PLINK: a toolset for whole genome association and population-based linkage analyses

Shaun Purcell[1,2], Benjamin Neale[1,2,3],
Kathe Todd-Brown[1], Lori Thomas[1],
Manuel A R Ferreira[1], David Bender[1,2],
Julian Maller[1,2], Paul I W de Bakker[1,2],
Mark J Daly[1,2], Pak C Sham[4]

[1] Center for Human Genetic Research, MGH, Boston, USA.
[2] Broad Institute of Harvard and MIT, Cambridge, USA.
[3] Institute of Psychiatry, University of London, London, UK.
[4] Genome Research Center, University of Hong Kong, Pokfulam, Hong Kong.

Correspondence: Shaun Purcell, Rm 6.254, CPZ-N, 185 Cambridge Street, Boston, MA, 02114, USA; Tel: 617-726-7642; Fax: 617-726-0830; shaun@pngu.mgh.harvard.edu
Keywords : Whole genome association studies, single nucleotide polymorphisms, identity-by-state, identity-by-descent, linkage analysis, computer software

Abstract

Whole-genome association studies (WGAS) bring new computational as well as analytic challenges to researchers. Many existing genetic analysis tools are not designed to handle such large datasets in a convenient manner and do not necessarily exploit the new opportunities that whole-genome data bring. To address these issues, we developed PLINK, an open source C/C++ WGAS toolset. Large datasets, comprising hundreds of thousands of markers genotyped for thousands of individuals, can be rapidly manipulated and analyzed in their entirety. As well as providing tools to make the basic analytic steps computationally efficient, PLINK also supports some novel approaches to whole-genome data, that take advantage of whole-genome coverage. We introduce PLINK and describe the five main domains of function: data management, summary statistics, population stratification, association analysis and identity-by-descent estimation. In particular, we focus on the estimation and use of identity-by-state and identity-by-descent information in the context of population-based whole-genome studies. This information can be used to detect and correct for population stratification and to identify extended chromosomal segments that are shared identical-by-descent between very distantly related individuals. Analysis of the patterns of segmental sharing has the potential to map disease loci that contain multiple, rare variants in a "population-based linkage analysis". PLINK is freely available from the author's website, `http://pngu.mgh.harvard.edu/purcell/plink/`

Despite a substantial body of research that spans decades, we have largely failed to elucidate the molecular genetic basis of most common, complex human traits and diseases. The genetic epidemiology of these outcomes has often convincingly demonstrated only two facts: that genetic factors play an important role, and that the genetic variation is not due to a single, Mendelian mutation. With this implication of polygenic effects in mind (many genes of small effect), researchers have increasingly been aware of the need to design larger linkage and association studies that have adequate power[1,2]. However, the strategies of the past decade have met with limited success[3,4]. One possible reason for the lack of identified complex trait disease genes is that studies have still been lacking in sample size and genome coverage.

Modern whole genome association studies (WGAS) represent a direct attempt to address these problems. Based on advances arising from large-scale genomic projects including the human genome sequence, single nucleotide polymorphism (SNP) discovery efforts and the HapMap, as well as new genotyping technology, it is only in the past one or two years that our understanding of variation and technical ability to assess it has enabled association to move from candidate gene to unbiased whole genome searches. The standard logic of the WGAS design implicitly assumes that common variants with modest effects on disease frequently exist and explain substantial proportions of variation (i.e. the common disease/common variant (CD/CV) hypothesis[5], but that previous studies either have not looked at them at all (not enough genetic markers tested) or have been underpowered to find significant associations (not enough samples). Eventually, WGAS should provide a

powerful and comprehensive test of the CD/CV hypothesis for any given disease. In this report we introduce a new analytic tool for WGAS and discuss some crucial analytic design considerations such as multiple testing, bias due to confounding, and the possibility that rare genetic variation underlies common disease.

New SNP genotyping technologies have enabled the next generation of genetic studies, with many WGAS either planned, underway or already completed. A typical WGAS, currently with hundreds of thousands of SNPs genotyped for thousands of individuals, represents a dataset that is several orders-of-magnitude larger than previous linkage and association studies. As such, WGAS present new computational and statistical challenges. Perhaps the most apparent challenge is from the increased multiple testing burden: the concern that, from a set of hundreds of thousands of tests, many highly significant results are expected by chance alone, making it hard to distinguish signal from noise. To a large extent, this problem can be assuaged by moderate increases in sample size: basic power calculations show that to maintain the same power when performing an exponentially-larger number of Bonferroni-corrected tests requires only a linear increase in sample size. For example, if five hundred individuals are needed to test a single SNP with adequate power, something in the order of two thousand will be required to test five hundred thousand SNPs even after Bonferroni correction. So although increased sample sizes are certainly required, even with the most conservative statistical approaches these will often be achievable, rather than order-of-magnitude, increases.

The size of these datasets will present a computational, as well as a statistical, testing burden, as many existing genetic analysis software programs were not designed with WGAS in mind. We have therefore developed a user-friendly software tool, PLINK, to facilitate the analysis of whole-genome data in a number of ways: by addressing the mundane but important need for easy ways to manage such data, by making routine analyses computationally efficient, and by offering new analyses that take advantage of whole-genome coverage. Considering a relatively small WGAS dataset of 100,000 SNPs genotyped on 350 individuals, for example, PLINK takes approximately 10 seconds to load, filter and perform association analysis for all SNPs; much larger datasets can also be straightforwardly handled.

Aside from computational challenges, larger datasets also exacerbate the problem of confounding in genetic association studies. With increasing power to detect true effects, comes increased potential for bias to affect results (i.e. the "power to detect" departure from the null hypothesis due to unaccounted confounders will also increase). One well-acknowledged source of confounding in population-based association studies is population stratification[6]. However, in the context of WGAS, this perhaps presents less of a problem, given the availability of hundreds of thousands of markers across the genome, allowing for a very accurate empirical assessment of stratification, via genomic control[7] and structured association methods[8,9,10]. Augmenting this set of approaches, we describe below our approach to stratification, implemented within PLINK and designed to work with whole-genome data.

Another, arguably more insidious, source of confounding in WGAS is from non-random genotyping failure[11,12]. This involves an individual's SNP genotype that is either incorrectly called or (more commonly) not called at all. If this is non-random with respect to genotype (e.g. some genotypes more likely to be uncalled) and to phenotype (e.g. cases have lower genotyping rates than controls on average) then false positive associations can occur. That certain genotypes for a given SNP are more likely to fail than others is almost

certainly the rule rather than the exception for any genotyping technology. Furthermore, it is probably also the exception rather than the rule that cases and controls are collected at exactly the same time and place and are handled similarly throughout the laboratory process; indeed, control data may come from a completely different study having been genotyped in a different laboratory and called with a different algorithm.

Even though genotyping rates might be very high in general, in large samples even a small proportion of non-random genotyping failure could induce a false positive (especially if it occurs for one of the many SNPs already showing a high level of association by chance). As normal screening procedures, based on measures such as overall genotyping rate and Hardy-Weinberg equilibrium, will often not detect these biased SNPs, it is important to look closely at patterns of genotyping failure for non-random effects (as well as visually inspecting the raw data, prior to calling genotypes). In PLINK, genotyping failure can be examined both with respect to phenotype and (potentially unobserved) genotype.

For some complex traits and diseases, an alternate hypothesis for the lack of identified genes is that common variants do not explain a substantial proportion of the phenotypic variation. Under this model, the considerable levels of heritability could reflect aggregates of very many, very rare variants (each potentially of moderate effect but accounting for virtually none of the variation at the population level) which we refer to as the multiple rare variant (MRV) hypothesis[13]. Standard association approaches will likely fail when the MRV hypothesis holds (power will be low even for high genotypic relative risks, e.g. > 5, if the frequency is very low, e.g. $\sim 1/10,000$, even before multiple testing corrections). Importantly though, the same data collected for WGAS studies (in particular, panels of common SNPs genotyped in population-based samples) can potentially be analyzed using different approaches that do not assume that common variation underlies disease. In particular, if multiple, rare disease variants exist *within the same gene or genomic region*, then instead of standard association one might consider an approach more akin to linkage analysis, but performed in population-based samples of unrelated individuals. Rather than directly test frequency differences of a variant, we propose examining ancestral sharing at a locus. That is, given ascertainment on disease, we might expect to see multiple copies of even very rare variants, that are moderately or highly penetrant, among the descendants of the founder in whom the mutational events occurred.

In standard association analysis undocumented relatedness can be another source of bias, although with whole-genome data and analytic tools such as those described below, one can unambiguously detect closely related individuals. However, more distant relatedness between individuals who share the same disease may convey additional information for gene-mapping. Analyses of the type we propose here might be able to leverage this information to provide a complementary approach to standard association analysis, using the same data already being collected for single SNP association studies.

If two individuals share the same rare variant, we would also expect that they share not just that variant but also the surrounding chromosomal region, particularly as rarer variants are more likely to be relatively recent. We propose to use panels of common SNPs to look for these regions of extended sharing (regions that are inherited identical by descent between seemingly unrelated individuals). If a particular region harbors multiple rare variants, we would expect to see inflated levels of segmental sharing between case/case pairs at that locus, compared to case/control and control/control pairs. A procedure for detecting

shared segments and testing for correlation between sharing and phenotypic similarity forms the basis of a "population-based linkage analysis", that is intended as a complementary approach to standard association analysis. This differs from standard haplotype analysis in that we do not try to infer phase explicitly or estimate haplotype frequencies so we can assess sharing of very rare, very long regions accurately; in addition, the subsequent unit of analysis is sharing at the locus rather than the frequency of one or more haplotypic variants.

