Published in final edited form as:

Stat Biosci. 2013 November ; 5(2): . doi:10.1007/s12561-013-9090-0.

Interpreting Whole-Genome Marker Data

Bruce S. Weir

Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195-7232, USA, Tel.: +1-206-221-7947, Fax: +1-206-543-3286

Bruce S. Weir: bsweir@uw.edu

Abstract

The challenges of whole-genome data, when genotypes are available from hundreds of thousands of genetic markers, are explored for four topics in statistical genetics: Hardy-Weinberg testing, estimating linkage disequilibrium from unphased genotypic data, association mapping and characterizing population structure.

Keywords

Hardy-Weinberg; Linkage Disequilibrium; Association Mapping; Population Structure; SNPs

1 Introduction

The study of human population genetics has been transformed by the explosion in data over the past few years. New technologies have made it possible to type thousands of individuals at millions of single nucleotide polymorphism (SNP) markers and then impute genotypes at many more millions. Meta-analyses are being published from such data collected from over 100,000 individuals. As valuable as these whole-genome data sets are, the resulting inferences can only be as good as the data: data cleaning has become a key component of statistical analyses. This discussion addresses some of the common activities of inference and points to the need for a careful examination of the data, with special attention to missing data. It is very unlikely that every study sample will be typed at every SNP, and checks of missingness rates are standard aspects of whole-genome data cleaning [1]. An extreme view would be that any missingness is an indicator of typing problems at that SNP, so that consideration could be given to discarding information from that SNP. Some empirical support for this view is shown below, although a degree of relaxation will be needed in practice, especially with data coming from DNA sequencing studies.

2 Hardy-Weinberg Testing

One of the earliest results in statistical genetics was the Hardy-Weinberg law [2], [3] which states that the probability of an individual carrying a specific pair of alleles is the product of the probabilities of each allele, with a factor of two when the alleles are of different types. The interpretation of whole-genome SNP data requires an examination of whether the data are consistent with the law for two different reasons. First there is a data cleaning aspect: the law is expected to hold in large human populations for which there is no population structure so that observed departures from the law may indicate poor quality data such as a sample of having no heterozygotes or nothing but heterozygotes. This data-cleaning aspect is appropriate if any observed departures from Hardy-Weinberg equilibrium (HWE) are scattered over the genome, whereas localized departures may indicate the effects of evolutionary forces such as selection. The second motivation for Hardy-Weinberg testing is that HWE is assumed for many tests for marker-trait association.

It is only recently that the limitations of Hardy-Weinberg testing have been fully appreciated [4], [5] with these limitations stemming from small sample sizes or small numbers of the least frequent allele. Certainly the goodness-of-fit test statistic, comparing observed and expected genotype counts, is known to be only approximately chi-square distributed under the null hypotheses of HWE and the exact test [6] has long been preferred. The difficulties with the exact test arise when continuous-variable theory is applied to discrete genotypic data. If the two alleles for a SNP marker are A and a the probability for the counts n_{AA} , n_{Aa} , n_{aa} of genotypes AA , Aa , aa , conditional on the allele counts $n_A = 2n_{AA} + n_{Aa}$ and $n_a = 2n_{aa} + n_{Aa}$, can be written as $\Pr(n_{Aa}|n_A)$

where

$$\Pr(n_{Aa}|n_A) = \frac{C\psi^{n_{Aa}}}{n_{AA}!n_{Aa}!n_{aa}!} \quad (1)$$

The quantity ψ is a combination of the proportions P_{AA} , P_{Aa} , P_{aa} of the three genotypes in the population: $\psi = P_{Aa} / \sqrt{(P_{AA}P_{aa})}$. Without loss of generality we can take A to be the minor allele, n_A the minor allele frequency (MAF). Then C is a normalizing constant to ensure that $\sum n_{Aa} \Pr(n_{Aa}|n_A) = 1$ and the sum is over the integers $0, 2, \dots, n_A$ when n_A is even and over $1, 3, \dots, n_A$ when n_A is odd. When the population is in HWE, $\psi = 2$. Equation 1 shows the counts n_{AA} , n_{aa} even though these follow from the sample size $n = n_{AA} + n_{Aa} + n_{aa}$ and from n_A : $n_{AA} = (n_A - n_{Aa})/2$, $n_{aa} = n - n_{AA} - n_{Aa}$.

