-up samples. (A) The associated tag SNP and a causal variant are in the same region of linkage disequilibrium in both the discovery and follow-up samples. Both the exact and local approaches may yield successful transfer. (B) The associated tag SNP and a causal variant are not in the same region of linkage disequilibrium in both the discovery and follow-up samples. The exact approach will fail for the original tag SNP but the local approach may succeed for other proxy SNPs if they are in the same region of linkage disequilibrium as the causal variant in the follow-up sample [Shriner et al., 2009]. (C) For a given index SNP, all possible proxy SNPs are identified. The matrix of pairwise correlation between all pairs of SNPs is decomposed and used to estimate a multiple comparisons correction factor that fully accounts for linkage disequilibrium.

investigators with the impression that there has been failure to replicate (Fig. 2).

Due to these challenges, investigators have begun to develop other approaches to assess GWAS findings [Shriner et al., 2009]. GWAS findings are conventionally reported with respect to index SNPs at independent loci [The Wellcome Trust Case Control Consortium, 2007]. Even though an associated locus is often referred to by the index SNP, an associated locus often contains multiple associated SNPs due to the presence of LD. Methods for *local replication* include evaluation of proxy SNPs, which are in LD with the originally identified index SNPs [Clarke et al., 2007].

Leveraging Reference LD

One way to identify proxies for local replication or transferability analysis is to use external reference data (e.g. HapMap or 1000 Genomes). Sets of proxy SNPs for an index SNP can be defined based on physical distance, genetic distance, or LD. These approaches have led to more powerful assessment of follow-up studies as well as improved variant annotation. Furthermore, follow-up of associations in populations with weaker LD offers the opportunity to fine map associations. Genomic findings gleaned from the use of these methodologies illustrating the power of the approach were initially documented in 2009 when 107 independent loci associated with stature in European- or Asian-ancestry populations had been reported in the literature. In a follow-up study comprised of African Americans, sets of proxy variants were bounded by pairwise correlation with the index SNP of $r^2 \geq 0.3$ within 250 kb of the index SNP [Shriner et al., 2009]. The exact replication approach yielded association for only seven loci whereas the transferability approach yielded association for 53 loci [Shriner et al., 2009]. A notable limitation was coverage, which ranged from 71.2% and 75.8% for variants identified in the HapMap CEU and CHB samples, respectively [Shriner et al., 2009].

The advantage of this approach for fine mapping was initially identified in a GWAS of serum uric acid in African Americans. The gene *SLC2A9* that had been associated with serum uric acid along with several other genes in multiple European-ancestry samples [Caulfield et al., 2008; Kamatani et al., 2010; van der Harst et al., 2010; Wallace