In summary, given the issues raised above, we designed the PLINK WGAS toolset to the meet the following requirements: a) to provide a simple way to handle large WGAS datasets; b) to assess confounding due to stratification and non-random genotyping failure, and produce a range of other summary statistics; c) to perform a variety of standard association tests efficiently on very large datasets (in populations or families, for disease or quantitative outcomes, allowing for covariates, haplotypic tests, etc); d) to provide a means of assaying rare variation using common SNP panels, thereby providing a mapping method that might perform better when the MRV model holds. In the rest of this report, we highlight some of PLINK's main features, briefly describing five domains of functions (data management, summary statistics, assessment of population stratification, association analysis and identity-by-descent estimation). All these methods are applicable to whole genome datasets. Below, we either describe or provide references for the tests implemented; these methods and other new ones being added are described in more detail in the online technical documentation being added to the PLINK website.

Data Management

We have developed a compact binary file format to represent SNP data, as well as tools to transform between this and standard text-based formats (including both a one-row-per-individual and a transposed one-row-per-SNP format). A simple interface is provided for reordering, recoding and filtering genotype information (extracting individuals and/or SNPs on certain criteria, such as physical position, genotyping rate or covariate values). It is also possible to merge two or more datasets (that can partially overlap in terms of both individuals and markers) and produce reports of discrepancies between overlapping datasets.

Summary statistics

Standard summary measures are available: genotyping rates, allele and genotype frequencies, Hardy-Weinberg equilibrium tests using asymptotic and exact[14] procedures and single SNP Mendel error summaries for family data. PLINK also estimates individual heterozygosity rates and provides an automatic gender-check facility based on X chromosome heterozygosity. We employ a per-SNP test of non-random genotyping failure with respect to phenotypic status, based on a simple chi-squared test of different rates of genotyping failure in cases versus controls.

We also test whether missingness at a site can be predicted by the local haplotypic background, to spot non-random genotyping failure with respect to genotype. Taking each SNP that has an above-threshold level of genotyping failure as the reference SNP, we ask whether the haplotypes formed by the two (or more) flanking SNPs can predict which individuals are missing at the reference SNP. The test is a simple haplotypic case/control test,

where the phenotype is the presence or absence of a called genotype at the reference SNP. If missingness at the reference SNP is not random with respect to the true (unobserved) genotype, we will often expect to see an association between missingness and flanking haplotypes. This test will often have higher specificity than sensitivity: it relies on linkage disequilibrium patterns to make an inference about the potentially unobserved reference allele, so it might miss many SNPs showing high, non-random levels of genotyping failure. However, used as a screening tool, SNPs that show highly non-random patterns of missing data could obviously be problematic and should be treated with caution.

For an example, we consider an Illumina whole-genome SNP dataset, freely available at the NINDS Repository at Coriell (see the Acknowledgments section below for full details), comprising 276 amyotrophic lateral sclerosis cases and 271 controls. We illustrate the above tests for one particular SNP, rs5742981. Genotyping failure for this SNP was not randomly distributed with respect to phenotype (10.5% in cases and 0.7% in controls, $p = 8 \times 10^{-7}$) or genotype. Flanking haplotypes (formed by SNPs rs1899025 and rs5743030) are very strongly associated with genotyping status at rs5742981 (e.g. the GA haplotype is associated with missing rs5742981 genotype at $p = 9 \times 10^{-56}$). In fact, if we divide the sample into individuals who are heterozygous for the background haplotype (N=213) and those who are not (N=333), then all 33 instances of genotyping failure at rs5742981 fall in the smaller, heterozygous group ($p = 7.6 \times 10^{-13}$). It would seem that, in this particular case, heterozygosity for the haplotypic background predicts heterozygosity at the reference SNP and that heterozygotes are preferentially dropped in cases only. This would also seem to generate the association between allele count at rs5742981 with disease ($p = 0.0043$, MAF 0.4% in cases, 2.6% in controls). Of course, such problems can often, but not always, be avoided by imposing appropriate thresholds on allele frequency, genotyping rate and Hardy-Weinberg threshold. In this case, most Hardy-Weinberg filters would not have excluded this SNP (in controls $p = 1$; in the total sample $p = 0.11$), although a missing data threshold of 5% would have excluded this SNP. In any case, having additional simple QC metrics, including the two presented here, automatically and quickly calculated by PLINK, will often help to flag problematic SNPs.

Population stratification

Based on the genome-wide average proportion of alleles shared identical-by-state (IBS) between any two individuals, PLINK offers tools to *a)* cluster individuals into homogeneous subsets, *b)* to perform classical multidimensional scaling to visualize substructure and provide quantitative indices of population genetic variation, and *c)* to identify outlying individuals. PLINK uses complete-linkage hierarchical clustering to assess population stratification, using whole-genome SNP data. This agglomerative procedure starts by considering every individual as a separate cluster of size 1, then repeatedly merges the two closest clusters. Complete-linkage clustering specifies that clusters are compared on the basis of their two most dissimilar members; clustering stops either when all individuals belong to one cluster, or based on pre-specified constraints (stopping rules).

Various optional constraints can be applied, with the specific goal of subsequent association analysis in mind, rather than an accurate description of population genetic variation *per se*. That is, we aim to ensure that all members of any derived cluster belong to the same subpopulation, rather than attempting to ensure that all members of the same

subpopulation belong to the same cluster. The purpose of the constraints is to select which solution to accept from the distance-based clustering approach (i.e. with no constraints, all solutions are considered, i.e. for N individuals, from N clusters each of size 1, to 1 cluster of size N).

One constraint PLINK applies is called the "pairwise population concordance" (PPC) test, similar to a method by Lee[15], such that for any putative new cluster, all pairs of individuals pass this test. For a given pair, we expect to see autosomal SNPs with two copies of each allele occur in a 2:1 ratio of IBS 2 $\{Aa, Aa\}$ to IBS 0 $\{AA, aa\}$ SNP pairs, if both members of the pair come from the same random-mating population. For SNPs selected far enough apart to be approximately independent (e.g. 500kb), a test of binomial proportions can suggest concordant or discordant ancestry for each pair of individuals. A pair from different populations is expected to show relatively more IBS 0 SNPs; a one-sided test for departure from a 2:1 ratio is given by the normal approximation to the binomial: for a particular pair, if L is the total number of informative, independent SNP pairs and L_2 is the subset that are IBS 2,

$$Z = \frac{\frac{L_2}{L} - \frac{2}{3}}{\sqrt{\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{L}}}.$$

One can choose to merge clusters only if no between-cluster pairs have a statistically significant PPC result, at a given significance threshold. As well as the PPC test, we have incorporated other constraints in the clustering procedure. As mentioned above, non-random genotyping failure is a possible source of confounding in genetic association studies. One possible constraint is to only cluster individuals that have similar profiles of missing data, or "identity by missingness", in which we specify a threshold for the maximum permissible proportion of sites for which two individuals are discordant in genotyping status (genotyped versus missing). For case/control samples, another possible constraint is that each cluster of 2 or more individuals has at least one case and one control (and so be informative for association analysis that conditions on cluster). Alternatively, the maximum cluster size or the number of clusters can be fixed. It is also possible to combine phenotype and cluster size constraints, by specifying that a cluster contains no more than 1 case and 3 controls, for example. Finally, one can also combine multiple external categorical and quantitative matching criteria (such as age, sex, other environmental variables or QC measures such as the genotype call rate for each individual) alongside the genetic matching. Categorical criteria can be either 'positive' or 'negative' such that only similarly-categorized or differently-categorized individuals can be merged. It is also possible to select only a single individual from a particular pre-specified group. The complete algorithm is as follows: the IBS distance between individual k (belonging to cluster i) and individual l (belonging to cluster j) is denoted d_{ijkl}; the between cluster distances are denoted D_{ij}.

1) **START:** Find valid i, j for $\min_{ij}(D_{ij})$ where $D_{ij} = \max_{kl}(d_{ijkl})$
2) Test (optional) constraints for this potential new cluster:
→ New cluster contains both cases and controls ?
→ Merged $i + j$ cluster smaller than maximum cluster size constraint?
→ Maximum number of cases or controls exceeded?
3) For every pair between i and j, test following (optional) constraints:

→ Pairable based on external constraints?
→ Non-significant PPC test?
→ Pass identity-by-missingness threshold?
→ Already selected an individual from this group?
4) **Satisfies constraints?** → MERGE CLUSTERS
5) **No remaining pairable clusters?** → STOP
6) Return to START for next best pair of clusters.

PLINK also provides an alternate way to look at population stratification: rather than clustering into discrete groups, it can use the data reduction technique of classical multidimensional scaling (MDS), to produce a k-dimensional representation of any substructure. Although the primary use of this approach is for visualization, the values for each of the k dimensions can be used as covariates in subsequent association analysis to control for stratification, instead of using discrete clusters. There is an option to use a Euclidean IBS distance metric ($\sqrt{\sum_l I_l^2}$ where I_l is 0,1,2 the number of alleles shared IBS at locus l) in place of the standard metric of proportional sharing; classical MDS based on a Euclidean distance metric is numerically identical to principal components analysis, which forms the basis of other methods[10].

Finally, PLINK also supports an IBS-based "nearest neighbor" analysis to detect outlying individuals who do not belong with any major cluster in the sample. For each individual, the distance to their nearest neighbor is calculated; this distribution is standardized (by the sample mean and variance of nearest neighbor distances) and can be inspected for outliers. The same procedure can also be applied to individuals' n^{th} nearest neighbor.