The exact test for HWE proceeds by evaluating equation 1 with $\psi = 2$ for the data at hand, to give a value P . The equation is also evaluated for all values of n_{Aa} consistent with the observed n_A and the sum of such values less than or equal to P provides the p -value for data. A conventional one-marker test would reject HWE if the p -value was less than 0.05.

It is convenient to display the results of genome-wide HWE tests by means of a QQ-plot, where the observed set of ranked p -values is plotted against the values expected if all the markers were in HWE. It is usually assumed that the p -values have a null distribution that is uniform over the interval (0,1). As interest is mainly in the smallest p -values, it is also usual to transform observed and expected p -values to $-\log_{10}$ of their values. A QQ-plot is shown as Figure 1 for 9208 SNPs on chromosome 1 scored in 50 individuals used as controls in a study of age-related macular degeneration (AMD) [7]. It can be seen that the observed p -values depart from expectation around $p = 0.01$, whereas the usual Bonferroni correction for a 5% family-wise significance level would suggest rejecting HWE when p was less than $0.05/9208 = 5 \times 10^{-6}$ or $-\log_{10}(p) > 5.3$. The test is clearly detecting errors in the data however: the two largest values of $-\log_{10}(p)$ occur for SNPs with no heterozygotes ($n_{AA} = 30, n_{Aa} = 0, n_{aa} = 15$) or (almost) no homozygotes ($n_{AA} = 2, n_{Aa} = 47, n_{aa} = 0$).

For any allele count n_A there are only $v = \lfloor n_A/2 \rfloor + 1$ values of n_{Aa} , where $\lfloor x \rfloor$ is the largest integer less than or equal to x . This means that there are only v values of equation 1 to be evaluated and only v possible p -values for any given sample size n and minor allele count n_A . For a sample of size 50, for example, the maximum value of n_A is 100 and this would allow 51 p -values. At the other extreme, a sample with a single minor allele, $n_A = 1$, has only the p -value of 1. Moreover, one of the v p -values must equal 1 so the null distribution of these values is far from uniform as would be the case for continuously-distributed test statistics. Assuming uniformity to give expected p -values may be quite misleading. Much of the non-uniformity of the distribution of p -values is caused by low minor-allele counts and it has been suggested [5] that p -value null distributions formed by weighting distributions by the minor allele frequencies (MAF) in a data set. However, if the SNPs in the AMD controls

were filtered by MAF, so that SNPs with fewer than six copies of the minor allele were excluded, reducing the number of SNPs from 9208 to 6929, no appreciable impact on the highest values of $-\log_{10}(p)$ was found. Even more drastic excluding of SNPs with fewer than 26 minor alleles had little effect on Figure 1.

A different SNP filtering approach led to Figure 2, where now there are only five SNPs that appear to have large departures from HWE. This filtering was on missingness: only those SNPs typed in every one of the 50 individuals were used. This reduced the 9208 SNPs to 7446. The rationale for this kind of extreme filtering is that if the genotype for a SNP could not be determined in even one individual there may be issues with typing of that SNP that render it unsuitable for further analysis. The mere fact of some missingness for a SNP suggests the kinds of technical errors that HWE is intended to detect. It is interesting to note that all five extreme SNPs indicated as outliers in Figure 2 had zero counts for the minor allele homozygote and had at least 32 heterozygotes: small MAF values was not the issue.