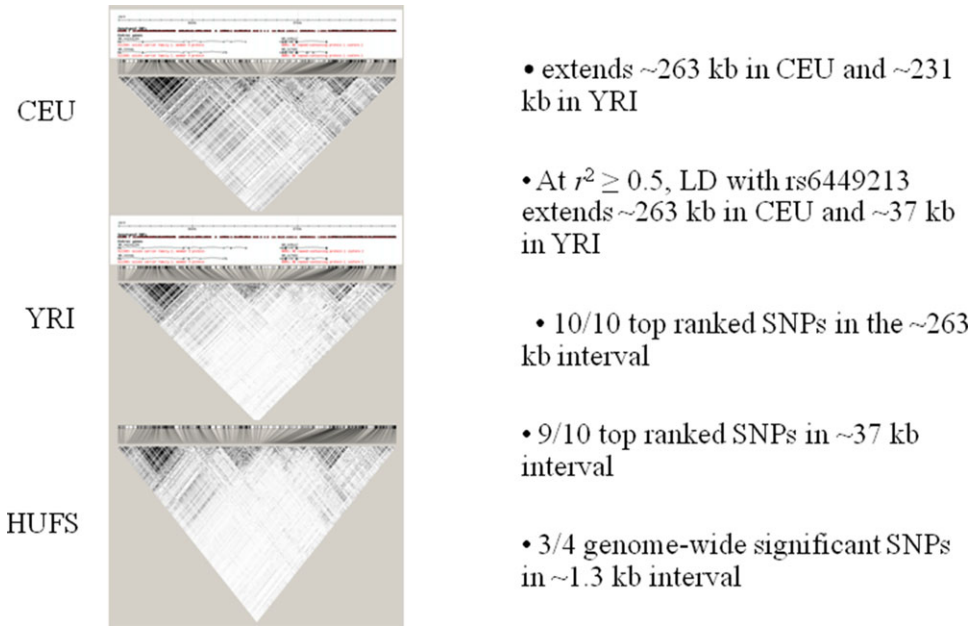

Figure 3. Comparison of linkage disequilibrium in the HapMap Phase II CEU, YRI, and HUFS sample (African Americans). Using SNPs with $r^2 \geq 0.5$, the LD interval encompassing the index SNP identified in the European-ancestry discovery population is larger in the HapMap CEU sample than the LD interval encompassing the index SNP in the HapMap YRI sample. Taking advantage of this discordant pattern of LD results in a reduction from 263 kb in CEU to 37 kb in YRI to 1.3 kb for 75% of the genome-wide significant SNPs in the African American sample, illustrating fine-mapping when LD is weaker in the follow-up population [Ashtari et al., 2011].

et al., 2008] was the only gene identified in the African American sample. The association of the index SNP rs6449123 in *SLC2A9* with serum uric acid was fine mapped in the African American, CEU and YRI samples [Charles et al., 2011]. The LD pattern for the African American sample was almost identical to that of the YRI sample. At a threshold of $r^2 \geq 0.5$ for SNPs in LD with rs6449123, the associated interval was 263 kb in the CEU sample while in the African American sample the associated interval was 37 kb (Fig. 3) [Charles et al., 2011]. Three of four SNPs associated with genome-wide significance in the African American sample lay within a 1.3 kb interval [Charles et al., 2011]. These results provide compelling evidence of the advantage of following up GWAS signals originally identified in populations with strong LD (European ancestry) in populations with weak LD (African ancestry). Similar methodologies have been used for fasting plasma glucose, coronary artery disease, and type 2 diabetes and are described below.

Fasting Plasma Glucose

Exact and local methods were applied to a study of fasting plasma glucose in African Americans [Ramos et al., 2011]. Replication was assessed for 20 SNPs and an additional 9 SNPs were assessed using local methods [Ramos et al., 2011]. The region of assessment for local transferability was determined using $r^2 \geq 0.3$ for SNPs within 250 kb of the index SNPs [Ramos et al., 2011]. SNPs were selected using the HapMap LD data from the CEU population with the exception of rs2166706, for which the reference sample was GIH to account for the discovery of this association in Asian Indians [Ramos et al., 2011]. A sample covariance matrix based on genotypes was decomposed and the effective degrees of freedom were denoted as $N_{eff} = \frac{(\sum_{k=1}^{K} \lambda_k)^2}{\sum_{k=1}^{K} \lambda_k^2}$, in which λ_k is the k^{th} eigenvalue from the $k \times k$covariance matrix for the K SNPs in the region [Bretherton et al., 1999]. The threshold for significance was defined as $\alpha = \frac{0.05}{N_{eff}}$, corresponding to a partial Bonferroni correction rather than the full Bonferroni correction of $\alpha = \frac{0.05}{K}$ since $1 \leq N_{eff} \leq K$ [Ramos et al., 2011]. Using this methodology, associations for SNPs in or near five genes were replicated and SNPs in four additional genes were found to be associated using transferability analysis [Ramos et al., 2011].

Coronary Heart Disease

Lettre et al. [2011] fine mapped in African Americans five loci discovered in European-ancestry individuals as associated with coronary artery disease and its risk factors. SNPs correlated ($r^2 \geq 0.5$) to the index SNP in the HapMap CEU sample but uncorrelated ($r^2 \leq 0.1$) in the HapMap YRI sample were evaluated [Lettre et al., 2011]. For each of these five loci, markers more strongly associated with the phenotype than the index SNPs were discovered in the African American sample [Lettre et al., 2011]. As an example, the index SNP rs174547 from a European-ancestry population is in strong LD with rs1535, the most strongly associated SNP in the locus

among African Americans. The two markers are in strong LD in the HapMap CEU sample ($r^2 = 1$) but not in the HapMap YRI sample ($r^2 \geq 0.09$) [Lettre et al., 2011]. LD surrounding rs174547 extends over 113 kb in the HapMap CEU sample but rs1535 is not strongly correlated with any other marker in the HapMap YRI sample [Lettre et al., 2011].

