10 Meta-Analysis of Genome-Wide Association Studies

Paul I.W. de Bakker,[1,2] Benjamin M. Neale,[2,3] and Mark J. Daly[2,3]

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115; [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142; [3]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114

INTRODUCTION

Individual genome-wide association studies (GWAS) have only limited power to find novel loci underlying complex traits and common diseases. With relatively modest sample and effect sizes, a true association between genotype and phenotype may never meet genome-wide statistical significance ($p < 5 \times 10^{-8}$) in a single study. Through meta-analysis, novel susceptibility loci can be discovered by effectively summing the statistical evidence of individually underpowered studies. Most genetic discoveries for complex traits are now made through meta-analysis collaborations (Barrett et al. 2008; Ferreira et al. 2008; Zeggini et al. 2008; Kathiresan et al. 2009; Newton-Cheh et al. 2009). These efforts so far have been restricted to single-locus analyses, testing for main effects at a single polymorphism at a time. A key benefit of this approach is that individual-level genotype (and phenotype) data do not need to be exchanged between research groups (which in practice can be a genuine obstacle). In this chapter, we focus only on meta-analysis at single single-nucleotide polymorphisms (SNPs), paying particular attention to how imputation uncertainty can be incorporated into the association analysis and subsequent meta-analysis.

Probably the most important aspect of genome-wide association meta-analysis is harmonization of the study results (de Bakker et al. 2008). Not surprisingly, studies differ in design, sample collection, genotyping platforms, association analysis methods, and so forth. The goal for meta-analysis is that the association results (per SNP) of each study can be formatted, exchanged, and analyzed in such a way that the statistical evidence can be combined appropriately and that no valuable information is lost. Without minimizing the importance of having a clear phenotype definition (and corresponding measurements), we will assume for the sake of brevity that investigators representing the various studies have made sensible agreements about phenotype definitions, necessary sample exclusions, and appropriate covariate modeling.

IMPUTATION FILLS IN MISSING GENOTYPE DATA

Genotyping platforms differ in terms of their SNP content (and more recently, also in terms of copy-number polymorphism [CNP] content). Thus, combining disparate data sets from different studies would seem to be a difficult task. Fortunately, linkage disequilibrium (LD) among nearby variants

makes it possible to predict (or "impute") polymorphisms that are not directly genotyped. As described elsewhere, the International HapMap Project (http://www.hapmap.org) has genotyped >3 million SNPs across the human genome in 270 DNA samples from four populations, providing 120 phased haplotypes for Utah residents with ancestry from northern and western Europe (CEU), 120 for Yoruba in Ibadan, Nigeria (YRI), and another 180 for Han Chinese from Beijing, China (CHB) and Japanese from Tokyo, Japan (JPT) (International HapMap Consortium 2007). A number of software tools are now available for imputation, including IMPUTE (Marchini et al. 2007), MACH (Li and Abecasis 2006), PLINK (Purcell et al. 2007), BIMBAM (Servin and Stephens 2007), and BEAGLE (Browning and Browning 2009). These programs require as input the genotype data collected in a sample and the HapMap genotypes (or haplotypes) as reference data and generate genotypes for all SNPs present on HapMap as output. Thus, the imputation procedure fills in the missing genotype data, effectively allowing data sets to be analyzed for a common set of SNPs.

Imputation Accuracy and Quality

Imputation accuracy is limited by two factors. First, the genotyping platform affects the imputation accuracy, because the SNP content directly determines the effective genome-wide coverage of variation. Second, imputation accuracy is limited by the SNP density and sample size of the reference panel (i.e., HapMap). By design, HapMap is strongly biased toward common variation (with good coverage of alleles with frequency >5%). Consequently, SNP genotyping arrays based on HapMap provide good coverage of common SNPs (Pe'er et al. 2006). In contrast, representation of rare variants (alleles with frequency <5%) is much less complete, because only a few haplotypes are observed per minor allele for those rare SNPs covered in HapMap. Therefore, the expectation is that the prediction accuracy for rare alleles will be worse than for common variation when using HapMap as the reference data set. Importantly, the accuracy of the (imputed) genotypes for a given SNP is likely to vary between studies, given that some will have directly genotyped it while others imputed it with varying degrees of success.