The case for extreme filtering prior to HWE testing is bolstered by some additional work with other genotyping platforms (S. Gogarten personal communication). With the Illumina Omni2.5 array applied to a sample of 1258 individuals, for example, using all 2,377,238 SNPs led to a QQ plot where departures from the expected $-\log_{10}(p)$ -values began around 2.5. Examining only the 2,066,422 SNPs with minor allele frequencies above 0.005 showed departures starting at 3.5, whereas examining the 339,574 SNPs with no missing data raised the departure to nearer 4.5 and less than 10 SNPs in the set with p values outside the expected range.

HWE is a statement about pairs of alleles within individuals at a single marker. Consistency of the data with HWE suggests that the alleles are independent, as will result if genotypes are determined without error in large randomly-mating populations when there are no disturbing forces such as selection, mutation or migration. Failure to reject the HWE hypothesis, of course, does not imply that the data are error free or that there have been no disturbing forces acting on the population in the past. This general observation that failure to reject does not prove the null is especially relevant for HWE testing because of the low power of tests for SNP data.

Consider a sample of size $n = 50$ for which $n_A = 35$. There are 18 possible sets of genotypes, with heterozygote counts ranging from 1 to 35, and the probabilities for these sets conditional on n_a are shown in Table 1 for a range of values of f . The central column, $f = 0$, is for HWE. The three columns to the left of center ($f < 0$) have homozygote excess and the three columns to the right of center ($f > 0$) have heterozygote excess. If a conventional 5% significance level is adopted, the rejection region for HWE is $n_{AA} \leq 15$ (probability under HWE of 0.0150) plus $n_{Aa} \geq 31$ (probability under HWE of 0.0116). The empirical significance level is 0.0266 and it is not possible to get closer to but less than 0.05. The last row of Table 1 shows the probabilities of obtaining a value in the rejection region under all seven considered values of f – all of these power values are small.

It is more usual to use the inbreeding coefficient f than the quantity Δ to characterize departures from HWE. This parameter can be defined from the relationship

$P_{AA} = p_A^2 + p_A p_a f$, and values of f are shown in the column headings for Table 1, assuming population allele frequencies of $p_A = 0.35$, $p_a = 0.65$. Values of $f = \pm 0.10$ for human populations are very large, yet will have little chance of being detected in samples as small as $n = 50$. Approximate sample-size determinations follow from regarding the estimated value $f = 1 - 2n_{Aa}/n_A n_a$ as being normally distributed: 90% power for 5% significance when n^2 is regarded as having a chi-square distribution with 1 df requires a sample size of at least

$10.5/\bar{f}^2$. A sample of at least 1,000 is needed to have good power for detecting $f=0.1$, and this number increases to 4,000 for $f=0.05$.

3 Linkage Disequilibrium

A major use of whole-genome marker data is to locate genes associated with human disease. Traditionally these mapping exercises were based on large pedigrees and used the transmission of marker alleles and disease status down the pedigree to estimate the recombination fraction between marker and disease loci. The precision of such studies depended on the numbers of meioses, or opportunities for recombination, in the pedigree and this was rarely more than a few hundred. Recombination fraction, maybe converted to genetic map distance, is regarded as a surrogate for physical distance on a chromosome. Association mapping, on the other hand, is based on samples from current populations without pedigree structure and uses linkage disequilibrium as a surrogate for physical distance: values for this parameter reflect recombination within the ancestors over many generations of all members of a study sample.

The conventional measure of gametic linkage disequilibrium is analogous to the inbreeding coefficient. If markers **A** and **B** have alleles A, a and B, b then the measure D_{AB} is the deviation of the joint (gametic) frequency of alleles A, B from the product of their separate frequencies: $D_{AB} = P_{AB} - p_A p_B$. It is more convenient to transform this to a correlation

coefficient $\rho_{AB} = D_{AB} / \sqrt{p_A(1-p_A)p_B(1-p_B)}$. The nature of this correlation is easiest seen by introducing indicator variables x, y for alleles A, B . Alleles at markers **A** or **B** are assigned values $x = 1$ or $y = 1$ if they are of type A or B respectively, and are assigned values $x = 0$ or $y = 0$ otherwise. Then D_{AB} is the covariance of x and y for alleles received by an individual from one parent and ρ_{AB} is the correlation of x, y . Similarly, the inbreeding coefficient f_a for marker **A** is the correlation of x values for the two **A** alleles received, one from each parent, by an individual. It may be helpful to introduce D_A and D_B as the covariances of the two **A** alleles and the two **B** alleles, respectively.

