Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*Curr Protoc Hum Genet*. Author manuscript; available in PMC 2017 October 11.

Published in final edited form as:

*Curr Protoc Hum Genet*. ; 91: 1.30.1–1.30.10. doi:10.1002/cphg.25.

# Analysis of Heritability Using Genome-wide Data

**Jacob B. Hall**[1] and **William S. Bush**[1]

[1]Institute for Computational Biology, Case Western Reserve University, OH

## Abstract

Most analyses of genome-wide association data consider each variant independently without considering or adjusting for the genetic background present in the rest of the genome. New approaches to genome analysis use representations of genomic sharing to better account for confounding factors like population stratification, or to directly approximate heritability through the estimated sharing of individuals in a dataset. These approaches use mixed linear models, which relate genotypic sharing to phenotypic sharing, and rely on the efficient computation of genetic sharing among individuals in a dataset. This unit describes the principles and practical application of mixed-models for the analysis of genome-wide association study data.

### Keywords

mixed-model analysis; heritability; GCTA

## KEY CONCEPTS

Genetic analyses are ultimately about sharing, with the fundamental question being “does sharing of genetic variation relate to sharing of phenotypic variation?”. Early geneticists relied on the principles of Mendelian segregation and independent assortment (Unit 1.4) [*Copy Editor: Here and throughout, the authors add helpful cross references to other CPHG units. Pleas keep these, but also add the equivalent literature citation to the CPHG units.] to generate a theoretical distribution of genetic sharing within related individuals. Based on these estimates of genetic sharing, and the observed rate of sharing of a phenotype, investigators can quantify the proportion of variance for a phenotype that is explained by sharing of genomic regions, also known as the trait *heritability*.

Twin studies are especially useful for heritability estimation because twins are exposed to nearly the same environmental factors (including shared intrauterine environment, parenting style, wealth, culture, and time period). This reduces envronmental variability thus genetic factors can be better interpreted and quantified. Using Falconer’s formula (Falconer et al., 1996) heritability can be calculated as twice the difference between monozygotic and dizygotic twin correlation ($r_{mz}$ and $r_{dz}$, respectively), as shown in Equation 1.

**Contact Author**. Jake Hall, Case Western Reserve University - Epidemiology & Biostatistics 2530 Wolstein Research Building 2103 Cornell Road, Cleveland, Ohio 44106, jakehall@case.edu.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$H^2 = 2(r_{mz} - r_{dz}) \quad (1)$$

Twin studies are still considered the “gold standard” for estimating heritability in humans, but this study design poses interesting challenges. For example, ascertainment requires infrastructure and registries for gathering larger numbers of twins and performing phenotyping (Boomsma et al., 2002). Furthermore, while some efforts have been proposed (Rahmioğlu and Ahmadi, 2010), difficult phenotypes such as drug response are often not amenable to this design due to the ethics of exposing individuals to medication without clinical need. Also due to disease prevalance, some phenotypes are probabilistically difficult to ascertain within twins.

Since 2005, genotyping technologies (and now DNA sequencing) have allowed investigators to directly capture much of the genetic variation within an individual (Abecasis et al., 2012; Barrett and Cardon, 2006), and have helped enable genome-wide association studies (GWAS) (Units 1.17, 1.19, 1.20, 1.25). Data generated from GWAS are typically analysed by examining the impact of each individual genotyped variant on the outcome of interest, resulting in new associations between genetic variants and phentoypes (Welter et al., 2014). More recently, genome-wide genetic data has also provided new opportunities to estimate genetic relatedness using *mixed-model analysis*.

The principal behind mixed-model analyses is the same as twin and family-based heritability analyses — heritability is estimated via correlation between genetic sharing and phenotypic sharing. The key difference is that rather than using the theoretical estimates of genetic sharing (ultimately based on Mendel’s laws), for mixed-model analysis, empirical estimates of genetic sharing are used and are directly observed from genotype data. This underlying concept is illustrated in Figure 1. Estimates of genetic sharing, often represented as a *genetic relationship matrix* (GRM), can then be used in a variety of statistical analyses, most notably for the estimation of “chip” heritability and to adjust single-variant statistical analyses for sharing (due to various forms of sample stratification). These analyses are typically based on data from very large scale genome-wide single nucleotide polymorphism (SNP) genotyping arrays (Unit 2.9)

In this unit, we describe the mathematical principles and computational tools used to estimate genetic relationship matrices, and their application for adjustment of traditional statistical analyses of GWAS data and heritability estimation. Finally, we discuss the advantages and disadvantages of these types of analyses.