To illustrate some of these effects on imputation quality, we explore here data from the Diabetes Genetics Initiative (http://www.broad.mit.edu/diabetes). In that study, 1464 cases (affected with type 2 diabetes) and 1467 controls (matched for age, gender, BMI, and geographic locale) were genotyped on the Affymetrix 500K platform (Saxena et al. 2007). Figure 10.1 shows a pie chart of all HapMap SNPs split into four bins according to the pairwise r^2 between each SNP and any of the genotyped SNPs on the 500K array. Approximately 46% are perfectly captured by either genotyping or perfect LD. About one quarter of all HapMap SNPs have weaker LD (pairwise $r^2 < 0.5$) and are thus more difficult to impute. Table 10.1 shows the number of SNPs in these four bins split by their minor allele frequency. Of the low-frequency SNPs, most (58%) are poorly captured by the genotyped SNPs. Generally, high-frequency SNPs are better captured than low-frequency SNPs,

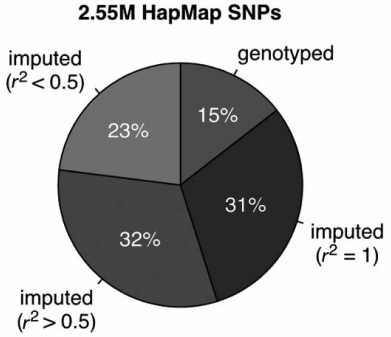

FIGURE 10.1 Breakdown of 2.5 million polymorphic SNPs in HapMap-CEU (release 21) by the pairwise correlation (r^2) to any of the genotyped SNPs on the Affymetrix 500K array. Almost two thirds of all HapMap SNPs can be imputed relatively straightforwardly as pairwise LD is strong: 31% of all HapMap SNPs are perfectly correlated ($r^2=1$), and another 32% have $r^2>0.5$ to a genotyped SNP; 23% of HapMap SNPs are more difficult to capture and likely will require haplotype information to be imputed accurately. (Based on data produced by the Diabetes Genetics Initiative [Saxena et al. 2007].)

TABLE 10.1 Number of SNPs by minor allele frequency split by pairwise r^2 to any of the genotyped SNPs on the Affymetrix platform

	Minor allele frequency (MAF)		
	Low frequency (MAF <5%)	Intermediate frequency (MAF 5%–20%)	High frequency (MAF 20%–50%)
r^2	[MAF<5%]	[MAF 5–20%]	[MAF 20–50%]
$r^2 = 1$	112,153 (33%)	291,171 (40%)	379,247 (34%)
$r^2 > 0.5$	33,724 (10%)	249,031 (34%)	530,915 (48%)
$r^2 < 0.5$	198,368 (58%)	194,498 (26%)	195,753 (18%)
Total	344,245 (100%)	734,700 (100%)	1,105,915 (100%)

consistent with previous observations that common SNPs in HapMap CEU tend to have more proxies (International HapMap Consortium 2005).

Starting with all quality control (QC)-passing SNPs from the 500K array (~380K SNPs) as input genotypes, we used the MACH program to impute all SNPs on HapMap CEU (release 21). As output, MACH computes the dosage (the estimated number of minor alleles per individual, ranging between 0 and 2) for every SNP in each individual (output in mldose files). The dosage is based on the posterior probabilities for each of the three genotype possibilities (AA, AB, BB) in each individual (output in mlprob files):

$$\text{dosage} = 1 \times p(\text{AB}) + 2 \times p(\text{BB}) , \tag{1}$$

where p(AB) and p(BB) are the posterior probabilities of the heterozygote (AB) and minor homozygote (BB), respectively.

After the imputations are done, MACH computes the average maximal posterior probability (called "Quality") for each SNP averaged over the entire sample and estimates the correlation (called "Rsq") between the imputed genotype and the actual genotype (which can be interpreted as a measure of the imputation uncertainty). This Rsq metric is equivalent to the ratio between the observed variance of the dosage and the expected (binomial) variance of the dosage. A low observed/expected dosage variance ratio would indicate poor imputation accuracy, whereas accurate genotypes should approach unity (fluctuation around 1 due to Hardy–Weinberg deviations). In our example, we find that common SNPs are generally imputed well, with a clear decrease in imputation accuracy for low-frequency SNPs (as reflected by the heavy tail toward low observed/expected dosage variance ratios in Fig. 10.2A).