A complexity of genetic analyses is that data are collected at the individual level, on genotypes rather than alleles or gametes. This requires an expansion of the two-allele measures of inbreeding and linkage disequilibrium to measures of three and four alleles. If $X = x_1 + x_2$ is the sum of the indicators for the two **A** alleles received by an individual, and Y the sum of the two **B** indicators, then the composite linkage disequilibrium parameter ρ_{AB} is defined to be the covariance of X and Y . The corresponding correlation coefficient is [8]

$$\rho_{AB}^c = \frac{\Delta_{AB}}{\sqrt{[p_A(1-p_A)(1+f_A)] [p_B(1-p_B)(1+f_B)]}}$$

The advantage of this measure is that it can be estimated directly from genotypic data rather than requiring phased data. The estimate r^2 is the squared sample correlation coefficient of the numbers (0,1,2) of A and B alleles carried by an individual. If both markers are in HWE then ρ_{AB}^c reduces to the usual (gametic) correlation coefficient. In Figures 3 and 4 the values for 100,000 random pairs of markers on chromosome 1 of the AMD control data are shown, first for markers anywhere on the chromosome and then for pairs markers within 1 Mb of each other. Only SNPs that were typed in all 50 control individuals were used. Although the linkage disequilibrium statistic r^2 clearly decreases with distance between markers, there is little further decrease after about 200 Kb = 0.2 Mb distance apart. Association mapping depends on significant linkage disequilibrium between marker and disease gene, so

significant association should localize the gene to less than 200 Kb. An approximate test for linkage disequilibrium when n individuals furnish a sample value r is provided by regarding nr^2 as having a chi-square distribution with 1 df. With a 5% significance level, linkage disequilibrium will be declared significantly different from zero when $r^2 > 3.84/n$ or 0.27 in this case.

To complete the linkage disequilibrium picture, the joint frequencies of sets of three or four alleles need to be considered. The probability that an individual is homozygous of allele A at one locus and receives allele B at another locus from one parent, without regard to the allele at that other locus from the second parent, is P_{AAB} . Similarly, P_{ABB} is the frequency of individuals homozygous for allele B at one locus and carrying allele A at another locus. The single-locus homozygote frequency P_{AA} can be written as the sum of two three-allele frequencies: $P_{AA} = P_{AAB} + P_{AAb}$ when only alleles B, b exist at locus **B**. The frequency of double homozygotes $AABB$ is written as P_{AABB} . With complete independence among alleles, these joint frequencies are $P_{AAB} = p_A^2 p_B$, $P_{ABB} = p_A p_B^2$ and $P_{AABB} = p_A^2 p_B^2$. Dependencies are quantified with measures D_{aAb}, D_{aBb} , and D_{aabb} defined by the following expressions:

$$P_{AAB} = p_A^2 p_B + p_A \Delta_{AB} + p_B D_A + D_{AAB}$$

$$P_{ABB} = p_A p_B^2 + p_B \Delta_{AB} + p_A D_B + D_{ABB}$$

$$P_{AABB} = p_A^2 p_B^2 + 2p_A D_{ABB} + 2p_B D_{AAB} + 2p_A p_B \Delta_{AB} + \Delta_{AB}^2 + p_A^2 D_B + p_B^2 D_A + D_A D_B + \Delta_{AABB}$$

The third of these equations is repeated as the first line of Table 2, which also shows the decomposition of all eight genotypic frequencies for markers **A** and **B**.

4 Association Mapping

There is now a large literature on the use of whole-genome marker data for mapping human disease genes. Here only the simple allelic case-control test will be discussed. The counts for the two marker alleles in each of the case and control samples can be arranged in a two by two contingency table and a contingency table chi-square test statistic calculated for testing the hypothesis that marker allele frequencies are the same in cases and controls: i.e. marker alleles are independent of disease status. What are the implications in terms of linkage disequilibrium parameters?