## ESTIMATING GENETIC RELATIONSHIP MATRICES

Calculating a genetic relationship matrix (GRM) yields a value for each pair of individuals in a dataset; this can be a computationally intensive process for datasets with large sample sizes. Generally, each pair of individuals is scored using a similarity function that compares the two genomes. In practice, the similarity function is typically based on the additive

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

sharing of alleles across all (N) genotyped single nucleotide polymorphisms (SNPs), according to Equation 2.

$$A_{jk}=\frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)} \quad (2)$$

The relative importance of each shared allele is probabilistically weighted by its overall frequency. Importantly, this model of relatedness does assume an additive effect per SNP. $A_{jk}$ is the assigned genetic relationship value for individuals $j$ and $k$; $N$ is the number of non-missing SNPs shared between two individuals; $x_{ij}$ is the number of copies of the reference allele for the i$^{th}$ SNP of the j$^{th}$ individual; $p_i$ is the frequency of the reference allele. The effects from each SNP are then summed ($\Sigma$) and weighted equally ($1/N$). The equal weighting assumes SNPs are independent, which may or may not be true depending on the design of the genome-wide genotype SNP array being used. Modern genotyping array designs typically have a wide, representative coverage of the genome. The resulting matrix of pair-wise relationships computed by Equation 2 is a variance-covariance matrix.

In addition to estimating genomic sharing using an additive function, some studies have also calcuated genetic relatedness that considers sharing under a dominant model (Hill et al., 2008). GRMs can also be specified theoretically using pedigree information within families (Hayes et al., 2009; Legarra et al., 2009).

Many methods are available to estimate genetic relationships within datasets (Table 1); Eu-Ahsunthornwattana et al. (Eu-Ahsunthornwattana et al., 2014) compare many of these methods and conclude that all of the linear mixed-model (LMM) methods perform similarly in most cases and that, ultimately, the method used should be selected based on factors such as usability and computation speed.

## MIXED-MODEL ANALYSES

Analyses that use GRMs must be fit using mixed models — some form of a generalized linear model that contains both *fixed effects* and *random effects*. There are multiple definitions in statistical literature for the distinction between fixed and random effects (Green and Tukey, 1960; Kreft et al., 1998; LaMotte, 1983; Robinson, 1991; Searle et al., 1992; Snijders and Bosker, 1999); in this context, regardless of terminology, the motivation is to partition the variance of a trait across a collection of directly observed values (fixed effects, such as clinical covariates or study-specific factors) and the degree of genomic sharing (a random effect based on the GRM). These models are typically fit using restricted maximum likelihood estimation (REML). With traditional maximum likelihood estimation procedures, the fixed effects in the model are ignored, resulting in a negative bias (Duchateau et al., 1998). REML is an iterative method that finds the best fit for a mixed linear model (Equation 3), where $Y$ is the phenotype, $X$ is any fixed variable, $\beta$ is the fixed variable effect size, $Z$ is the GRM, $\gamma$ is the vector of random effects from the GRM, and $\varepsilon$ is the residual random effect (representing environmental, non-genetic effects).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$Y = X\beta + Z\gamma + \varepsilon \quad (3)$$

Two primary analyses have emerged from the application of Equation 3 (and similar derivations) depending on whether a fixed effect or a random effect is the focus of the analysis. For some analyses, the information captured by the GRM is included for the adjustment of a fixed effect; this is generally done to correct traditional single-SNP analyses for sample stratification (Unit 1.22). Other analyses focus on the random effect based on the GRM, attempting to quantify the proportion of trait variance explained ("chip" heritability) by cumulative sharing of genomic loci. These two types of analyses are explained in more detail below.