Type 2 Diabetes

Palmer et al. [2012] evaluated 37 index SNPs associated with type 2 diabetes in African Americans. Sets of SNPs surrounding the index SNPs were based on LD blocks defined by D' using the HapMap CEU data [Gabriel et al., 2002]. Locus-specific association P-values were adjusted for the effective number of SNPs as calculated using the method of Li and Ji [2005]. One index SNP was significantly associated using the exact approach and one additional SNP was significantly associated using a local approach [Palmer et al., 2012].

A worked example for serum bilirubin [Chen et al., 2011; Johnson et al., 2009; Kang et al., 2010; Sanna et al., 2009) is provided in Appendix 2 Text Box.

Contrasting LD in Two Samples

If correlations between a pair of SNPs in two populations result in D' values that are positive in one sample and negative in another sample, both the type I and type II error rates in association studies can increase [Teo et al., 2009b]. To quantify variation in LD between samples, Teo et al. devised a signed r^2 statistic defined as $(-1)^{\mathrm{I}(P_{AB}<P_AP_B)}\frac{(P_{AB}-P_AP_B)^2}{P_AP_aP_BP_b}$, in which $\mathrm{I}(P_{AB} < P_AP_B)$ is an indicator function equaling one when true and equaling zero otherwise [Ong and Teo, 2010; Teo et al., 2009a; Teo et al., 2009b]. This statistic is the conventional r^2 statistic with the sign of D'. Matrices M_1 and M_2 are constructed using the signed r^2 values between all pairs of SNPs for samples 1 and 2. Let Δ_i represent the diagonal matrix of sorted eigenvalues from the decomposition of M_i. The VarLD score is defined as trace $|\Delta_1 - \Delta_2|$, with significance of the score assessed by sampling with replacement [Teo et al., 2009a]. Examples of regions for which the VarLD score indicates variation in LD between samples include *VKORC1*, *NRG1*, *LCT*, *SLC24A5*, *HBB*, and MHC [Ong and Teo, 2010].

The VarLD score can empirically demonstrate that LD variation in specific genomic regions is associated with variation in haplotype frequencies and that nontransferability may be more common in these regions [Teo et al., 2009a]. Additionally, loci under different selection pressures in two populations are more likely to have discordant LD patterns between populations [Teo et al., 2010]. A weakness of the VarLD score is exclusion of the eigenvectors, which leads to a loss of power. To illustrate, suppose two samples have LD patterns that are mirror images (Fig. 4). The VarLD score is 0 because the sorted eigenvalues are identical. However, if the eigenvectors are first matched by pairing each eigenvector from sample 1 with the most closely parallel eigenvector from sample 2 (based on the angle between two vectors equaling the arccosine of the dot product of two orthonormal eigenvectors), then the VarLD score reaches its maximum value. Krzanowski used a two-stage approach assessing both eigenvalues and eigenvectors to assess the equality of matrices of the correlation coefficients [Krzanowski, 1979; Krzanowski, 1993]. Zaykin also uses this two-stage approach, while calculation of the VarLD score includes only assessment of the eigenvalues (calculating the absolute difference of the eigenvalues). There are many well-established tests of equality that have been described in textbooks on multivariate statistical analysis [Anderson, 1958], however, we have not observed instances of their application to use of LD to interrogate the genomes of diverse populations during this review. It is not clear why these classical approaches have not been used in the field to date.

Multivariate Normality

Zaitlen et al. [2010] described an approach for testing association for multiple correlated SNPs from diverse populations. For each study, association Z scores are assumed to follow the multivariate normal distribution, $\mathbf{Z} \sim \mathrm{MVN}(\boldsymbol{\lambda}\sqrt{\mathbf{n}}, \boldsymbol{\Sigma})$, in which $\boldsymbol{\lambda}\sqrt{\mathbf{n}}$ is the vector of noncentrality parameters and $\boldsymbol{\Sigma}$ is the matrix of haplotype-based correlation coefficients [Zaitlen et al., 2010]. The Z scores are weighted across studies, yielding test statistics $\mathbf{T}$ that are also multivariate normal, $\mathrm{T} \sim \mathrm{MVN}(\frac{\sum \sqrt{n}\boldsymbol{\lambda}\sqrt{n}}{\sqrt{\sum n}}, \frac{\sum n\boldsymbol{\Sigma}}{\sum n})$ [Zaitlen et al., 2010]. Clark et al. [2010] described a Bayesian test comparing the hypothesis that two covariance matrices are equal to the hypothesis that two covariance matrices are different. Due to the assumption of multivariate normality and the use of conjugate priors, the Bayesian test can be computed efficiently in closed form and therefore does not require permutation testing.