The frequency of marker allele A among cases depends on the unseen disease genotype. If the disease locus has alleles T, t and the probabilities of being affected are G_{TT}, G_{Tt}, G_{tt} for genotypes TT, Tt, tt then

$$\begin{aligned} \Pr(A|\text{Case}) &= \Pr(AA|\text{Case}) \\ &+ \frac{1}{2} \Pr(Aa|\text{Case}) = \frac{1}{\mu_T} \{ [G_{TT} \Pr(AATT) + G_{Tt} \Pr(AATt) + G_{tt} \Pr(AAtt)] + \frac{1}{2} [G_{TT} \Pr(AaTT) + G_{Tt} \Pr(AaTt) + G_{tt} \Pr(Aatt)] \} \end{aligned}$$

where $\mu_T = G_{TT}\Pr(TT) + G_{Tt}\Pr(Tt) + G_{tt}\Pr(tt)$ is the population prevalence of the disease. The expressions of Table 2 for two-locus genotypes can be substituted into this equation and the corresponding equation for $\Pr(A|\text{Control})$. The difference in marker allele frequencies between cases and controls becomes [8]

$$\Pr(A|\text{Case}) - \Pr(A|\text{Control}) = \frac{\Delta_{AT}}{\mu_T(1 - \mu_T)} [p_T(G_{TT} - G_{Tt}) + p_t(G_{Tt} - G_{tt})] + \frac{D_{ATT}}{\mu_T(1 - \mu_T)} [G_{TT} - 2G_{Tt} + G_{tt}]$$

If the disease gene action is additive, $(G_{TT} - 2G_{Tt} + G_{tt}) = 0$, and the case-control test is testing for *composite* linkage disequilibrium between marker and disease alleles. This quantity reflects both gametic linkage disequilibrium and departures from HWE. Because data are collected on genotypes, departures from HWE affect the tests even when the tests are conducted on alleles.

There are several ways of performing the allelic case-control test. If cases and controls have a and b copies, respectively of marker allele A among them and they have c and d copies of the alternative marker allele, then a contingency table test statistic is $n(ad - bc)^2/D$ where $n = a + b + c + d$ and $D = (a + c)(b + d)(a + b)(c + d)$. Under the null hypothesis of no marker-disease association this has a chi-square distribution with 1 df.

The results for the allelic case-control test are shown in Figures 5 and 6 for the AMD chromosome 1 data: first along the chromosome and then as a QQ-plot. In both cases only for those SNPs scored in all 146 individuals were used. The test statistic was assumed to have a chi-square distribution with 1 df and that distribution was used to determine the p -values. The horizontal line in Figure 5[7] is at $-\log_{10}(p) = 6.32$, the Bonferroni cut-off for genome-wide significance when 103,611 SNPs are used (as was the case in [7]). This approach is known to be conservative and it may be better to consider the other SNPs that lie above the expected line of the QQ-plot. The clearly-significant value under either criterion was for the SNP with id rs380390, as was reported in [7]. How much of this test statistic reflects departures from HWE? The plot in Figure 7 shows the HWE test statistics along the chromosome and there are no elevated values at position 193.9893 MB, the location of rs380390. The test seems to be detecting only association between the marker and the disease alleles association although, if alleles at the disease locus do not act additively, it is not known whether the association is for one or two disease alleles with the marker allele (Δ_{AT} of D_{ATT}). The authors of [7] showed that SNP with id rs10272438 on chromosome 7 was also significantly associated with the disease. However this SNP was substantially out of HWE, which suggests the association was spurious. In light of the earlier discussion, it is interesting to note that this SNP had substantial missing data, and when subsequent work restored the missing marker genotypes the disease association was no longer significant.