## Adjusting single-SNP analyses for sample stratification using GRMs

Analysis of single SNPs for a trait of interest are often affected by various forms of confounding due to stratification within the sample. The most common of these is population stratification (Unit 1.22), where differences in the genetic ancestry of cases and controls can lead to systematic differences in SNP allele frequency, and subsequently, spurious statistical associations (Liu et al., 2013). Similar forms of stratification can be caused by differences in experimental handling and genotype calling, or batch effects, which artificially distort allele frequency differences across samples (Miclaus et al., 2010) (Unit 1.19). Cryptic relatedness can also occur within a sample, where individuals participating within a study are distantly related due to the geographic proximity of sample ascertainment (Malinowski et al., 2015).

The standard approach of correcting for these types of stratification is to use principal components analysis (PCA) of the genotype data to generate eigenvectors that reflect the strongest axes or components of variation within the dataset. These eigenvectors typically capture population-based differences among the samples, and can be used as covariates in statistical models to adjust for stratification (Price et al., 2006) (Unit 1.19).

Similar to PCA, inclusion of a random variable that models genomic sharing (based on a GRM) can effectively adjust for systematic similiarities among samples. Correction for stratification using this mixed-model approach can be more statistically powerful than PCA (Yu et al., 2006). Furthermore, mixed-model analysis eliminates the need to pre-select the number of principal components to be included as fixed effects in a regression model. Published studies include anywhere from three to twenty PCs for adjustment, and there are no accepted standards for selecting the number for inclusion, although three is the mostly commonly used number of included PCs.

Of the tools listed in Table 1, EMMAX, FaST-LMM, GEMMA, and GRAMMAR-Gamma were specifically designed for GRM adjustment of single SNP associations. GCTA recently added options for performing mixed-linear model association (or MLMA) of single SNPs (Yang et al., 2014).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

## ESTIMATING TRAIT VARIANCE EXPLAINED USING GRMS

As mentioned previously, GRMs use genotyped SNPs from GWAS datasets to estimate genomic sharing among individuals in a study sample. Using these estimates of genomic sharing, investigators can statistically model the proportion of trait variance explained by this sharing. If the information captured by the GWAS dataset represented 100% of all genetic variation, this analysis would yeild a perfectly accurate estimate of trait heritability. Because genotyping technologies do not capture all genetic variation, the estimates of genomic sharing are limited to genetic variants directly genotyped (or variants in strong linkage disequilibirum). Therefore, when properly adjusted for confounding factors, the variance explained by genetic sharing of GWAS-genotyped SNPs — often referred to as chip heritability or pseudo-heritability — can be considered a surrogate for narrow sense heritability (or heritability due to additive genetic effects). This assumes, however, that the genetic variants captured by the GWAS SNPs tag rare and structural variation with enough accuracy to properly estimate their effects along with those of common SNPs.

Analyses that focus on quantifying the trait variance explained by GRMs can be conducted in multiple implementations of mixed-model regression. The most commonly used implementation is a stand-alone tool called Genome-wide Complex Trait Analysis (GCTA) (Yang et al., 2011a) (http://cnsgenomics.com/software/gcta). The online GCTA forum can be used as a valuable resource when errors are encountered or questions arise (http://gcta.freeforums.net). GCTA is a command-line tool with options similar to the commonly used PLINK software (Purcell et al., 2007).

GCTA will accept input genotypes formatted as PLINK (Chang et al., 2015) binary files (bfile) and MACH (Li et al., 2010) genotype imputation software output files (dosage-mach and dosage-mach-gz for uncompressed and compressed MACH files, respectively). Genotype data can then be filtered in various ways, such as filtering in/out individuals (keep/remove), filtering by specific chromosome or all autosomes (chr/autosome), filtering in/out SNPs (extract/exclude), filtering by minor allele frequency (maf/max-maf), and filtering for imputed data based on imputation quality (imput-rsq).