Metaanalysis

Teo et al. [2010] compared inverse-variance weighted fixed effects meta-analysis with Fisher's combined P-value to fine map disease-associated variants. Between-study heterogeneity (effect size variation) violates the assumption of uniformity of effect sizes required for a fixed effects model to be most powerful [Cochran, 1954; de Bakker, Ferreira et al., 2008; Mantel and Haenszel, 1959]. For populations with discordant LD structure, a random effects model is more appropriate than either of these approaches. However, the commonly employed random effects model effectively assumes between-study heterogeneity under the null hypothesis and yields conservative P-values compared to the fixed effects model. Inclusion of between-study heterogeneity in the alternative hypothesis, rather than in the null hypothesis, results in greater power for detecting association but is also more sensitive to confounding [Han and Eskin, 2011].

Morris developed a Bayesian partition model in which populations are clustered based on shared ancestry [Morris, 2011]. This model is more powerful than the fixed effects model if there is heterogeneity and more powerful than the random effects model if populations are related. Potential effects of LD are not formally part of this model but are

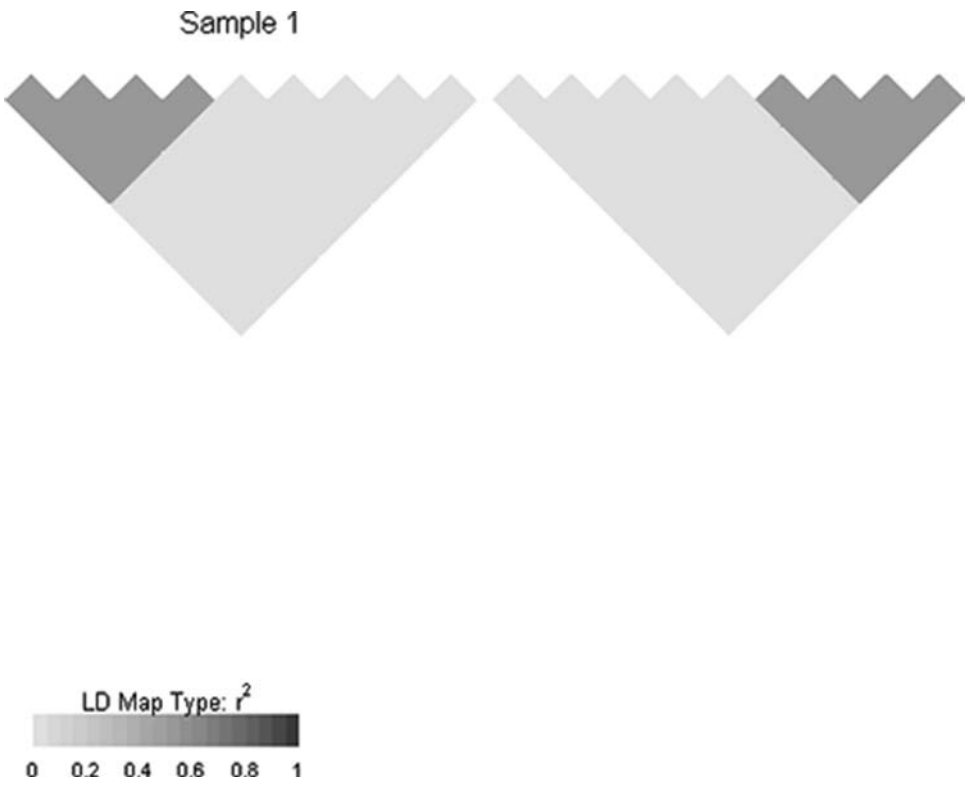

Figure 4. Error associated with using the trace of the eigenvalues without matching the eigenvectors. Sample 1 consists of 10 SNPs with pairwise $r^2 = 0.5$ for all pairs of SNPs 1–5 and $r^2 = 0$ for all other pairs of SNPS. Sample 2 is the mirror image of sample 1, with pairwise $r^2 = 0.5$ for all pairs of SNPs 6–10 and $r^2 = 0$ for all other pairs of SNPs. Without matching the eigenvectors, the sorted eigenvalues are identical.

indirectly captured through cluster-specific allelic effects. Using this approach three of 19 loci associated with type 2 diabetes showed evidence of heterogeneity in allelic effects among European American, African American, Hispanic, Japanese American, and Native Hawaiian ancestry [Morris, 2011].