For the analyses presented here, we excluded 326 discordant sibships and kept only unrelated cases and matched controls, thus allowing us to use the 1-degree-of-freedom chi-square test for association:

$$\chi^2 = \frac{\left(p_{\text{case}} - p_{\text{control}}\right)^2}{\left(\frac{1}{n_{\text{case}}} + \frac{1}{n_{\text{control}}}\right)\left(p\left(1-p\right)\right)} , \tag{2}$$

where p_{case} and p_{control} are the minor allele frequency of a given SNP in cases and controls, respectively, p is the combined allele frequency, and n_{case} and n_{control} are the number of chromosomes in cases and controls, respectively.

Chi-Squared Correction

Although the genomic inflation factor λ (described elsewhere) for the genotyped SNPs in our analysis was modest (1.04), we observed that the test statistic was conservative, resulting in a substantial

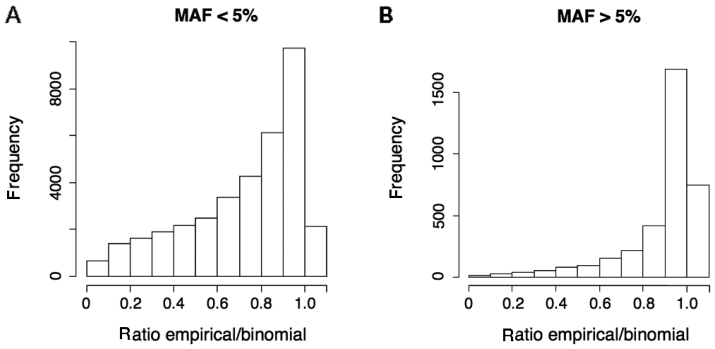

FIGURE 10.2 Histogram of the observed/expected dosage variance ratios for (*A*) rare SNPs (MAF <5%) and (*B*) common SNPs (MAF >5%). (MAF) Minor allele frequency.

deflation of the distribution for poorly imputed SNPs (Fig. 10.3). This illustrates that the χ^2 test statistic based on raw dosages (when uncertain) does not behave properly under the null hypothesis. We propose here a simple correction for this conservative behavior. With increasing imputation uncertainty, the variance of the dosage decreases due to lack of information. To correct for the uncertainty, we need to appropriately reduce the variance term in the denominator of the χ^2 formulation above. We can achieve this by replacing the binomial variance term $p(1 - p)$ with the empirically observed variance. (To compute the empirically observed variance, we calculate the mean dosage for the total sample and sum the square deviations.)

We validate this statistical correction for a subset of poorly imputed SNPs in our example to mimic a worst-case scenario. Figure 10.3 shows the quantile–quantile (Q-Q) plot for 198,368 SNPs of frequency <5% and pairwise $r^2 < 0.5$ to any of the genotyped SNPs. Before correction, we observe a strongly deflated distribution of the test statistic. After correcting for the variance deflation, we recover a test statistic distribution consistent with the null distribution (along the diagonal of the Q-Q plot).

To further explore the impact of imputation accuracy on the association analysis, a subset of common (~8000) and rare (~36,000) genotyped SNPs was hidden from the imputation. For these SNP sets, we compared the test statistic based on the "true" genotype to the corrected test statistic based on the imputed dosage. Figure 10.4 shows the distribution of the respective *Z* scores (we converted the chi-square into a *Z* score for the sake of visualization). There is a clear positive trend