Following genotype filtering, GRMs can be calculated and output in binary compressed format (make-grm). Importantly, GRMs are typically created using autosomal SNPs only, as the GRM calcuation expects diploid chromosomes in its model. However GCTA provides an option that uses a dosage-modified calculation (for male-male, male-female, and female-female pairs) (Yang et al., 2011a) to make a separate GRM for SNPs on the X-chromosome (make-grm-xchr). Y-chromosome and mitochondrial SNPs should be removed prior to GRM creation. Inbreeding coefficients can also be estimated by GCTA using multiple methods (Yang et al., 2011b).

Phenotypes must be specified in a file separate from genotypes (using the pheno option). Both continuous and dichotomous outcomes may be used in GCTA. Dichotomous traits are transformed by a liability threshold model (Dempster and Lerner, 1950; Falconer, 1967); liability threshold models are dependent on the population prevalence of the dichotomous trait (specified with the prevalence option). Disease prevalence is best estimated from prior

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

published studies of large cross-sectional epidemiological cohorts, such as NHANES estimates of obesity (Flegal et al., 2012). Adjustments to the liability threshold model that control for case-control ascertainment may improve statistical power, but currently is not recommended for datasets with high relatedness, or for traits that require fixed-effect adjustments (Hayeck et al., 2015).

As noted earlier, a genetic relatedness estimated from genome-wide data may capture multiple types of artifactual genetic similarity beyond a shared genetic component of disease. Specifically, the use of imputed genotypes may introduce stratification due to differences in the underlying genotyping platforms and differential genotyping missingness across different regions of the genome that may not impute well (Verma et al., 2014). Other types of genotype batch effects, and more commonly population stratification, can inflate estimates of genetic relatedness. When the goal of mixed-model analysis is to quantify the contribution of this shared genetic component of disease, these extraneous factors must be adjusted out. GCTA performs PCA in the same way as EIGENSTRAT (Price et al., 2006), the standard tool for calculating principal components. The output is an *.eigenvec file, which includes principal component (PCs) values that can be included as covariates in any analyses. Adjusting for PCs can correct for batch and population stratification effects.

GCTA can be sensitive to linkage disequilibrium (LD) and heritability can be underestimated or overestimated in influential regions with high or low LD. Lee et al. (Lee et al., 2011) and Purcell et al. (Purcell et al., 2009) suggest that LD has a relatively minimal effect and propose a minor allele frequency (MAF) stratification approach. Speed et al. (Speed et al., 2012) shows, through simulations, that GCTA-type analyses are robust as long as LD is similar for causal and non-causal regions. In regions where there is high LD near causal variants, heritability is overestimated — the opposite is true in areas of low LD. A modified method, linkage disequilibrium adjusted kinships (LDAK; www.ldak.org), can be used as an alternative method of generating a GRM, which generates a modified kinship matrix by weighting SNPs based on local LD patterns (Speed et al., 2012). Alternatively, PLINK's built-in LD pruning option can be used to filter SNPs based on LD, keeping only representative SNPs using a given LD threshold (Purcell et al., 2007), lessening the potential for confounding due to LD.

Once the fixed effects and random effects (GRM) are specified, the model is fit using REML to generate an estimate of variance explained by the GRM adjusted for all fixed effects. A likelihood ratio test (LRT) is performed by default, examining the significance of the random effect for the GRM on the fit of the model, yielding a p-value. Example output from the GCTA REML process is shown in Table 2. In Table 2, V(G) represents the variation from additive genetic effects, V(e) represents variatoin from residual effects, Vp represents the total phenotypic variation, and V(G)/Vp represents the proportion of genotypic to phenotypic variation and is interpreted as the proportion of variation explained (PVE) for quantatative traits or proportion of risk explained (PRE) for case-control datasets.

While the general application of mixed-model analyses uses a GRM generated from all SNPs in a genome-wide dataset, GRMs can also be created from SNP subsets. For example, Yang et al. partitioned genetic variance for human height over the 22 autosomes, generating

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

a separate GRM for each (Yang et al., 2011a). Subsequent studies have generated GRMs to partition trait variance by allele frequency and functional annotation (Davis et al., 2013; Hall et al., 2015). For these analyses, a large model with multiple random effects (one for each GRM) is fit to generate estimates of trait variance explained for each. Likelihood ratio tests can similarly be performed dropping each GRM from the model to estimate its effect on the trait. Example findings from studies that have applied GCTA to estimate trait variance explained are shown in Table 3.