5 Population Structure

Whole-genome marker data collected from two or more populations also contain information about the history of those populations. It has recently become clear that attention needs to be paid to missing data, just as has been shown above for HWE and association testing. The classic approach to interpreting data from more than one population has been to calculate “ F -statistics”, particularly the quantity F_{ST} that makes statements about pairs of alleles within populations relative to pairs of alleles between populations. The work of [9] and [10] is relevant.

Suppose that samples of n_i alleles are available from a series of populations i , $i = 1, 2 \dots r$. The sample frequencies of minor alleles A in these samples are written as p_i , all of which

have expected value p_A , the frequency in the overall population. The F -statistics within (i) and between (ij) populations allow the expression of variances and covariances of sample allele frequencies as follows [10].

– Variance of sample allele frequencies within populations:

$$\text{Var}(\tilde{p}_i) = p_A(1 - p_A) \left[\theta_i + \frac{1 - \theta_i}{n_i} \right]$$

– Covariances of sample frequencies in different populations:

$$\text{Cov}(\tilde{p}_i, \tilde{p}_j) = p_A(1 - p_A) \theta_{ij}$$

The heterozygosities within sample i and between sample pair i, j are

$$\tilde{H}_i = \frac{2n_i}{n_i - 1} \tilde{p}_i(1 - \tilde{p}_i)$$

$$\tilde{H}_{ij} = \tilde{p}_i(1 - \tilde{p}_j) + (1 - \tilde{p}_i)\tilde{p}_j$$

and these are used to construct estimates of the θ_i and θ_{ij} relative to the average of the pairwise values $\theta_A = \frac{1}{r(r-1)} \sum_{i,j} \theta_{ij}$. The estimates and their expected values are

$$\hat{\beta}_i = 1 - \frac{\tilde{H}_i}{\tilde{H}_A}, \varepsilon(\hat{\beta}_i) = \beta_i = \frac{\theta_i - \theta_A}{1 - \theta_A}$$

$$\hat{\beta}_w = 1 - \frac{\tilde{H}_w}{\tilde{H}_A}, \varepsilon(\hat{\beta}_w) = \beta_w = \frac{\theta_w - \theta_A}{1 - \theta_A}$$

$$\hat{\beta}_{ij} = 1 - \frac{\tilde{H}_{ij}}{\tilde{H}_A}, \varepsilon(\hat{\beta}_{ij}) = \beta_{ij} = \frac{\theta_{ij} - \theta_A}{1 - \theta_A}$$

where θ_w is the average of the single values $\theta_i = \sum_j \theta_{ij} / (r-1)$.

The approach of [9] used sample sizes as weights under the assumption that all populations i had the same value θ_i and that all pairs of populations were independent $\theta_{ij} = 0$. The resulting estimate $\hat{\beta}_{WC}$ was for θ_A for all sample sizes. If the assumptions about θ_i, θ_{ij} are relaxed, however, it may be better to use unweighted estimates.

For two populations, $r=2$, there are three sample heterozygosities H_1, H_2 and H_{12} as above and two population-specific estimates β_1, β_2 of $(\theta_1 - \theta_2)/(1 - \theta_2)$ and $(\theta_2 - \theta_1)/(1 - \theta_1)$. Each is proportional to the time since that population diverged from the population ancestral to the two current populations, assuming that divergence is due to genetic drift. Their

average is for $[(1 + 2)/2 - 12]/(1 - 12)$ whereas the traditional estimate w_C is for $[(1 + 2)/2 - 12]/(1 + Q - 12)$ where $Q = -(1 - 2)(n_1 - n_2)/(n_1 + n_2)$. The term Q is zero if either $n_1 = n_2$ or $1 = 2$.

The consequences of using the weighted estimates of [9] or unweighted estimates $w = (1 + 2)/2$ are shown in Figure 8 for the 11 populations in the HapMapIII data (www.sanger.ac.uk/resources/downloads/human/hapmap3.html), described in Table 3. Only SNPs on chromosome 1 were used here. The two sets of estimates are indistinguishable for 45 of the 55 pairs of populations. The remaining 10 pairs all involved the JPT population, for which over 1,000 of the SNPs were typed for only two individuals. This severe imbalance led to the differences shown in Figure 8.