Rare Variant Analysis

The common disease-common variant hypothesis is the paradigm upon which GWAS were initially designed; however, variants identified by most GWAS account for a small fraction of the heritability of complex diseases under investigation. The common disease-rare variant hypothesis posits that single or multiple rare variants with moderate to large effects are the source of the some of the missing heritability. While a review of techniques for rare variant analysis is beyond our scope, two points are relevant. First, a causal rare variant that is in strong LD with a noncausal common variant may result in the (indirect) association of the noncausal common variant with the phenotype [Dickson et al., 2010]. Second, some methods for rare variant analysis assume that rare variants are not in LD whereas other methods relax this assumption, including methods designed to jointly assess common and rare variants [Shriner and Vaughan, 2011; Yuan et al., 2012].

Discussion

In this review, we have provided a comprehensive resource of statistical methods utilizing the most common measures of LD developed for addressing the challenges and opportunities associated with following-up in African-ancestry populations, GWAS initially conducted in European-ancestry populations or, more generically, between populations with discordant LD patterns. We provided examples of the application of these procedures to phenotypes such as coronary heart disease, fasting plasma glucose, serum uric acid, stature, bilirubin, and type 2 diabetes. We also described the strengths and weaknesses of each approach (Table 1). This information provides investigators with a resource for selecting the most appropriate analytic method based on their sample and research question when seeking to compare populations with discordant LD patterns.

Each method adds valuable information to the literature. The LD contrast test was designed to compare pairwise LD matrices between cases and controls while maintaining robustness to departure from HWE. The LD contrast test has been modified for case-parent data, for quantitative traits, and to eliminate confounding caused by background LD. The VarLD score quantifies differences in LD patterns between two samples but requires knowledge of phase and is underpowered. Transferability analysis is an extension of

Table 1. Select methods for leveraging LD in populations with discordant LD patterns

Test	LD contrast test[49]	Modified LD contrast[51]	VarLD[75]	Transferability test[53]	Transethnic fine-mapping[60]	Localization success rate[77]	Bayesian partitioning[82]
Study Design	Case/control	Case/control Continuous traits	Case/control Continuous traits	Case/control Continuous traits	Metaanalysis	Metaanalysis	Metaanalysis
Population LD pattern	Concordant	Discordant	Discordant	Concordant/Discordant	Discordant	Discordant	Concordant/Discordant
Data/Methods	Tag SNP markers, GWAS, sequence	Tag SNP markers, GWAS, sequence	Tag SNP markers, GWAS, sequence	Tag SNP markers, GWAS, sequence	Fine-mapping	Tag SNP Markers fine-mapping	GWAS, Fine-mapping
Strengths	Increased power compared to haplotype-based test or multilocus tests based on SNP scores in populations with concordant LD between cases and controls Robust to departures from HWE Test statistic for comparing LD plots	Increased power Accounts for confounding due to background LD	Models correlation between regional genomic variation and haplotype diversity Models correlation between selection pressure and LD Robust to departures allelic homogeneity of effect	Improved variant localization Can be employed for use with exact, or local replication, transferability and fine mapping Can applied when comparing multiple samples with different ancestral backgrounds Ease of use	Improved variant localization	Improved variant localization Software available	Robust to departures from allelic heterogeneity of effect Models between-population allelic heterogeneity Increased power compared to random and fixed effects models Software available by request
Weaknesses	Less powerful than haplotype-based tests/multilocus tests when the case/control populations have discordant LD Requires permutation testing Confounding due to background LD	NA	Eigenvectors exclusion resulting in loss of power Non-GWAS data requires permutation Indels, variation in coverage or genotyping quality decrease power	None	Between-study heterogeneity violates assumption of homogeneity resulting in loss of power	Violation of the assumption of multivariate normality results in decreased power	NA

replication analysis designed to overcome the challenges associated with transferring results of a GWAS from a discovery sample to a follow-up sample with discordant LD. Depending on the samples under investigation, transferability analysis may also be used for fine mapping. This approach often results in remarkable resolution when contrasting genomic findings from European-ancestry populations with findings from African-ancestry populations [Charles et al., 2011; Chen et al., 2011; Ramos et al., 2011]. Random effects models of metaanalyses have been described but have not yet achieved widespread use.