Here we illustrate how these methods can differentiate between two quite similar populations and control for between-population differences in tests of association, using freely-available HapMap data (http://www.hapmap.org). From the 90 Asian individuals (45 unrelated Han Chinese from Beijing, China, labelled CHB, and 45 unrelated Japanese from Tokyo, Japan, labelled JPT) in the Phase II dataset, we extracted the set of autosomal SNPs on the Affymetrix GeneChip 500K Mapping Array. Figure 1 shows the results of a multidimensional scaling analysis, which clearly separates out two clusters (the left and right clusters correspond to CHB and JPT respectively). The color coding in the three panels shows the classification of individuals according to increasingly liberal PPC thresholds: 0.01, 0.001 and 0.0001 (left to right), which result in 7, 5 and 4 classes respectively. In all solutions, two of the classes represent two single Japanese individuals: the nearest-neighbor diagnostics identified these two individuals (NA18987 and NA18992) as clear outliers, more than 3 standard deviations below the mean.

To mimic a dataset showing stratification effects, we next created a dummy phenotype to represent genotype at rs2976396, a SNP that shows strong ($p = 2.7 \times 10^{-8}$) allele frequency differences between CHB (frequency of A allele is 0.24) versus JPT (frequency of A allele is 0.66) populations. Specifically, individuals with the GG genotype are designated cases, AA and AG individuals are designated controls. The dummy phenotype therefore shows a marked difference in prevalence between CHB and JPT populations. Performing a standard association analysis (methods described below in next section) for all 500K SNPs with this phenotype, the genomic control inflation factor is 1.163, indicative of some infla-

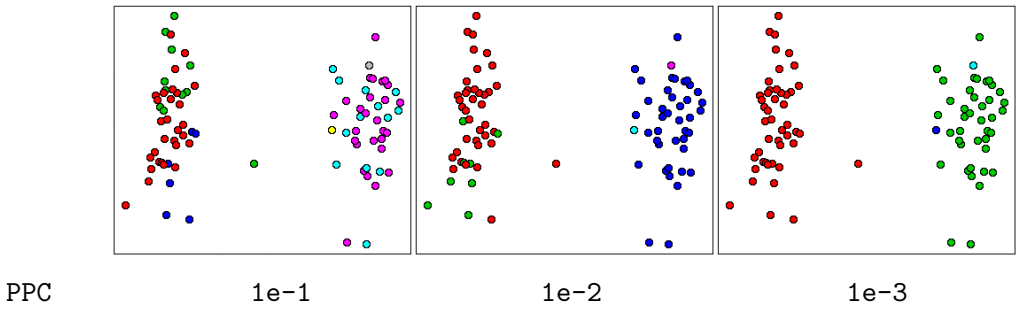

Figure 1. **Multidimensional scaling and classification of Asian HapMap individuals.** Multidimensional scaling reveals two clear clusters which correspond to Chinese and Japanese HapMap populations. The figure's three panels differ only in the color scheme, which (left to right) represents classification according to PPC thresholds of 0.01, 0.001 and 0.0001.

tion (importantly, this would be sufficient to impact the tail of the test statistic distibution, perhaps pushing what would have been marginally significant results to genome-wide significance). Performing tests of association controlling for between-strata effects reduced the genomic control inflation factor to virtually 1 for all three solutions (1.00474, 1.00315, 1.00098 for PPC 0.01, 0.001 and 0.0001 solutions, respectively; if actual population membership, CHB versus JPT, was used instead of the empirically-derived classification, the inflation factor was 1.00044).

Given the range of methods and options for detecting and correcting for population stratification (in addition to a genomic control procedure[7] implemented in PLINK), further work is clearly required to determine and quantify the typical performance of these approaches, in terms of the type I and type II error rates in subsequent association analysis. Further work is also needed to assess how best to apply these types of genetic matching procedures when controls are selected from large, pre-existing panels (i.e. as opposed to cases and controls being collected and genotyped together). Towards this goal, we note that several of the current authors are involved in work that provides direct comparisons of PLINK with other methods in real whole genome data[16,17].

Association analysis

As well as the standard case/control allelic test, PLINK offers a Cochran-Armitage trend test, Fisher's exact test, genotypic tests (general, dominant and recessive models) and Cochran-Mantel-Haenszel tests for stratified tables[18], which allow for tests of association conditional on any cluster solution or other categorization of samples. The Breslow-Day[18] and homogeneity of odds ratio[19] tests are supported, as are tests for quantitative traits using a standard linear regression of phenotype on allele dosage. The standard disease and quantitative trait association tests are implemented for speed; the same analyses are also framed as more general linear and logistic regression models which allow for multiple binary or continuous covariates having both main effects and interactions. One can test for joint effects, or perform a scan conditional on a given SNP or set of SNPs, for example; also, gene-gene and gene-environment interaction tests for quantitative and disease traits can be

performed.

For family data, the standard TDT[20] is provided. The permutation procedure applied to the TDT flips the transmitted and untransmitted alleles of all individuals in a nuclear family for all SNPs per permutation, thereby preserving in each permuted dataset the possible non-independence of transmissions across SNPs and across multiple offspring due to linkage disequilibrium and linkage. We also implement the sib-TDT[21] for nuclear families to include sibships without parents as well as unrelated individuals (called the DFAM test within PLINK). We break pedigrees into nuclear families and classify them as those where both parents are genotyped and those where they are not. For the first class of families we obtain the allele count of the minor allele (A, major allele is a) among affected children in family f, labeled S_f. Under the null hypothesis of no association the binomial distribution gives the expected value and variance of S_f, given the parental genotypes (i.e. sampling parental transmissions with replacement). If family f contains D affected offspring, then $E(S_f) = DP$ where P is $\frac{1}{2}$, 1 or $1\frac{1}{2}$ for $aa \times Aa$, $Aa \times Aa$ and $AA \times Aa$ parental mating types respectively, and $Var(S_f) = D/4$ unless both parents are heterozygous in which case $Var(S_f) = D/2$. For the second class of families, we also obtain S_f, the count of minor alleles in affected offspring and its expected value and variance (under H_0, based on the genotypes of all siblings in the family); these are given by the multivariate hypergeometric distribution (i.e. sampling genotypes without replacement). The use of the genotype-based multivariate hypergeometric distribution in sibships accounts for the fact that not all allelic combinations are possible within a sibship (e.g. an individual cannot have two paternal alleles). The number of all offspring with genotypes AA, Aa and aa in family f are labeled N_{AA}, N_{Aa} and N_{aa} respectively (which sum to N); the equivalent numbers in affected offspring are D_{AA}, D_{Aa} and D_{aa}, which sum to D; therefore $S_f = 2D_{AA} + D_{Aa}$. The expected allele count in affecteds in family f under the null is $E(S_f) = 2E(D_{AA}) + E(D_{Aa}) = (2N_{AA} + N_{Aa})(D/N)$ and the variance, obtain using the multivariate hypergeometric distribution, is

$$\begin{aligned} Var(S_f) &= 4Var(D_{AA}) + Var(D_{Aa}) + 4Cov(D_{AA}, D_{Aa}) \\ &= 4D\left(\frac{N_{AA}}{N}\right)\left(1 - \frac{N_{AA}}{N}\right)\left(\frac{N-D}{N-1}\right) \\ &\quad +D\left(\frac{N_{Aa}}{N}\right)\left(1 - \frac{N_{Aa}}{N}\right)\left(\frac{N-D}{N-1}\right) \\ &\quad -4\left(\frac{N_{AA}N_{Aa}}{N^2}\right) \end{aligned}$$

Summing over families, a test statistic is $\frac{\left(\sum S_f - \sum E(S_f)\right)^2}{\sum Var(S_f)}$ which follows a χ^2_1 distribution under the null hypothesis. Additional unrelated cases and controls, potentially stratified into clusters, can be included within this framework by treating them as sibships except using the standard hypergeometric distribution (i.e. sampling *alleles* without replacement), which is equivalent to the standard Cochran-Mantel-Haenszel test mentioned above.

For quantitative traits, PLINK provides an implementation of the between/within model[22,23], which uses a permutation procedure (permuting genotype rather than phenotype) to control for the non-independence of individuals within the same family (the QFAM test). The analysis of phenotype-genotype association is a standard regression of phenotype

on genotype, ignoring family structure. Significance is based on the following permutation procedure: genotypes are decomposed into between and within family components, following the models referenced above; these two components are then permuted independently at the level of the family and then summed to form new pseudo-genotype scores for each individual. That is, between components are swapped between families; within components have their sign swapped with 50% chance (similarly for all members of the same family). This approach provides tests that give correct type I error rates, accounting for the relatedness between individuals. Despite the necessity of permutation, one advantage is that non-normal and dichotomous phenotypes can be appropriately analyzed. While the basic test is of total association, the between and within components can also be tested separately. Information on parental phenotypes can also be combined in these analyses[24,25].

There is support for haplotype-based case/control tests and TDT based on the expected haplotype distribution for each individual obtained from E-M phasing. Either prespecified lists or sliding windows are used to specify the particular haplotype tests; precomputed lists of efficient sets of tests for common WGAS products based on the HapMap[26] are available from the PLINK website and can be immediately applied to these datasets. Also, two non-haplotypic multilocus, "gene-based" or "set-based" tests are available: using sum-statistics[27] and, for case/control samples, Hotelling's T^2.