6 Discussion

This is an exciting time for statisticians working with genetic data. They now have access to virtually all the genetic data there are. This discussion has looked at SNP data which, although substantial, are but part of the whole-genome sequence data that are emerging. Along with the opportunities offered by large data sets come the challenges of ensuring the data are of high quality and that statistical issues such as those surrounding multiple testing are addressed. Some data are shown here that provide support for an extreme level of data filtering: discarding any SNP that is not typed in all individuals. Work is ongoing to determine to what extent this stringency can be relaxed.

Many topics of interest and importance have not been discussed here. Genetic structuring of populations can have profound effects on many analyses, for example. A population consisting of several subpopulations, all in HWE, will itself be out of HWE to the extent that allele frequencies vary among sub-populations and this can lead to rejection of the HWE hypothesis from data sampled from the whole population. Varying allele frequencies can also cause hypotheses of linkage equilibrium to be rejected from whole-population samples even if there is linkage equilibrium within each subpopulation. If disease status is confounded with population structure, there can be false indications of marker-disease association if population structure is ignored. Care is also needed for admixed populations where individuals may have ancestors on two or more different ancestral populations, as is the case for the African-American population for example.

The high dimensionality of current whole-genome data suggest that the single-SNP analyses described here be replaced by analyses on regions rather than single SNPs. These regions may need to be defined empirically from levels of linkage disequilibrium or shared ancestry but they may allow some smoothing out of the noisy patterns such as that in Figure 4, and of course they would reduce dimensionality.

Acknowledgments

This work was supported in part by NIH Grant GM 07591

References

1. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs K, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, Zheng X, Weir BS. GENEVA Investigators. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*. 2010; 34:591–602. [PubMed: 20718045]
2. Hardy GH. Mendelian proportions in a mixed population. *Science*. 1908; 28:49–50. [PubMed: 17779291]

3. Weinberg W. über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg. 1908; 64:368–382.
4. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet.* 2005; 76:887–893. [PubMed: 15789306]
5. Rohlfs R, Weir BS. Distributions of Hardy-Weinberg equilibrium test statistics. *Genetics.* 2008; 180:1609–1616. [PubMed: 18791257]
6. Haldane JBS. An exact test for randomness of mating. *Journal of Genetics.* 1954; 8:299–309.
7. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. *Science.* 2005; 308:385–389. [PubMed: 15761122]
8. Weir BS. Linkage disequilibrium and association mapping. *Annual Reviews of Genomics and Human Genetics.* 2008; 9:129–142.
9. Weir BS, Cockerham CC. Estimating F^2 -statistics for the analysis of population structure. *Evolution.* 1984; 38:1358–1370.
10. Weir BS, Hill WG. Estimating F -statistics. *Annual Review of Genetics.* 2002; 36:721–750.

Fig. 1.
QQ-plot for HWE exact test on AMD Control data: no SNP filtering.

HWE Test: SNPs Filtered on Missingness

Fig. 2.
QQ-plot for HWE exact test on AMD Control data: SNPs filtered on missingness.

AMD Control Linkage Disequilibrium

Fig. 3.
Plot of linkage disequilibrium along Chromosome 1 for the AMD Control data.

AMD Control Linkage Disequilibrium

Fig. 4.
Plot of linkage disequilibrium along Chromosome 1 for the AMD Control data.

AMD Case Control Test

Fig. 5.
Plot of case-control test statistic along Chromosome 1 for the AMD data.

Fig. 6.
QQ-plot of case-control test statistics for the AMD Chromosome 1 data.

Control HWE Test: With SNP Filtering

Fig. 7.
Plot of HWE test statistics along Chromosome 1 for the AMD data.

HapMap Fst estimates: no SNP filtering

Fig. 8.
 F_{ST} estimates for HapMap III, using all 87,592 SNPs on chromosome 1.