An increasing number of genomics projects, such as the 1000 Genomes Project, the Singapore Genome Variation Project, and the Human Heredity and Health in Africa Initiative (H3Africa.org), include non-European populations. This type of research requires strategies for addressing the analytic challenges spawned from the diversity inherent in human genomes. Resources presented in this review will help to ensure that results are not marred by false positive or false negative findings. The analytic methods incorporating LD described in this review are applicable to the analysis of both genotype and sequence data and should maintain their relevance for the foreseeable future.

Acknowledgments

Grants and fellowships supporting the writing of the paper: Intramural Research Program of the Center for Research on Genomics and Global Health supported by the National Human Genome Research Institute (NHGRI), the National Institute of Diabetes and Digestive and Kidney Diseases, the Center for Information Technology, the Office of the Director at the National Institutes of Health (Z01HG200362) and the NHGRI Health Disparity Postdoctoral Fellowship.

The authors have no conflicts of interest to disclose.

References

Adeyemo A, Gerry N, Chen G, Herbert A, Doumatey A, Huang H, Zhou J, Lashley K, Chen Y, Christman M and others. 2009. A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genet* 5(7):e1000564.

Affymetrix. 2011. *SNP genotyping using the Affymetrix Axiom Genome-Wide Pan-African (PanAFR) Array Set. Affymetrix Inc. Affymetrix.* Affymetrix, Santa Clara.

Anderson TW. 1958. *Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Ashtari M, Cyckowski LL, Monroe JF, Marshall KA, Chung DC, Auricchio A, Simonelli F, Leroy BP, Maguire AM, Shindler KS and others. 2011. The human visual cortex responds to gene therapy—mediated recovery of retinal function. *J Clin Invest* 121(6):2160–2168.

Barrett JC, Cardon LR. 2006. Evaluating coverage of genome-wide association studies. *Nat Genet* 38(6):659–662.

Bensen JT, Xu Z, Smith GJ, Mohler JL, Fontham ETH and Taylor JA. 2012. Genetic polymorphism and prostate cancer aggressiveness: a case-only study of 1,536 GWAS and candidate SNPs in African-Americans and European-Americans. *The Prostate* 73(1):11–22.

Bretherton CS, Widmann M, Dymnikov VP, Wallace JM and Bladé I. 1999. The effective number of spatial degrees of freedom of a time-varying field. *J Climate* 12:1990–2009.

Campbell MC, Tishkoff SA. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genom Hum Genet* 9:403–433.

Cantor RM, Lange K and Sinsheimer JS. 2010. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum Genet* 86:6–22.

Caulfield MJ, Munroe PB, O'Neill D, Witkowska K, Charchar FJ, Doblado M, Evans S, Eyheramendy S, Onipinla A, Howard P and others. 2008. SLC2A9 is a high-capacity urate transporter in humans. *PLoS Med* 5(10):e197.

Charles B, Shriner D, Doumatey A, Chen G, Zhou J, Huang H, Herbert A, Gerry N, Christman M, Adeyemo A and others. 2011. A genome-wide association study of serum uric acid in African Americans. *BMC Med Genomics* 4(1):17.

Chen G, Ramos E, Shriner D, Doumatey A, Gerry N, Herbert A, Huang H, Zhou J, Christman M, Adeyemo A and others. 2011. UGT1A1 is a major locus influencing bilirubin levels in African Americans. *Eur J Hum Genet* 20(4):463–468.

Clark TG, Campino SG, Anastasi E, Auburn S, Teo YY, Small K, Rockett KA, Kwiatkowski DP and Holmes CC. 2010. A Bayesian approach using covariance of single nucleotide polymorphism data to detect differences in linkage disequilibrium patterns between groups of individuals. *Bioinformatics* 26(16): 1999–2003.

Clarke GM, Carter KW, Palmer LJ, Morris AP and Cardon LR. 2007. Fine mapping versus replication in whole-genome association studies. *Am J Hum Genet* 81:995–1005.

Cochran WG. 1954. The combination of estimates from different experiments. *Biometrics* 10(1):101–129.

Colhoun HM, McKeigue PM and Smith GD. 2003. Problems of reporting genetic associations with complex outcomes. *Lancet* 361:865–872.

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA and Pritchard JK. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 38(11):1251–1260.

de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S and Voight Benjamin F. 2008. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 17(R2):R122–R128.

Delano D, Eberle M and Rosenow C. 2010. *Array Differences in Genomic Coverage and Data Quality Impact GWAS Success.* Illumina, Inc., San Diego.