For many tests, a number of permutation procedures are available: "adaptive" permutations which give up early on clearly non-significant results[28]; "max(T)" permutation to correct for multiple tests[29]; a rank-ordered permutation in which the n^{th} best original result is compared against the n^{th} best in each permuted dataset; gene-dropping for family-based tests; finally, the between/within permutation scheme described above. A range of multiple test corrections are also available, including those based on Bonferroni correction and false discovery rate[30].

Identity-by-descent estimation

The final domain of function concerns identity-by-descent (IBD) estimation. In homogeneous samples, PLINK provides options to estimate genome-wide IBD sharing coefficients between seemingly unrelated individuals from whole-genome data. These metrics (probabilities of sharing 0, 1 or 2 IBD) can be particularly useful for quality control, by diagnosing pedigree errors, undetected relationships and sample swap, duplication and contamination events.

PLINK has a simple procedure to find extended stretches of homozygosity in whole-genome data (regions spanning more than a certain number of SNPs and/or kilobases, allowing for a certain amount of missing genotypes and/or occasional heterozygote calls) which occur relatively frequently and can provide a powerful approach to map recessive disease genes[31,32]. Via permutation, an empirical p-value can be calculated for each SNP, based on a test for whether there is a higher rate of homozygous segments spanning that position in cases versus controls. PLINK also has options to determine distinct sets of overlapping (and, optionally, allelically-matching) segments, thereby allowing for further inspection of the data.

PLINK also calculates inbreeding coefficients for each individual. Specifically, for a particular SNP with known allele frequencies p and q, the probability that individual i is homozygous equals $f_i + (1 - f_i)(p^2 + q^2)$ or the probability of being autozygous (homozygous

by descent) (f_i) plus the probability of being homozygous by chance. If individual i has L_i genotyped autosomal SNPs, let O_i be the number of observed homozygotes and E_i be the number expected by chance; then $O_i = f_i \cdot L_i + (1 - f_i)E_i$ which gives $f_i = (O_i - E_i)/(L_i - E_i)$. When allele frequencies are not known but estimated from the sample, an unbiased estimator of E_i is based on the sum over all SNPs non-missing for individual i, $\sum_{j=1}^{L_i} 1 - 2p_j q_j \cdot T_{Aj}/(T_{Aj} - 1)$ where T_{Aj} is twice the number of non-missing genotypes for SNP j.

We have also implemented a novel method to detect extended chromosomal segmental IBD sharing between pairs of distantly related individuals using a hidden Markov model, in which the underlying hidden IBD state is estimated given the observed IBS sharing and genome-wide level of relatedness between the pair. We also provide a test for correlation between segmental chromosomal sharing and phenotypic sharing. This test, a population-based linkage analysis, potentially offers a complementary approach to whole-genome data that does not assume the common variant hypothesis of disease-related genetic variation. We describe our approach in three steps: estimation of genome-wide relatedness; estimation of local segmental sharing; relating pairwise segmental sharing to phenotypic similarity.

Estimation of genome-wide identity-by-descent sharing

We use a method of moments approach to estimate the probability of sharing 0, 1 or 2 alleles IBD for any two individuals from the same homogeneous, random-mating population. If we denote IBS states as I and IBD states as Z (in both cases the possible states being 0, 1 and 2) then we can express the prior probability of IBS sharing

$$P(I = i) \quad = \quad \sum_{z=0}^{z=i} P(I = i|Z = z)P(Z = z) \tag{1}$$

As described in detail below, for each SNP, we specify $P(I|Z)$ in terms of the allele frequency; averaging over all SNPs, we obtain the expected value for $P(I|Z)$. Then, re-arranging the three equations implied by Eq. 1, we solve for $P(Z = 0)$, $P(Z = 1)$ and $P(Z = 2)$ and calculate $\hat{\pi} = P(Z = 1)/2 + P(Z = 2)$, the proportion of alleles shared IBD.

For all SNPs, we calculate allele frequencies (based only on founders if family information is present). For any one marker, $P(I|Z)$ is a function of allele frequency (for alleles A and a, these are p and $q = 1 - p$). If p and q were known with certainty, then $2p^2q^2$ would, for example, be an unbiased estimator of $P(I = 0|Z = 0)$ (i.e. this requires that both individuals have opposite homozygotes, either $\{AA/aa\}$ with probability $p^2 \times q^2$ or $\{aa/AA\}$ with probability $q^2 \times p^2$). However, because p and q are only estimated from a finite sample, there is a bias that we take into account as follows. Let X and Y equal the counts of the two alleles in the sample for a particular SNP, and so $p = X/T_A$ and $q = Y/T_A$ where T_A is twice the number of non-missing genotypes. There are $T_A(T_A - 1)(T_A - 2)(T_A - 3)$ possible ways of selecting 4 distinct alleles from T_A alleles; of these, $X(X - 1)Y(Y - 1)$ will be $\{AA/aa\}$ genotype pairs and $Y(Y - 1)X(X - 1)$ will be $\{aa/AA\}$. Therefore, $P(I = 0|Z = 0) = \frac{2X(X-1)Y(Y-1)}{T_A(T_A-1)(T_A-2)(T_A-3)}$ which, re-expressed in terms of the original probabilities and a correction factor based on allele counts, equals $2p^2q^2 \cdot \left(\frac{X-1}{X} \cdot \frac{Y-1}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$. Following a similar logic, the full set of $P(I|Z)$ is given in Table 1.

I	Z	$P(I\|Z)$
0	0	$2p^2q^2 \cdot \left(\frac{X-1}{X} \cdot \frac{Y-1}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$
1	0	$4p^3q \cdot \left(\frac{X-1}{X} \cdot \frac{X-2}{X} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$
		$+4pq^3 \cdot \left(\frac{Y-1}{Y} \cdot \frac{Y-2}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$
2	0	$p^4 \cdot \left(\frac{X-1}{X} \cdot \frac{X-2}{X} \cdot \frac{X-3}{X} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$
		$+q^4 \cdot \left(\frac{Y-1}{Y} \cdot \frac{Y-2}{Y} \cdot \frac{Y-3}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$
		$+4p^2q^2 \cdot \left(\frac{X-1}{X} \cdot \frac{Y-1}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$
0	1	0
1	1	$2p^2q \cdot \left(\frac{X-1}{X} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2}\right) + 2pq^2 \cdot \left(\frac{Y-1}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2}\right)$
2	1	$p^3 \cdot \left(\frac{X-1}{X} \cdot \frac{X-2}{X} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2}\right) + q^3 \cdot \left(\frac{Y-1}{Y} \cdot \frac{Y-2}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2}\right)$
		$+p^2q \cdot \left(\frac{X-1}{X} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2}\right) + pq^2 \cdot \left(\frac{Y-1}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2}\right)$
0	2	0
1	2	0
2	2	1

Table 1: Calculation of $P(I|Z)$, the probability of IBS given IBD state for a given SNP, as a function of SNP allele frequency (p and $q = 1 - p$) and incorporating an ascertainment correction, where T_A is the total number of non-missing alleles, X and Y are the number of A and a alleles, so $p = X/T_A$ and $q = Y/T_A$.

Conditional on IBD state $Z = z$ for the entire genome, the expected count of SNPs with IBS state $I = i$ is given by $N(I = i|Z = z) = \sum_{m=1}^{L} P(I = i|Z = z)$, where the summation is over all SNPs with genotype data on both individuals. Then from Eq. 1 we can obtain global IBD estimates of $P(Z)$ for that pair by the method of moments, substituting into

$$
\begin{aligned}
P(Z=0) &= \frac{N(I=0)}{N(I=0|Z=0)} \\
P(Z=1) &= \frac{N(I=1) - P(Z=0) \cdot N(I=1|Z=0)}{N(I=1|Z=1)} \\
P(Z=2) &= \frac{N(I=2) - P(Z=0) \cdot N(I=2|Z=0) - P(Z=1) \cdot N(I=2|Z=1)}{N(I=2|Z=2)}
\end{aligned}
$$

These estimates of $P(Z)$ are not bounded $0 \leq x \leq 1$ and are also not constrained to biologically plausible values (e.g. 0.5, 0, 0.5 are not plausible values for IBD states 0, 1 and 2). In practice, we bound these estimates as follows. If $P(Z = 0) > 1$ then $P(Z = 0)$ is set to 1 and $P(Z = 1)$ and $P(Z = 2)$ are set to 0. If $P(Z = 0) < 0$ then $P(Z = 0)$ is set to 0 and $P(Z = 1)$ and $P(Z = 2)$ are set to $P(Z = 1)/S$ and $P(Z = 2)/S$ respectively, where $S = P(Z = 1) + P(Z = 2)$.

If $\pi^2 \leq P(Z = 2)$ where $\pi = P(Z = 1)/2 + P(Z = 2)$, to constrain IBD estimates to biologically plausible values (assuming a homogeneous, random-mating population), we

find a new value for $P(Z = 1)$, which we label $P_*(Z = 1)$, which satisfies the equation $(P_*(Z = 1) + 2\pi^2)/2 = \pi$, which gives $P_*(Z = 1) = 2\pi(1 - \pi)$. The transformed IBD probabilities, used in all subsequent calculations, are therefore

$$\begin{aligned} P_*(Z = 0) &= (1 - \pi)^2 \\ P_*(Z = 1) &= 2\pi(1 - \pi) \\ P_*(Z = 2) &= \pi^2. \end{aligned}$$

When not constrained to biologically plausible values, genome-wide IBD sharing estimates can be used for quality control, to indicate and diagnose sample and genotyping errors including swaps, duplications, and contamination events as well as mis-specified or undetected familial relationships. For example, values of $P(Z = 2)$ near 1 clearly indicate duplicated samples (or MZ twins). Alternatively, if an experiment is conducted on two separate chips (e.g. two 250K SNP arrays comprising a 500K array) values near 0.5, 0 and 0.5 might represent an individual duplicated for one 250K array only.