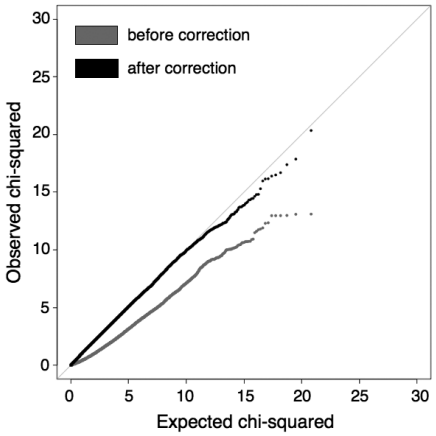

FIGURE 10.3 Quantile–quantile plot for the test statistic distribution of the chi-square test for SNPs of frequency <5% and pairwise r^2 to any of the genotyped SNPs before and after correction for the imputation uncertainty. The uncorrected distribution is deflated relative to the expected null distribution. The proposed correction is able to recover a proper null distribution along the diagonal.

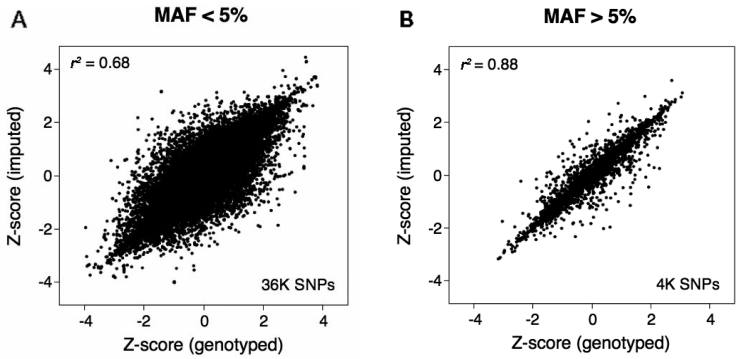

FIGURE 10.4 Comparison of the test statistic (as a *Z* score) between imputed genotypes and experimentally observed genotypes for randomly selected SNPs: (*A*) Minor allele frequency (MAF) < 0.05 (n = 36,000) and (*B*) MAF > 5% (n = 8,000). The lower correlation coefficient indicates overall lower imputation accuracy for the low-frequency SNPs. (Based on data produced by the Diabetes Genetics Initiative [Saxena et al. 2007].)

between the correlation of the *Z* scores and the allele frequency. The Pearson correlation between the *Z* scores is 0.9 for high-frequency SNPs and 0.7 for low-frequency SNPs. The correlation drops progressively with lower frequency, with a $r^2 = 0.3$ for SNPs with frequency <1%. So far, we have demonstrated that it is possible to obtain proper test statistics from imputation dosages by using a simple correction to the chi-squared test and that, as expected, the performance for rare variant imputation is worse than for common SNPs.

Incorporating Imputation Uncertainty into Meta-Analysis

The depression in the variance of the test statistic not only underestimates the true significance, but also causes further problems in the context of meta-analysis. Combining uncorrected test statistics causes an initial loss in power, but the weighting of a study is also typically proportional to its sample size or the inverse of the variance (of the effect). By not correcting the sample size, we allow for too great a contribution of an uncertain genotype. Therefore, an important consideration is how imputation uncertainty is incorporated into the meta-analysis (de Bakker et al. 2008). No single study should be able to distort (or disproportionately contribute to) the meta-analysis. Only when these conditions are met, can the association information be safely combined across multiple studies in the meta-analysis.

PERFORMING THE META-ANALYSIS

Having ensured that the test statistic distribution looks good, we are ready to do the actual meta-analysis. Probably the simplest approach is to sum the association evidence across studies where every study is weighted by the effective sample size. First, we convert *p* values to *Z* scores as follows (in R code):

```
#
# this is the routine to convert a pvalue into a zscore.
#
# the direction of the odds ratio (or) determines the sign of
# the resulting z-score
#
convert.pvalue <- function(pval, or) {
```

```
 if ( or > 1 ) {
 z <- qnorm( pval / 2 );
 } else {
 z <- -(qnorm( pval / 2 ));
 }

 return(z);
}
```

The resulting Z scores can then be summed weighted by the sample size, as follows:

$$z_{\text{meta}} = \sum z_i \times w_i \,, \tag{3}$$

where $w_i = \sqrt{N_i / N_{\text{total}}}$ and N_i and N_{total} are the study sample size and the total sample size, respectively.