Dickson SP, Wang K, Krantz I, Hakonarson H and Goldstein DB. 2010. Rare variants create synthetic genome-wide associations. *PLoS Biol* 8(1):e1000294.

Dumitrescu L, Carty CL, Taylor K, Schumacher FR, Hindorff LA, Ambite JL, Anderson G, Best LG, Brown-Gentry K, Bůžková P and others. 2011. Genetic determinants of lipid traits in diverse populations from the population architecture using genomics and epidemiology (PAGE) study. *PLoS Genet* 7(6): e1002138.

Freedman BI, Kopp JB, Langefeld CD, Genovese G, Friedman DJ, Nelson GW, Winkler CA, Bowden DW and Pollak MR. 2010. The apolipoprotein L1 (APOL1) gene and nondiabetic nephropathy in African Americans. *J Am Soc Nephrol* 21(9):1422–1426.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart, Maura and others. 2002. The structure of haplotype blocks in the human genome. *Science* 296(5576):2225–2229.

Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR and others. 2005. Complement factor H variant increases the risk of age-related macular degeneration. *Science* 308(5720): 419–421.

Han B, Eskin E. 2011. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* 88(5): 586–598.

Hayes DF, Thor AD, Dressler LG, Weaver D, Edgerton S, Cowan D, Broadwater G, Goldstein, Martino LJ, Ingle S, Ingle JN and others. 2007. HER2 and response to paclitaxel in node-positive breast cancer. *New Engl J Med* 357(15):1496–1506.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS and Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* 106(23): 9362–9367.

Huang W, He Y, Wang, H, Wang Y, Liu Y, Wang Y, Chu X, Wang Y, Xu L, Shen, Y and others. 2005. Linkage disequilibrium sharing and haplotype-tagged SNP portability between populations. *Proc Natl Acad Sci* 103(5):1418–1421.

Ioannidis JP. 2007. Non-replication and Inconsistency in the genome-wide association setting. *Hum Hered* 64:203–213.

Ioannidis JPA, Ntzani EE, Trikalinos TA and Contopoulos-Ioannisis DG. 2001. Replication validity of genetic association studies. *Nat Genet* 29:306–309.

Johnson AD, Kavousi M, Smith AV, Chen MH, Dehghan A, Aspelund T, Lin JP, van Duijn CM, Harris TB, Cupples LA and others. 2009. Genome-wide association meta-analysis for total serum bilirubin levels. *Hum Mol Genet* 18(14): 2700–2710.

Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, Nakamura Y and Kamatani N. 2010. Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat Genet* 42(3):210–215.

Kang TW, Kim HJ, Ju H, Kim JH, Jeon YJ, Lee HC, Kim KK, Kim JW, Lee S, Kim JY and others. 2010. Genome-wide association of serum builirubin levels in Korean population. *Hum Mol Genet* 19(18):3672–3678.

Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST and others. 2005. Complement Factor H Polymorphism in age-related macular degeneration. *Science* 308(5720):385–389.

Krzanowski W. 1979. Between-groups comparison of principal components. *J Am Stat Assoc* 74(367):703–707.

Krzanowski W. 1993. Permutational tests for correlation matrices. *Stat Comput* 3:37–44.

Lettre G, Palmer CD, Young T, Ejebe KG, Allayee H, Benjamin EJ, Bennett F, Bowden DW, Chakravarti A, Dreisbach A and others. 2011. Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. *PLoS Genet* 7(2):e1001300.

Li J, Ji L. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 95(3):221–227.

Los M, Roodhart JML and Voest EE. 2007. Target practice: lessons from phase III trials with bevacizumab and vatalanib in the treatment of advanced colorectal cancer. *Oncologist* 12(4):443–450.

Mantel N, Haenszel W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 22(4):719–748.

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11(7):499–511.

Morris AP. 2011. Trans-ethnic meta-analysis of genome-wide association studies. *Genet Epidemiol* 35(8):809–822.

Need AC, Goldstein DB. 2006. Genome-wide tagging for everyone. *Nat Genet* 38(11):1227–1228.

Need AC, Goldstein DB. 2009. Next generation disparities in human genomics: concerns and remedies. *Trends Genet* 25(11):489–494.

Norris JM, Langefeld CD, Talbert ME, Wing MR, Haritunians T, Fingerlin TE, Hanley AJG, Ziegler JT, Taylor KD, Haffner SM and others. 2009. Genome-wide association study and follow-up analysis of adiposity traits in hispanic Americans: the IRAS family study. *Obesity* 17(10):1932–1941.