If DNA from one or more individual contaminates other samples, this can lead to a distinctive pattern of contaminated samples showing high IBD with all other individuals. This is because contamination induces false heterozygote calls (e.g. AA pooled with CC may well be typed as AC) and heterozygotes cannot be IBS 0 with any other SNP genotype, which artificially inflates IBD estimates. Furthermore, contaminated samples will show strong, negative inbreeding coefficients, indicating more heterozygotes than expected.

Estimation of local, segmental identity-by-descent sharing

Analogous to the traditional Lander-Green algorithm for multilocus analysis[33], we use a hidden Markov model (HMM) approach to provide multipoint estimates of allele-sharing IBD, for each pair of individuals in a homogeneous sample, at any arbitrary position along the chromosome, given the observed pattern of IBS sharing. Note that, unlike full inheritance vectors, IBD states along the chromosome do not actually satisfy the Markov property. Nevertheless, we have used a hidden Markov model because it is computationally tractable and likely to give a good approximation.

We require the conditional probability of IBD for $z = 0$, 1, or 2, at a particular position, given the marker genotypes, M, of all K markers on a chromosome, $P(Z = z|M)$. This can be re-expressed using Bayes Theorem as

$$P(Z = z|M) = \frac{P(M|Z = z)P(Z = z)}{P(M)} = \frac{P(M|Z = z)P(Z = z)}{\sum_{z'=0}^{2} P(M|Z = z')P(Z = z')}$$

Here $P(Z = z)$ is the global IBD sharing probability for the whole genome, and the summation is over the 3 possible IBD states. Because of the Markov property, the probability $P(M|Z = z)$ can be factorized as the product $P(M_L|Z = z) \cdot P(M_R|Z = z)$, where M_L and M_R are the marker genotypes to the left and to the right of the position, respectively. Suppose the position is between markers l and $l+1$, then the Markov property ensures that

$$P(M_L|Z = z) = \sum_{z_l, z_{l-1}, \ldots, z_1} P(Z_l = z_l|Z = z)P(M_l|Z_l = z_l)$$

$$\times P(Z_{l-1} = z_{l-1}|Z_l = z_l)P(M_{l-1}|Z_{l-1} = z_{l-1})$$
$$...$$
$$\times P(Z_1 = z_1|Z_2 = z_2)P(M_1|Z_1 = z_1)$$

where the summation is over all possible IBD states for all markers. Writing the 3×3 diagonal matrix of marker genotype probabilities conditional on IBD state for marker l as $\mathbf{M_l}$ and the 3×3 transition matrix between marker l and $l+1$ as $\mathbf{T_l}$, (where element t_{ij} is the conditional probability of marker l having IBD state j given that marker $l+1$ has IBD state i), this summation can be written in matrix form as

$$\begin{aligned} P(M_L|Z = z) &= \mathbf{z}^{'}\mathbf{T_L}\mathbf{M_l}\mathbf{T_l}\mathbf{M_{l-1}}\mathbf{T_{l-1}}...\mathbf{T_2}\mathbf{M_2}\mathbf{T_1}\mathbf{M_1}\ \mathbf{1} \\ &= (\mathbf{1}^{'}\ \mathbf{M_1}\mathbf{T}_{\mathbf{1}}^{'}\mathbf{M_2}\mathbf{T}_{\mathbf{2}}^{'}...\mathbf{M_l})\ \mathbf{T}_{\mathbf{L}}^{'}\ \mathbf{z} \end{aligned}$$

where $\mathbf{T_L}$ is the transition matrix between marker l and the position, $\mathbf{1}$ is a 3×1 vector of 1s and $\mathbf{z}$ is a 3×1 column vector which has value 1 for element z and value 0 for the others. The elements of $\mathbf{M}$ and $\mathbf{T}$ are given in Tables 2 and 4 respectively and are described in the section below.

The expression $P(M_L|Z)$ represents the "left conditional" probability, based on markers 1 through l; the same procedure is used to calculate the right conditional probability from markers K back through $l+1$

$$P(M_R|Z = z) = (\mathbf{1}^{'}\ \mathbf{M_K}\mathbf{T}_{\mathbf{K-1}}^{'}\mathbf{M_{K-1}}\mathbf{T}_{\mathbf{K-2}}^{'}...\mathbf{T}_{\mathbf{l+1}}^{'}\mathbf{M_{l+1}})\ \mathbf{T}_{\mathbf{L+1}}^{'}\ \mathbf{z}$$

where where $\mathbf{T_{L+1}}$ is the transition matrix between marker $l+1$ and the position.

For computational efficiency, the "left conditionals" ($\mathbf{M_1}\mathbf{T}_{\mathbf{1}}^{'}\mathbf{M_2}\mathbf{T}_{\mathbf{2}}^{'}...\mathbf{M_l}$) and "right conditionals" ($\mathbf{M_K}\mathbf{T}_{\mathbf{K-1}}^{'}\mathbf{M_{K-1}}\mathbf{T}_{\mathbf{K-2}}^{'}...\mathbf{T}_{\mathbf{l+1}}^{'}\mathbf{M_{l+1}}$) are pre-computed for different values of l, and used in the computation for conditional IBD probabilities of all positions. In practice, normalization steps must be performed at intervals to avoid underflow.

Based on these estimates of $P(Z = z|M)$, shared segments are defined as any contiguous region above 50% chance of having at least 1 pair of chromosomes shared IBD (we can ignore the negligible probability of distantly-related individuals sharing 2 segments IBD).

Description of hidden Markov model emission and transition parameters, $\mathbf{M}$ *and* $\mathbf{T}$

The elements of $\mathbf{M}$, the probability of the pair's genotypes for that SNP, conditional on IBD state, $P(M|Z)$, are calculated according to Table 2. These values are a function of allele frequency, including an ascertainment correction term following the procedure described above in the calculation of global IBD probabilities. Markers with missing genotypes are assigned an identity matrix $\mathbf{M}$.

The elements of $\mathbf{T}$, the transition probabilities between two IBD states at neighboring loci, are precalculated in terms of the recombination fraction estimated from a specified genetic map and global relatedness for that pair, $P(Z)$. With dense SNP maps, the method is not particularly sensitive to the precise genetic map used: in practice, a basic $1cM = 1Mb$ approximation appears to work well, although one could also use the fine-scale recombination map[34]. For each pair of individuals, we estimate the least number of meioses that separate the two genomes and use these estimates to specify a transition matrix for unobserved IBD states along the chromosome. Specifically, we consider two chromosomes, or

$M = G_1, G_2$	Z	$P(M\|Z)$
AA, AA	0	$p^4 \cdot \left(\frac{X-1}{X} \cdot \frac{X-2}{X} \cdot \frac{X-3}{X} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$
AA, Aa	0	$4p^3q \cdot \left(\frac{X-1}{X} \cdot \frac{X-2}{X} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$
AA, aa	0	$2p^2q^2 \cdot \left(\frac{X-1}{X} \cdot \frac{Y-1}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$
Aa, Aa	0	$4p^2q^2 \cdot \left(\frac{X-1}{X} \cdot \frac{Y-1}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$
Aa, aa	0	$4pq^3 \cdot \left(\frac{Y-1}{Y} \cdot \frac{Y-2}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$
aa, aa	0	$q^4 \cdot \left(\frac{Y-1}{Y} \cdot \frac{Y-2}{Y} \cdot \frac{Y-3}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2} \cdot \frac{T_A}{T_A-3}\right)$
AA, AA	1	$p^3 \cdot \left(\frac{X-1}{X} \cdot \frac{X-2}{X} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2}\right)$
AA, Aa	1	$2p^2q \cdot \left(\frac{X-1}{X} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2}\right)$
AA, aa	1	0
Aa, Aa	1	$p^2q \cdot \left(\frac{X-1}{X} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2}\right)$ $+pq^2 \cdot \left(\frac{Y-1}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2}\right)$
Aa, aa	1	$2pq^2 \cdot \left(\frac{Y-1}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2}\right)$
aa, aa	1	$q^3 \cdot \left(\frac{Y-1}{Y} \cdot \frac{Y-2}{Y} \cdot \frac{T_A}{T_A-1} \cdot \frac{T_A}{T_A-2}\right)$
AA, AA	2	$p^2 \cdot \left(\frac{X-1}{X} \cdot \frac{T_A}{T_A-1}\right)$
AA, Aa	2	0
AA, aa	2	0
Aa, Aa	2	$2pq \cdot \left(\frac{T_A}{T_A-1}\right)$
Aa, aa	2	0
aa, aa	2	$q^2 \cdot \left(\frac{Y-1}{Y} \cdot \frac{T_A}{T_A-1}\right)$

Table 2: Calculation of genotypic state M given IBD state Z for a particular SNP, $P(M|Z)$, as a function of allele frequency (p and $q = 1-p$ for alleles A and a respectively). T_A is the total number of non-missing alleles, X and Y are the number of A and a alleles, so $p = X/T_A$ and $q = Y/T_A$.

haploid genomes, (C_1 and C_2) which share a common ancestor (C_C, a diploid genome), with C_1 and C_2 being separated by m meioses. If C_1 and C_2 are present in distinct individuals then $m \geq 2$, whereas if C_1 and C_2 are in the same person, then $m \geq 3$. At a particular locus, U, the probability that C_1 and C_2 are IBD is $(1/2)^{(m-1)}$. That is, if we label the allele transmitted in the first meiosis from C_C as u_1, there is probability $(1/2)^{(m-1)}$ that all the other $(m-1)$ meioses also transmit u_1. Now consider a second locus V that is linked to U with recombination fraction θ. Let the alleles at V present in C_C be v_1 and v_2, with allele v_1 being in coupling phase with allele u_1. Figure 2 shows some examples of possible transmission patterns and the corresponding IBD states for C_1 and C_2, which we outline here.