Genetic studies often focus on finding missing heritability; another useful partitioning approach is to specify one GRM to capture known associated variants (including variants in LD) and a second GRM to capture remaining genetic effects not associated with known associated variants, therefore estimating the amount of heritability that may exist but is not explained by known associated variants (Hall et al., 2015). In addition to an estimate of trait variance explained, the fitted model also produces best linear unbiased predictions (BLUPs) for individual SNPs used to generate the GRM. These predictions are different from a typical linear regression – they are conditioned on the effect of all other genotyped SNPs, simultaneously. These can be produced in GCTA by first estimating the cumulative effect of each SNP per individual (reml-pred-rand option), then calculating BLUP solutions for each SNP (blup-snp option), taking the individual cumulative BLUP prediction as input. These estimated SNP effects are analogous to beta coefficents for the additive effect of each SNP, and may be used to generate or validate the fitted mixed-model in an independent dataset.

## DISCUSSION OF STRATEGIES

Overall, multiple software implementations of mixed-model analyses provide powerful tools that are relatively easy to use. Genome-wide association studies (GWAS) can have false-positive results due to geographic population structure, family relatedness, or cryptic relatedness (Yang et al., 2014) (Unit 1.19). Traditional approaches for correcting for these effects require external quality control analyses; principal components analysis is used for adjustments for populations stratification and batch effects (Price et al., 2006), and visualization of relationships (Abecasis et al., 2001) or calulations of relatedness/inbreeding values to estimate cryptic relatedness (Stevens et al., 2012). Mixed-model analysis of single SNPs avoids these confounders by utilizing the genetic structure within a dataset and performing a single model test. In addition, mixed model analyses of single SNPs condition on non-candidate loci, increasing power for datasets regardless of whether or not population structure is present (Yang et al., 2014).

Mixed-model estimation of heritability has many advantages over more traditional approaches. Most critically, mixed models allow the use of population-based datasets for heritability estimation, rather than the larger cost and intensive efforts needed for twin or family-based studies. Even within family-based study designs, mixed models allow an empirical estimate of genomic sharing rather than relying on theoretical distributions; this approach reduces variability and provides more precise estimates.

GRMs empirically estimate genomic sharing from observed data; however, while the intention of calculating GRMs is to capture (and model) sharing that is related to a disease

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

phenotype, other factors can contribute to observed genomic sharing within a sample. Systematic differences in genotyping quality or imputation quality will result in differential genetic similarity among samples. Differences in genetic ancestry will also result in differential sharing by ancestry (i.e. population stratification).

Currently a key drawback of mixed model analysis is the focus on sharing of common variation. While GRMs can be computed over both common and rarely occuring variants, the extent to which the occurance of low frequency variants influence analysis is currently unknown. The genetic models that drive the effects of rare variants may be different, and thus an estimate of the narrow sense heritability (additive only effects) may not sufficiently capture the effects of these variants.

Finally, genetic risk scores are an alternative to the mixed-model approach. Multiple studies have been published that use an *en masse* approach, attempting to generate a genome-wide estimate of cumulative risk or effect by summing (by sample) the effects of alleles estimated from prior studies (Bush et al., 2010; Purcell et al., 2009). Risk scores have also been applied to assess the cumulative effect of GWAS-associated SNPs (Fritsche et al., 2013). The benefit of this approach relative to mixed model analysis is an easier interpretation (as SNPs have been previously associated) and the use of a second, independent dataset to quantify the impact of fitted effects.

## SUMMARY

This unit reviews the current state of mixed-model analysis for genome-wide genetic data. While mixed-models can be more computationally expensive, they do offer a gain in statistical power over traditional methods of stratification adjustment, and are currently one of the best approaches for empircially estimating heritability from GWAS data. As study sample sizes grow and computational refinements improve the perfomance of GRM estimation, the use of mixed models for both stratification adjustment and heritability estimation are likely to grow.

## Acknowledgments

This work was supported in part by the Consortium for Alzheimer's Sequence Analysis (CASA), funded by the National Institute on Aging (1UF01) AG07133, and an ocular genomics training grant, T32EY021453, from the National Institutes of Health).