Table 1

Heterozygote probabilities conditional on allele counts

f	0.10	0.05	0.01	0.00	-0.01	-0.05	-0.10
	1.61	1.79	1.96	2.00	2.04	2.24	2.51
n_{Aa}	$Pr(n_{Aa} n_A)$						
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
11	0.0029	0.0009	0.0004	0.0003	0.0002	0.0001	0.0000
13	0.0154	0.0063	0.0028	0.0023	0.0018	0.0007	0.0002
15	0.0544	0.0277	0.0147	0.0124	0.0103	0.0048	0.0016
17	0.1298	0.0820	0.0517	0.0455	0.0397	0.0219	0.0091
19	0.2125	0.1665	0.1250	0.1148	0.1049	0.0692	0.0360
21	0.2414	0.2346	0.2096	0.2012	0.1922	0.1515	0.0992
23	0.1905	0.2297	0.2443	0.2450	0.2445	0.2304	0.1896
25	0.1037	0.1552	0.1964	0.2058	0.2146	0.2417	0.2502
27	0.0383	0.0711	0.1071	0.1173	0.1278	0.1720	0.2239
29	0.0093	0.0214	0.0384	0.0439	0.0500	0.0804	0.1316
31	0.0014	0.0040	0.0085	0.0102	0.0121	0.0233	0.0480
33	0.0001	0.0004	0.0011	0.0013	0.0016	0.0038	0.0097
35	0.0000	0.0000	0.0001	0.0001	0.0001	0.0003	0.0008
Power	0.0745	0.0394	0.0276	0.0266	0.0261	0.0330	0.0603

Table 2
Frequencies of two-locus genotypes allowing for Hardy-Weinberg disequilibrium ^{*}.

Genotype	1	2D _{ABB}	2D _{AAB}	2 ^{AB}	Δ ² _{AB}	D _B	D _A	D _A D _B	AABB
AABB	$p_A^2 p_B^2$	p_A	p_B	$p_A p_B$	1	p_A^2	p_B^2	1	1
AABb	$2p_A^2 \pi_B$	$-2p_A$	B	$p_A B$	-2	$-2p_A^2$	$2B$	-2	-2
AAbb	$p_A^2 p_b^2$	p_A	$-p_b$	$-p_A p_b$	1	p_A^2	p_b^2	1	1
AaBB	$2\pi_A p_B^2$	A	$-2p_B$	$-A p_B$	-2	$2A$	$2p_B^2$	-2	-2
AaBb	$4A B$	$-2A$	$-2B$	$A B$	4	$-4A$	$-4B$	4	4
Aabb	$2\pi_A p_b^2$	A	$2p_b$	$-A p_b$	-2	$2A$	$-2p_b^2$	-2	-2
aaBB	$p_a^2 p_B^2$	$-p_a$	p_B	$-p_a p_B$	1	p_a^2	p_B^2	1	1
aaBb	$2p_a^2 \pi_B$	$2p_a$	B	$-p_a B$	-2	$-2p_a^2$	$2B$	-2	-2
aabb	$p_a^2 p_b^2$	$-p_a$	$-p_b$	$p_a p_b$	1	p_a^2	p_b^2	1	1

^{*} $A = p_A p_a$, $B = p_B p_b$, $A = (p_a - p_A)$, $B = (p_b - p_B)$

Table 3

HapMap III Populations.

Code	Population Description	Sample size	SNPs typed
ASW	African ancestry in Southwest USA	142	61566
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection	324	56218
CHB	Han Chinese in Beijing, China	160	51946
CHD	Chinese in Metropolitan Denver, Colorado	140	50517
GIH	Gujarati Indians in Houston, Texas	166	55710
JPT	Japanese in Tokyo, Japan	168	53020
LWK	Luhya in Webuye, Kenya	166	60026
MXL	Mexican ancestry in Los Angeles, California	142	57937
MKK	Maasai in Kinyawa, Kenya	342	61057
TSI	Toscani in Italia	154	55881
YRI	Yoruba in Ibadan, Nigeria	326	58559