For C_1 and C_2 to be IBD at V, they must either both share allele v_1 or allele v_2. In order for C_1 and C_2 to both share allele v_1, given that C_1 and C_2 are IBD at U for allele u_1 that is in coupling phase with v_1 in the common ancestor C_C, all the m meioses must

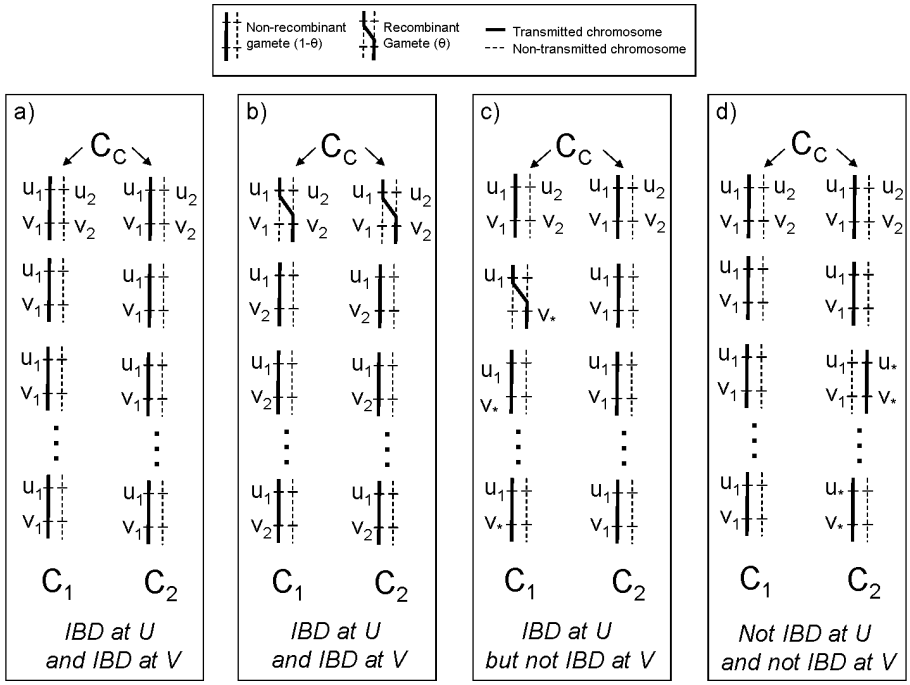

Figure 2. **Example transmissions and corresponding IBD states.** For two haploid genomes, C_1 and C_2, the figure illustrates four (of many) possible patterns of transmission and the corresponding IBD states at two positions, U and V. The main text describes how consideration of these possible scenarios leads to the specification of transition matrices for IBD state along the chromosome.

be non-recombinants, so that v_1 is co-transmitted with u_1 all the way down to C_1 and C_2. The probability of this is $(1-\theta)^m$. In order for C_1 and C_2 to both share allele v_2, given that C_1 and C_2 are IBD at U for allele u_1 that is in repulsion phase with v_2 in the common ancestor C_C, the two meioses of C_C must both be recombinants (so that v_2 crosses over to be co-transmitted with u_1) and all the remaining $(m-2)$ meioses must be non-recombinants, so that v_2 is co-transmitted with u_1 all the way down to C_1 and to C_2. The probability of this is $\theta^2(1-\theta)^{(m-2)}$. Taking these two possibilities together, the probability that C_1 and C_2 are IBD at V, given that they are IBD at U, is

$$P(IBD_V|IBD_U) = (1-\theta)^m + \theta^2(1-\theta)^{(m-2)}$$

which can be rewritten as $(1-\theta)^{(m-2)}(\theta^2 + (1-\theta)^2)$. The probability that C_1 and C_2 are IBD at V, given that they are not IBD at U, is given by Bayes theorem as follows:

$$P(IBD_V|\overline{IBD_U}) = \frac{P(\overline{IBD_U}|IBD_V)P(IBD_V)}{P(\overline{IBD_U})}$$

which simplifies to

$$(1-(1-\theta)^{(m-2)}(\theta^2 + (1-\theta)^2))/(2^{(m-1)} - 1).$$

This equation involves two parameters, θ and m. Because of the assumption of very closely spaced markers, we use the Morgan map function $\theta = d$, where d is the genetic distance between the loci in Morgans. For m, we consider the two pairs of haploid genomes (each pair containing a haploid genome from each individual) that may have common ancestry and estimate the numbers of meioses (m_A and m_B) that separate the two pairs. If x_A and x_B are the probabilities that the two pairs of haploid genomes are IBD, then

$$P(Z = 2) = x_A x_B$$

and

$$P(Z = 0) = (1 - x_A)(1 - x_B).$$

Substituting $x_B = P(Z = 2)/x_A$ into the second expression we obtain

$$x_A^2 - (P(Z = 1) + 2P(Z = 2))x_A + P(Z = 2) = 0.$$

Solving this equation gives x_A and x_B. These IBD probabilities (x) are related to the number of meioses separating the haploid genomes (m) by $x = (1/2)^{(m-1)}$. Therefore, if $P(Z = 2) = 0$ then

$$\begin{aligned} m_A &= 1 - \log(P(Z = 1))/\log(2) \\ m_B &= 0, \end{aligned}$$

otherwise

$$\begin{aligned} m_A &= 1 - \log(x_A)/\log(2) \\ m_B &= 1 - \log(x_B)/\log(2), \end{aligned}$$

Note that the quadratic equation

$$x_A^2 - (P(Z = 1) + 2P(Z = 2))x_A + P(Z = 2) = 0$$

has real roots only if

$$(P(Z = 1) + 2P(Z = 2))^2 \geq 4P(Z = 2).$$

This gives rise to the inequality $P(Z = 2) \leq \pi^2$ in the previous section on constraining global IBD estimates to biologically plausible values. The form of the transition matrix for pairs of haploid genomes (denoted **A** and **B**) is given in Table 3; these are combined to form the full transition matrix for a diploid genome in Table 4.

Application of segmental sharing analysis

One requirement of this approach is that SNPs are in approximate linkage equilibrium in the population, otherwise many small regions of high linkage disequilibrium (LD) will be called as shared IBD segments. One approach is to prune the SNP panel to a reduced subset of approximately independent SNPs. As outlined in the next section, for the purpose of population-based linkage analysis, we expect most segments surrounding shared rare, recent variants to be relatively large, and therefore detectable with a less dense SNP panel. We use a repeated, sliding window procedure, recursively pruning SNPs based on pairwise SNP

	$h_0^{(l+1)}$	$h_1^{(l+1)}$
$h_0^{(l)}$	$1 - \frac{1-(1-\theta)^{m-2}\xi}{2^{m-1}-1}$	$\frac{1-(1-\theta)^{m-2}\xi}{2^{m-1}-1}$
$h_1^{(l)}$	$1-(1-\theta)^{m-2}\xi$	$(1-\theta)^{m-2}\xi$

Table 3: Structure of transition submatrices **A** (where $m = m_A$) and **B** (where $m = m_B$) for two haploid genomes: probability of haploid IBD states 0 (h_0) and 1 (h_1) at locus $l+1$ conditional on state at locus l; m is the estimate of the least number of meioses for that haploid pair of genomes and $\xi = \theta^2 + (1-\theta)^2$.

	$Z_{l+1} = 0$	$Z_{l+1} = 1$	$Z_{l+1} = 2$
$Z_l = 0$	$a_{00}b_{00}$	$a_{00}b_{01} + a_{01}b_{00}$	$a_{01}b_{01}$
$Z_l = 1$	$\frac{a_{00}b_{10}+a_{10}b_{00}}{2}$	$\frac{a_{00}b_{11}+a_{01}b_{10}+a_{10}b_{01}+a_{11}b_{00}}{2}$	$\frac{a_{01}b_{11}+a_{11}b_{01}}{2}$
$Z_l = 2$	$a_{10}b_{10}$	$a_{10}b_{11} + a_{11}b_{10}$	$a_{11}b_{11}$

Table 4: Full transition matrix **T** specifying $P(Z_{l+1}|Z_l)$ for diploid IBD state from locus l to locus $l+1$; a_{ij} and b_{ij} represent the corresponding elements of the 2×2 transition matrices **A** and **B** shown in Table 3, transitioning from IBD state i to state j for a single pair of chromosomes.

r^2 values, and/or the variance inflation factor (VIF), which is defined as $1/(1-R^2)$ where R^2 is the multiple correlation coefficient between a SNP and all other SNPs in the window based on allele counts. In the context of the population-based linkage test described below, failure to completely prune all sample-level LD should not be particularly troublesome. At worst, it will simply mean that a number of more common extended segments (that are perhaps better tested in a standard association design) will be included in the analysis, possibly reducing efficiency.