## LITERATURE CITED

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. GRR: graphical representation of relationship errors. Bioinformatics. 2001; 17:742–743. [PubMed: 11524377]

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]

Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. Nat. Genet. 2006; 38:659–662. [PubMed: 16715099]

Boomsma D, Busjahn A, Peltonen L. Classical twin studies and beyond. Nat. Rev. Genet. 2002; 3:872–882. [PubMed: 12415317]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Bush WS, Sawcer SJ, de Jager PL, Oksenberg JR, McCauley JL, Pericak-Vance MA, Haines JL. Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come. Am. J. Hum. Genet. 2010; 86:621–625. [PubMed: 20362272]

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4:7. [PubMed: 25722852]

Davis LK, Yu D, Keenan CL, Gamazon ER, Konkashbaev AI, Derks EM, Neale BM, Yang J, Lee SH, Evans P, et al. Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. PLoS Genet. 2013; 9:e1003864. [PubMed: 24204291]

Dempster ER, Lerner IM. Heritability of Threshold Characters. Genetics. 1950; 35:212–236. [PubMed: 17247344]

Duchateau L, Janssen P, Rowlands J. Linear mixed models. An introduction with applications in veterinary research (ILRI (aka ILCA and ILRAD)). 1998

Eu-Ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SMB, Blackwell JM, Cordell HJ. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. PLoS Genet. 2014; 10:e1004445. [PubMed: 25033443]

Falconer DS. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. Ann. Hum. Genet. 1967; 31:1–20. [PubMed: 6056557]

Falconer DS, Mackay TF, Frankham R. Introduction to quantitative genetics (4th edition). Trends Genet. 1996; 12.7:280.

Flegal KM, Carroll MD, Kit BK, Ogden CL. Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999–2010. JAMA. 2012; 307:491–497. [PubMed: 22253363]

Fritsche LG, Chen W, Schu M, Yaspan BL, Yu Y, Thorleifsson G, Zack DJ, Arakawa S, Cipriani V, Ripke S, et al. Seven new loci associated with age-related macular degeneration. Nat. Genet. 2013; 45:433–439. 439e1–439e2. [PubMed: 23455636]

Green BF, Tukey JW. Complex analyses of variance: General problems. Psychometrika. 1960; 25:127–152.

Hall JB, Cooke Bailey JN, Hoffman JD, Pericak-Vance MA, Scott WK, Kovach JL, Schwartz SG, Agarwal A, Brantley MA, Haines JL, et al. Estimating cumulative pathway effects on risk for age-related macular degeneration using mixed linear models. BMC Bioinformatics. 2015; 16:329. [PubMed: 26467978]

Hayeck TJ, Zaitlen NA, Loh P-R, Vilhjalmsson B, Pollack S, Gusev A, Yang J, Chen G-B, Goddard ME, Visscher PM, et al. Mixed model with correction for case-control ascertainment increases association power. Am. J. Hum. Genet. 2015; 96:720–730. [PubMed: 25892111]

Hayes BJ, Visscher PM, Goddard ME. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. (Camb). 2009; 91:47–60. [PubMed: 19220931]

Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 2008; 4:e1000008. [PubMed: 18454194]

Kreft IGG, Kreft I, Leeuw Jde. Introducing Multilevel Modeling (SAGE Publications). 1998

LaMotte LR. Fixed-, Random-, and Mixed-Effects Models (Wiley StatsRef: Statistics Reference Online). 1983

Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. 2011; 88:294–305. [PubMed: 21376301]

Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 2009; 92:4656–4663. [PubMed: 19700729]

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 2010; 34:816–834. [PubMed: 21058334]

Liu Y, Nyunoya T, Leng S, Belinsky SA, Tesfaigzi Y, Bruse S. Softwares and methods for estimating genetic ancestry in human populations. Hum. Genomics. 2013; 7:1. [PubMed: 23289408]