Ober C. 2005. Perspectives on the past decade of asthma genetics. *J Allergy Clin Immunol* 116(2):274–278.

Ong RT-H, Teo, Y-Y. 2010. varLD: a program for quantifying variation in linkage disequilibrium patterns between populations. *Bioinformatics* 26(9):1269–1270.

Palmer ND, McDonough CW, Hicks PJ, Roh BH, Wing MR, An SS, Hester JM, Cooke JN, Bostrom MA, Rudock ME and others. 2012. A genome-wide association search for type 2 diabetes genes in African Americans. *PLoS One* 7(1):e29202.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 69(1):1–14.

Ramos E, Chen G, Shriner D, Doumatey A, Gerry N, Herbert A, Huang H, Zhou J, Christman M, Adeyemo A and others. 2011. Replication of genome-wide association studies (GWAS) loci for fasting plasma glucose in African-Americans. *Diabetologia* 54(4):783–788.

Rich S, Goodarzi M, Palmer N, Langefeld C, Ziegler J, Haffner S, Bryer-Ash M, Norris J, Taylor K, Haritunians T and others. 2009. A genome-wide association scan for acute insulin response to glucose in Hispanic-Americans: the Insulin Resistance Atherosclerosis Family Study (IRAS FS). *Diabetologia* 52(7):1326–1333.

Rotimi CN, Jorde LB. 2010. Ancestry and disease in the age of genomic medicine. *New Engl J Med* 363(16):1551–1558.

Sanna S, Busonero F, Maschio A, McArdle PF, Usala G, Dei M, Lai S, Mulas A, Piras MG, Perseu L and others. 2009. Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia. *Hum Mol Genet* 18(14):2711–2718.

Shriner D, Adeyemo A, Gerry NP, Herbert A, Chen G, Doumatey A, Huang H, Zhou J, Christman MF, Rotimi CN and others. 2009. Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. *PLoS One* 4(12):e8398.

Shriner D, Vaughan LK. 2011. A unified framework for multi-locus association analysis of both common and rare variants. *BMC genomics* 12:89.

Slatkin M. 2008. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9(6):477–485.

Teo Y-Y, Ong RTH, Sim X, Tai E and Chia K-S. 2010. Identifying candidate causal variants via trans-population fine-mapping. *Genet Epidemiol* 34(7):653–664.

Teo YY, Fry AE, Bhattacharya K, Small KS, Kwiatkowski DP and Clark TG. 2009a. Genome-wide comparisons of variation in linkage disequilibrium. *Genome Res* 19(10):1849–1860.

Teo YY, Small KS, Fry AE, Wu Y, Kwiatkowski DP and Clark TG. 2009b. Power consequences of linkage disequilibrium variation between populations. *Genet Epidemiol* 33(2):128–135.

The Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145):661–678.

Tin A, Woodward OM, Kao WHL, Liu C-T, Lu X, Nalls MA, Shriner D, Semmo M, Akylbekova EL, Wyatt SB and others. 2011. Genome-wide association study for serum urate concentrations and gout among African Americans identifies genomic risk loci and a novel URAT1 loss-of-function allele. *Hum Mol Genet* 307:1–13.

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M and others. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 271(5254):1380–1387.

van der Harst P, Bakker SJ, de Boer RA, Wolffenbuttel BH, Johnson T, Caulfield MJ and Navis G. 2010. Replication of the five novel loci for uric acid concentrations and potential mediating mechanisms. *Hum Mol Genet* 19(2):387–395.

Wallace C, Newhouse SJ, Braund P, Zhang F, Tobin M, Falchi M, Ahmadi K, Dobson RJ, Marcano AC, Hajat C and others. 2008. Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am J Hum Genet* 82(1):139–149.

Weir BS. 2008. Linkage disequilibrium and association mapping. *Ann Rev Genomics Hum Genet* 9:129–142.

Yuan A, Chen G, Zhou Y, Bentley A and Rotimi C. 2012. A novel approach for the simultaneous analysis of common and rare variants in complex traits. *Bioinformatics Biol Insights* 6:1–9.

Zaitlen N, Pasaniuc B, Gur T, Ziv E and Halperin E. 2010. Leveraging genetic variability across populations for the identification of causal variants. *Am J Hum Genet* 86:23–33.