The results of applying this method to Phase II HapMap data, in terms of the typical distribution and extent of extended segmental sharing and its relationship to rare variation, are described in the International HapMap Consortium Phase II analysis manuscript[35]. Here we apply the method to CEU (Utah residents with European ancestry from the CEPH collection) HapMap individuals. For illustrative purposes, here we focus only on a particular region of chromosome 9 shared between two families (parent-offspring trios). Genome-wide data on the full CEU panel were still used to calculate allele frequencies (founders only) and global IBD sharing estimates. We selected SNPs with complete genotyping and MAF greater than 1%, and then iteratively removed SNPs showing local LD; the final chromosome 9 dataset comprised 6513 SNPs (approximately 1 per 20kb). In the CEU sample as a whole, there is virtually no LD in the pruned dataset: for the entire chromosome 9, only 6 LD blocks are identified, comprising 15 SNPs in total (3 blocks containing 2 SNPs and 3 blocks containing 3 SNPs).

We present here a single segment shared between two CEU offspring, NA10863 and NA06991. This segment spans approximately 3.7Mb and, in the pruned dataset, 272 SNPs. The Phase II HapMap has over 4500 SNPs in this region (CEU panel) and dozens of estimated recombination hotspots. We selected a segment shared between two offspring for illustrative purposes: given the region is shared IBD, we would also expect to see a pattern of sharing consistent with transmission from one and only one parent in each family, as

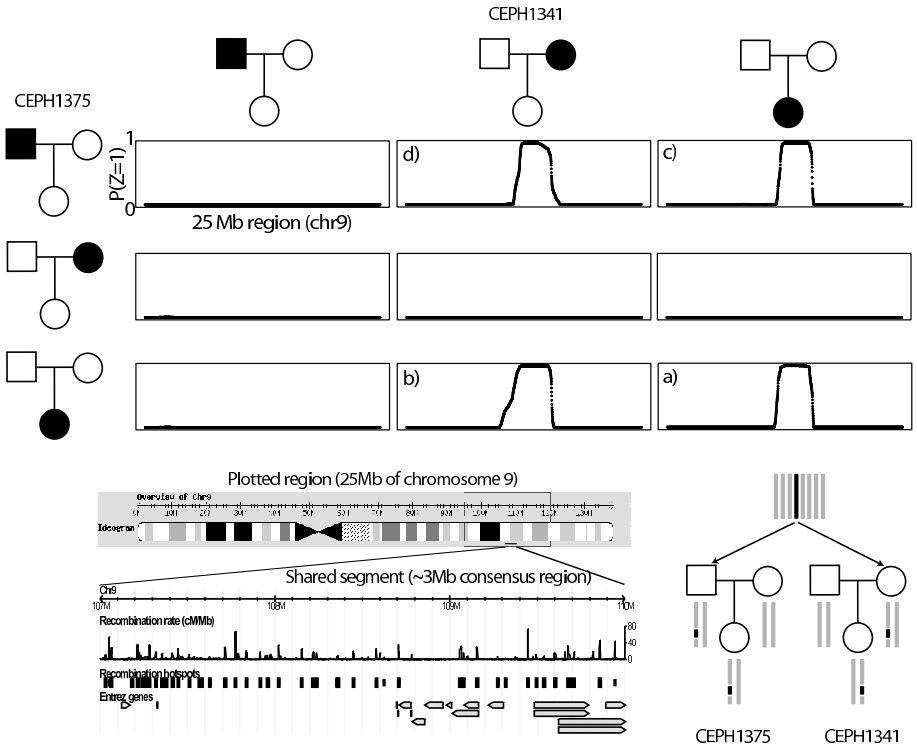

Figure 3. **Example segment shared IBD between two HapMap CEU offspring individuals and their parents.** The main set of plots show the multipoint estimate of IBD sharing, $P(Z = 1)$, for a 25Mb region of chromosome 9, for the pairs of individuals between two families (CEPH1375 and CEPH1341). The region was selected because the two offspring (NA10863 and NA06991) showed sharing in this region, shown in plot (a). The 3 other segments shared between seemingly unrelated individuals are shown, i.e. between the offspring in one family and a parent in the other family, two plots labelled (b), and between those two parents, labelled (c). The lower left diagram illustrates the region shared; this extended haplotype spans multiple haplotype blocks and recombination hotspots in the full Phase II data. The lower right diagram depicts the pattern of gene flow for this particular region: i.e. a segment of the original common chromosome (shown in dark) appears in the two families as shown.

illustrated in Figure 3. That is, we also observe the same segment shared between each offspring and one parent of the other family and also between these two parents. (Naturally, parent-offspring pairs within the same family are always IBD 1; these basic intra-familial relationships are not shown in the Figure).

No other pairs of individuals show any extended segmental sharing in this same region. This type of rare, extended segment is an example of shared genetic variation that is outside the standard heuristic and analytic framework of LD involving only short, common "haplotype blocks" separated by recombination hotspots. Naturally, this approach can also be applied to detecting the much longer segments shared between very closely related individuals. There are, in fact, a number of close relationships between HapMap founders: in the CEU and YRI (Yoruba individuals from Ibadan, Nigeria) panels there are a number of cousins and closer relationships (excluding known parent-offspring relationships of course).

For example, two CEU individuals, NA12154 and NA12264, have a global $P(Z = 1)$ of 0.14; shared segment analysis reveals at least 33 segments over 1Mb for this single pair; 6 segments are over 10Mb and the longest is 128Mb (they share virtually all of one copy of chromosome 11). For this pair, the total proportion of the autosomes spanned by these segments (in terms of physical distance) is approximately 0.12, which is close to the global probability of sharing 1 copy IBD of 0.14.

Population-based linkage analysis

In a population-based sample of cases and controls, having determined the extent and location of pairwise segmental sharing using the approach described above, one might also want to inquire whether patterns of segmental sharing are related to phenotypic similarity between individuals. In this section we describe a test based on the premises that a) shared rare variants will typically reside on shared extended segments and b) there will be an inflation in the rate of segmental sharing at the disease locus in case/case pairs if rare variants influence disease risk. This represents a first-generation approach that can no doubt be extended and improved in numerous ways.

A sample of N_A cases and N_U controls contains $N_{AA} = N_A(N_A-1)/2$ case/case pairs and $N_{!AA} = N_A N_U + N_U(N_U-1)/2$ case-control and control/control pairs. At a particular position p, the number of segments shared at a particular locus (i.e. spanning that position) is denoted S_p for all case/case pairs and T_p for case/control and control/control pairs. A standard test for a difference in rate of sharing between these two groups is complicated by the fact that not all pairs are independent (as the same individuals will possibly feature in multiple pairs); also, not all pairs have similar degrees of global relatedness. To account for the dependence, we use permutation to generate empirical significance values by label-swapping individuals' phenotypes and recalculating the pairwise phenotypic concordance metrics (i.e. rather than permuting the pairwise concordance terms directly). The test statistic is framed as a one-sided test (i.e. greater sharing in case/case pairs) and adjusts for the average level of global sharing in the two categories; for position p of L positions,

$$\left(S_p - \frac{\sum_{p'} S_{p'}}{L} \right) / N_{AA} - \left(T_p - \frac{\sum_{p'} T_{p'}}{L} \right) / N_{!AA}$$

which is bounded at 0 (in addition, $S_p - \frac{\sum_{p'} S_{p'}}{L}$ and $T_p - \frac{\sum_{p'} T_{p'}}{L}$ are also bounded at 0). The permutation procedure is computationally feasible, as the IBD segments do not need to be recomputed each replicate.

Given regions that show statistically significant levels of increased sharing among cases, one can use PLINK to determine the allelic identity of the specific sets of overlapping segments present (although without inferring phase when both individuals are heterozygous). The results of such an analysis should be similar in principle to those from a linkage analysis, except signals will be localized on a finer scale.

Performance

Very large WGAS datasets can be analyzed using fairly standard hardware and there are no fixed limits on the number of samples or SNPs. For example, a Linux workstation with

2GB RAM and 3.6GHz dual processor can handle more than 5000 individuals genotyped on 500,000 SNPs. On the same machine, one can load (using a binary format fileset) the entire Phase II HapMap (3.6 million SNPs on 270 individuals), filter on genotyping rate, then calculate and output allele frequencies for all SNPs in under 6 minutes.

Although analyses that involve pairwise comparisons between all individuals can take a long time in large samples (in particular, calculating genome-wide IBD and IBS for all individuals), if a cluster-computing environment is available, such jobs are easily parallelized and PLINK provides some options to facilitate this, potentially reducing analyses that might take days to one or two hours with little extra work on the part of the user.

Graphical user interface

We have also developed a separate, optional Java-based graphical user interface, gPLINK, to initiate, track and record PLINK jobs. In addition, gPLINK provides integration with Haploview[36]; version 4 of Haploview offers extensive tools for tabulating, filtering, sorting, merging and visualizing PLINK WGAS output files in the context of HapMap linkage disequilibrium and genomic information. gPLINK can also extract filtered subsets of WGAS data for viewing in Haploview with just a few mouse clicks. gPLINK can either be used to direct local analyses (with data and computation residing on the same, local machine) or remotely using secure shell networking (with data and computation performed by the remote server, but initiation and viewing of results done locally). Figure 4 illustrates the relationship between PLINK, gPLINK and Haploview.

Summary

In summary, PLINK offers a powerful, user-friendly tool for performing many common analyses with whole-genome data. There is comprehensive web-based documentation including a tutorial; an e-mail list and web-based version-check inform users of updates and problems. As methods for WGAS evolve, we expect that PLINK will be updated. Future directions include enhanced tools for browsing annotated WGAS results in their full genomic context and the incorporation of copy number variation data.

We have also implemented an approach based on the multiple rare variant (MRV) hypothesis, that is designed to be a complement, rather than an alternative, to association analysis. In this manuscript we have outlined our analytic approach and described an implementation of the method that is appropriate for whole-genome SNP data. Following decades of work mapping Mendelian disease genes, this approach uses haplotypes of common alleles to measure very rare variation. This is an example of how one can take multiple approaches to existing high density SNP array data, rather than needing to embark on a completely orthogonal data collection, to execute a MRV-oriented test.