Malinowski J, Goodloe R, Brown-Gentry K, Crawford DC. Cryptic relatedness in epidemiologic collections accessed for genetic association studies: experiences from the Epidemiologic Architecture for Genes Linked to Environment (EAGLE) study and the National Health and Nutrition Examination Surveys (NHANES). Front. Genet. 2015; 6:317. [PubMed: 26579192]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Miclaus K, Wolfinger R, Vega S, Chierici M, Furlanello C, Lambert C, Hong H, Zhang L, Yin S, Goodsaid F. Batch effects in the BRLMM genotype calling algorithm influence GWAS results for the Affymetrix 500K array. Pharmacogenomics J. 2010; 10:336–346. [PubMed: 20676071]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 2006; 38:904–909. [PubMed: 16862161]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007; 81:559–575. [PubMed: 17701901]

Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460:748–752. [PubMed: 19571811]

Rahmioğlu N, Ahmadi KR. Classical twin design in modern pharmacogenomics studies. Pharmacogenomics. 2010; 11:215–226. [PubMed: 20136360]

Robinson GK. That BLUP Is a Good Thing: The Estimation of Random Effects. Stat. Sci. 1991; 6:15–32.

Searle SR, Casella G, McCulloch CE. Variance Components. 1992

Snijders TAB, Bosker RJ. Introduction to multilevel analysis (London: Sage). 1999

Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. 2012; 91:1011–1021. [PubMed: 23217325]

Stevens EL, Baugher JD, Shirley MD, Frelin LP, Pevsner J. Unexpected relationships and inbreeding in HapMap phase III populations. PLoS One. 2012; 7:e49575. [PubMed: 23185369]

Verma SS, de Andrade M, Tromp G, Kuivaniemi H, Pugh E, Namjou-Khales B, Mukherjee S, Jarvik GP, Kottyan LC, Burt A, et al. Imputation and quality control steps for combining multiple genome-wide datasets. Front. Genet. 2014; 5:370. [PubMed: 25566314]

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42:D1001–D1006. [PubMed: 24316577]

Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 2011a; 88:76–82. [PubMed: 21167468]

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. 2011b; 43:519–525. [PubMed: 21552263]

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. 2014; 46:100–106. [PubMed: 24473328]

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 2006; 38:203–208. [PubMed: 16380716]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