Standard association will be much more powerful when a single common causal variant is directly assayed or well captured by a tag SNP. When the CD/CV hypothesis does not hold, however, we hope that this approach would perform better. In this case, straightforward association approaches are unlikely to succeed, as the rare variants will most likely be neither identified, genotyped, nor tagged with sufficient precision; in any case, there will typically be too few observations to provide adequate statistical power for standard association tests of any one rare variant.

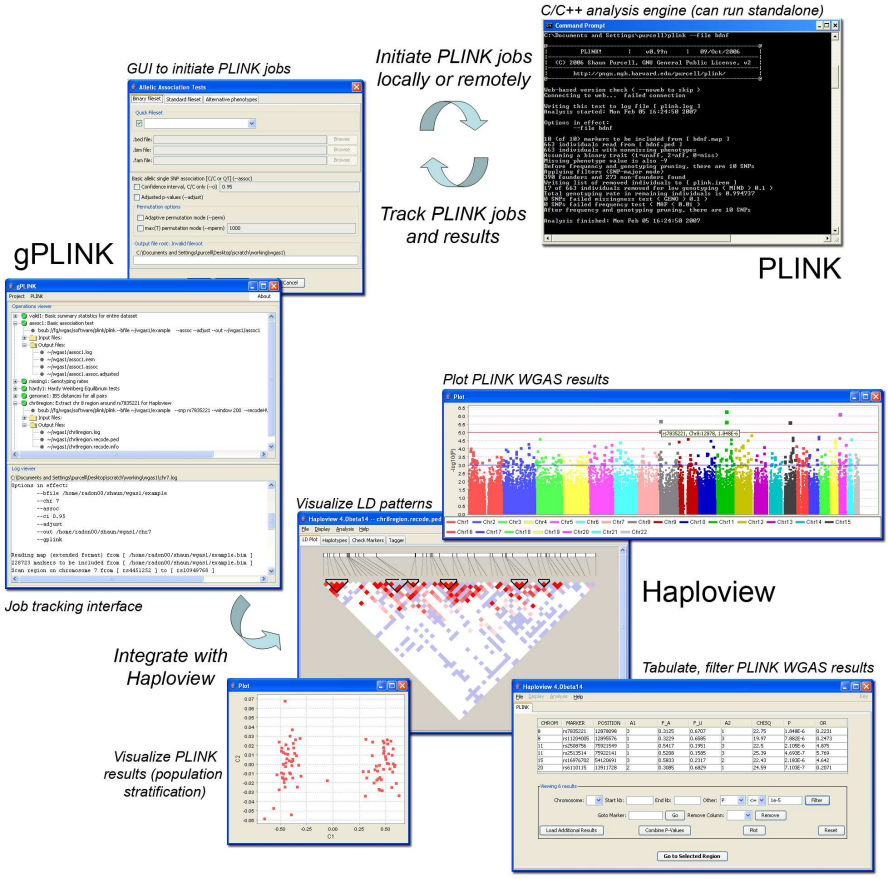

Figure 4. **Schema of integration between PLINK, gPLINK and Haploview.** PLINK is the main C/C++ WGAS analytic engine, that can either run as a standalone tool (from the command line or via shell scripting) or in conjunction with gPLINK, a Java-based graphical user interface. gPLINK also offers a simple project management framework to track PLINK analyses and facilitates integration with Haploview. It is easy to configure these tools such that the whole-genome data and PLINK analyses (i.e. the computationally expensive aspects of this process) can reside on a remote server, but all initiation and viewing of results is done locally, e.g. on a user's laptop, connected to the whole-genome data via the internet, using gPLINK's secure shell (SSH) networking.

We are currently embarking on the next step, to determine the potential power of such an approach under a range of scenarios and to determine the best way to apply this method to real data. Possible extensions of this approach include allowing for LD between SNPs, genotyping error and inbreeding in the IBD estimation. This approach could also be applied to detecting autozygous segments within a single individual, suggesting the possibility of population-based homozygosity mapping to map recessive disease loci.

Acknowledgments

We would like to acknowledge grant support from the NIH/NHLBI ENDGAME project (U01 HG004171; SP, MJD and PIWdB), from NIH grant EY-12562 (SP and PCS) and from NIH/NIMH grant R03 MH73806-01A1 (SP).

We would also like to thank the NINDS Repository at Coriell, for the data from the Laboratory of Neurogenetics, of the intramural program of the National Institute on Aging, National Institutes of Health, for making these data freely available. These data were deposited by John Hardy and Andrew Singleton; we accessed the data (upload IDs 7 and 8) at the Queue portal at the Coriell Institute

Finally, we would like to thank PLINK users both within the Broad Institute Medical and Population Genetics Program and elsewhere for all feedback.

Web Resources

PLINK and gPLINK:
`http://pngu.mgh.harvard.edu/purcell/plink/`
Haploview:
`http://www.broad.mit.edu/mpg/haploview/`
HapMap:
`http://www.hapmap.org`
Queue portal at the Coriell Institute:
`https://queue.coriell.org/q`

References

1. Lander ES & Schork NJ (1995) Genetic dissection of complex traits. Science, 265: 2037-2048.

2. Risch N & Merikangas K (1996) The future of genetic studies of complex human diseases. Science, 273(5281): 1516-7.

3. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med, 4(2): 45-61.

4. Ioannidis JP, Trikalinos TA, Khoury MJ (2006) Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. Am J Epidemiol, 164(7):609-14.

5. Reich DE & Lander ES (2001) On the allelic spectrum of human disease. Trends Genet, 17(9):502-10.

6. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, et al (2004) Assessing the impact of population stratification on genetic association studies. Nat Genet, 36(4): 388-93.

7. Devlin B & Roeder K (1999) Genomic Control for Association Studies. Biometrics, 55: 997-1004.

8. Pritchard JK, Stephens M, Donnelly PJ (2000) Inference of population structure using multilocus genotype data. Genetics, 155:945-959.

9. Purcell S, Sham PC (2004) Properties of structured association approaches to detecting population stratification. Hum Hered, 58(2):93-107.

10. 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.

11. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, et al (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet, 37(11):1243-6.

12. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet, 6(2): 95-108.

13. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet, 69(1):124-37.

14. Wigginton JE, Cutler DJ, Abecasis GR (2005) A Note on Exact Tests of Hardy-Weinberg Equilibrium. Am J Hum Genet, 76:887-893.

15. Lee WC (2003) Detecting population stratification using a panel of SNPs. Int J Epidemiol, 32:1120.

16. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University and Novartis Institutes for BioMedical Research (in press) Genome-wide association analysis identifies novel loci for type 2 diabetes and triglyceride levels. *Science*

17. Plenge R, Cotsapas C, Davies L, Price AL, de Bakker PIW, Maller J, Pe'er I, Burtt N, Blumenstiel B, DeFelice M (submitted) Whole genome association study using shared controls identifies rheumatoid arthritis risk locus near TNFAIP3.

18. Agresti A (1990) *Categorical Data Analysis*. New York: John Wiley and Sons; pp. 100-102.

19. Fleiss JL (1981) Statistical Methods for Rates and Proportions, 2nd Ed., New York: Wiley.

20. Ewens WJ & Spielman RS (1995) The Transmission/Disequilibrium Test: History, Subdivision, and Admixture. Am J Hum Genet, 57:455-465.

21. Spielman RS & Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet, 62(2), 450-458.

22. Fulker DW, Cherny SS, Sham PC & Hewitt JK (1999) Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet, 64(1):259-267.

23. Abecasis GR, Cardon LR & Cookson WO (2000) A General Test of Association for Quantitative Traits in Nuclear Families. Am Journal Hum Genet 66:279-292.

24. Purcell S, Sham PC & Daly MJ (2005) Parental phenotypes in family-based association analysis. Am J Hum Genet, 76(2):249-259.

25. Ferreira MAR, Sham PC, Daly MJ & Purcell S (in press) Ascertainment through family history of disease often decreases the power of family-based association studies. *Behav Genet*.

26. Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D & Daly MJ (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet, 38(6): 605-6.

27. Hoh J, Wille A & Ott J (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res, 11:2115-2119.

28. Besag J & Clifford P (1991) Sequential Monte Carlo p-values. Biometrika, 79:301-304.

29. Churchill GA, & Doerge RW (1996) Empirical Threshold Values for Quantitative Triat Mapping. Genetics, 142: 285-294.

30. Benjamini Y & Hochberg Y (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. J R Stat Soc Ser B, 57:289-300.

31. Broman KW & Weber JL (1999) Long Homozygous Chromosomal Segments in Reference Families from the Centre d'Etude du Polymorphisme Humain. Am J Hum Genet, 65:1493-1500.

32. Puffenberger EG, Hu-Lince D, Parod JM, Craig DW, Dobrin SE, Conway AR, Donarum EA, Strauss KA, Dunckley R, Cardenas JF, et al (2004) Mapping of sudden infant death with dysgenesis of the testes syndrome (SIDDT) by a SNP genome scan and identification of TSPYL loss of function. Proc Natl Acad Sci USA, 101(32):11689 - 11694.

33. Lander ES & Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA, 84:2363-2367

34. Myers S, Bottolo L, Freeman C, McVean G & Donnelly P (2005) A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. Science, 310(5746):321-324.

35. International HapMap Consortium (submitted) The Phase II HapMap.

36. Barrett JC, Fry B, Maller J & Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2): 263-265.