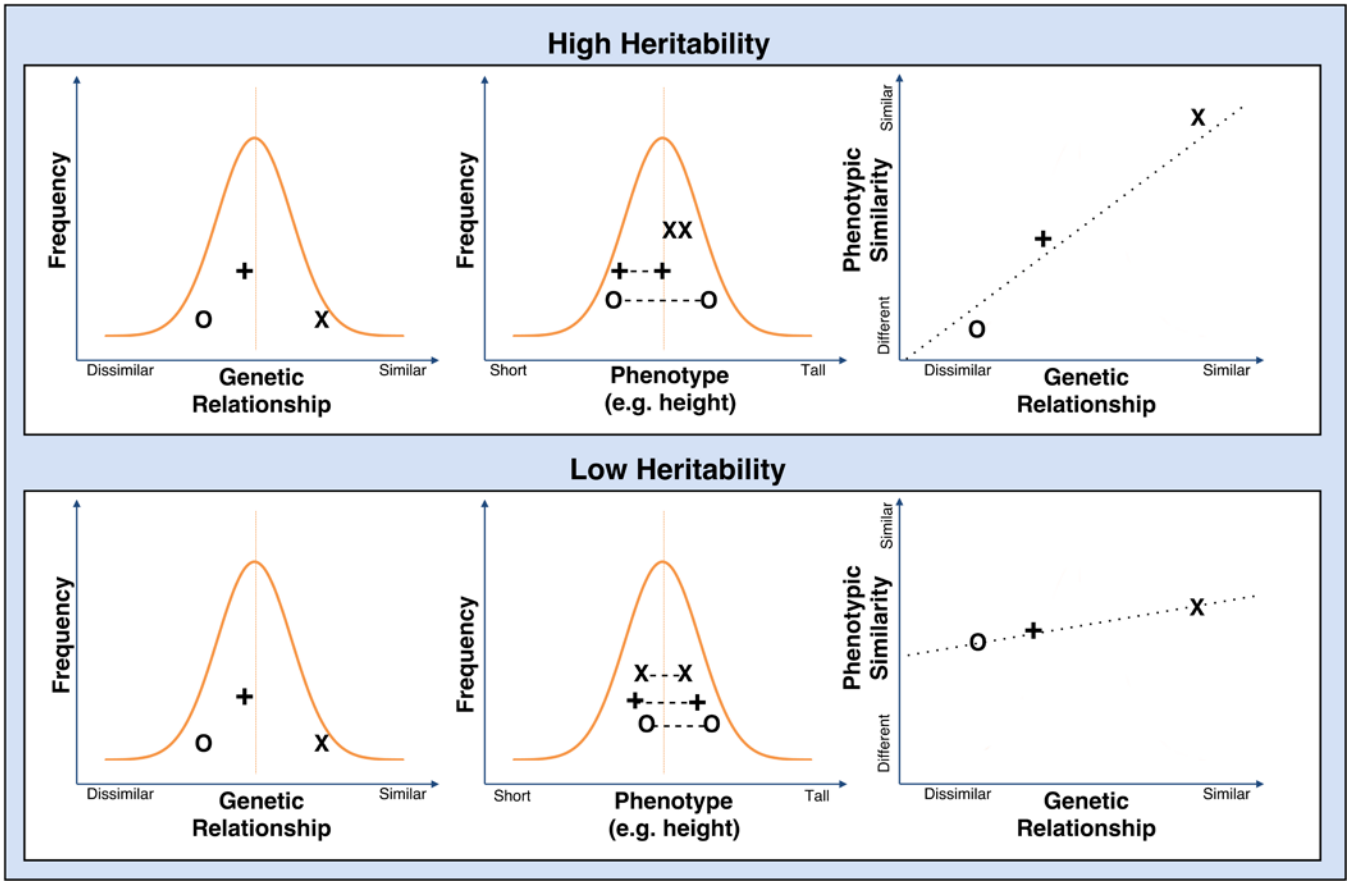


**Figure 1. Conceptual overview of mixed model analysis**

In the left-most plot, three pair-wise relationships are plotted for three pairs of samples (o, x, and +). In the middle plot, the difference in phenotypes between the two individuals of the pair are shown. In the right-most plot, the relationship between genetic similarity and phenotypic similarity is shown. A stronger correlation between genetic and phenotypic similarity indicates higher heritability.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 1**

Examples of methods to account for relatedness using linear mixed models (LMM).

| Method | PubMed ID |
|---|---|
| GCTA | 21167468 |
| REACTA | 24403537 |
| EMMAX | 20208533 |
| FaST-LMM | 21892150 |
| GEMMA | 22706312 |
| GRAMMAR-Gamma | 22983301 |

Table adapted from (Eu-Ahsunthornwattana et al., 2014).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 2**

Example output from a GCTA REML anlaysis.

| Source | Variance | Std. Err. | Component |
|---|---|---|---|
| V(G) | 0.08355 | 0.00219 | Genetic |
| V(e) | 0.12969 | 0.00188 | Residual |
| Vp | 0.21324 | 0.00165 | Phenotypic |
| V(G)/Vp | 0.39183 | 0.00897 | Proportional |

| | | |
|---|---|---|
| logL | 10919.76 | # Log Likelihood of Fitted Model |
| logL0 | 8923.723 | # Log Likelihood of Null Model |
| LRT | 3992.078 | # Likelihood Ratio Test Chi-Square |
| Pval | < 0.001 | # P-value from LRT |

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**Table 3**

Example findings from studies using GCTA.

| Trait Studied | PVE* | PubMed ID |
|---|---|---|
| Childhood Adiposity | 30% | 23528754 |
| Drug Dependence | 36% | 25424661 |
| Height | 45% | 20562875 |
| Intelligence (From age 11) | 62% | 22258510 |
| Multiple Myeloma | 15.2% | 26208354 |
| Psoriasis (in Han Chinese) | 45.7% | 26172869 |
| Pulmonary Function | 41.6–71.2% | 25745850 |
| Schizophrenia | 39% | 26198764 |

* Proportion of Variance Explained