Often, studies have many more controls than cases, and power will be primarily limited by the number of cases. (The power improvement for including additional controls saturates after adding three to four times the number of cases.) In these cases, we need to compute an "effective" sample size (as weighting a study by its total sample size would overestimate its true contribution). One possible way to do this is to compute the noncentrality parameter (NCP) for a given disease model (including case and control numbers) and use the resulting NCP as the N_i (see Appendix A).

To take into account imputation uncertainty, we can scale the sample size by the observed/expected variance ratio (or a similar information metric output by the imputation program). As imputation quality varies per SNP, this must be done on a per SNP basis. Because there is random fluctuation of the observed/expected variance ratio (even for accurate genotype calls), we should only use the test statistic correction (if using the χ^2 test for a binary trait above) and sample size scaling when the average maximal posterior probability is below some threshold (e.g., 0.99).

The resulting meta-analysis Z scores can be converted back to p values as follows (R code):

```
#
# to convert the meta-analytic z-scores into p-values
#
pmeta <- pnorm(-(abs(zmeta))) * 2;
```

Instead of meta-analysis by sample size weighting, we can perform meta-analysis by weighting by the inverse variance for linear and logistic regression association results. This is relatively straightforward. First, we compute the weighted β and standard error (SE) terms based on the estimated β coefficient and corresponding SE that have been calculated for each study:

$$\langle \beta \rangle = \frac{\sum_i \left[\beta_i / (\text{SE}_i)^2 \right]}{\sum_i \left[1 / (\text{SE}_i)^2 \right]} \tag{4}$$

and

$$\langle \text{SE} \rangle = \sqrt{1 / \sum_i \left[1 / (\text{SE})^2 \right]} \,. \tag{5}$$

From these two weighted terms, we can compute a meta-analysis Z score:

$$z = \langle \beta \rangle / \langle \text{SE} \rangle \,. \tag{6}$$

Using linear or logistic regression as the primary association analysis approach has a key advantage. Although count-based test statistics (such as the χ^2 test) require explicit corrections (as we pro-

posed above), linear and logistic regression models can implicitly deal with imputation uncertainty as the variance of the allele (or dosage) frequency becomes deflated with growing uncertainty. This variance deflation is automatically absorbed by regression modeling by generating a large(r) standard error of the estimated β coefficient. This means that, in principle, the genome-wide distribution of the test statistic by linear or logistic regression modeling is overall well behaved. For other methods that more explicitly deal with imputation uncertainty, see, for example, Guan and Stephens (2008).

Conventional meta-analysis literature advocates the preferred use of random-effects models (rather than fixed-effects models). For genetic discovery, the main focus is on finding novel associated loci, not on accurate estimation of the effect size. A random-effects meta-analysis effectively penalizes the test statistic for observed heterogeneity between studies, thus lowering power for discovery. This is why all genome-wide meta-analysis efforts to date are based on a fixed-effects (and not random-effects) model. However, we note that there are certainly situations where heterogeneity can be informative (especially when the number of contributing studies is large).

The 1000 Genomes Project (http://www.1000genomes.org) is an ongoing effort to build a genome-wide inventory of all segregating sequence variation down to a frequency of 1% (0.1% in genic regions) through sequencing in large population samples and to enable imputation-based approaches for rare variants. We expect that imputation-based meta-analysis will continue to have an important role in genetic discovery as more complete resources such as 1000 Genomes are being developed.

Finally, in support of this chapter, we have made available a meta-analysis example based on the genome-wide association results of the type 2 diabetes scan performed by the Wellcome Trust Case Control Consortium and the Diabetes Genetics Initiative (http://www.broad.mit.edu/~debakker/meta_t2d.html). R scripts and Perl code that were developed for and used for a meta-analysis of electrocardiographic QT interval duration (Newton-Cheh et al. 2009) can be found on the same web page.

APPENDIX A

The routine below (in R code) computes the noncentrality parameter (NCP) for a given disease model, assuming risk allele frequency, relative risks for the heterozygote and major homozygote, and sample size.

```
#
# this routine computes NCP given the following parameters:
#
# fA = risk allele frequency
# k = population prevalence of trait
# rAa = relative risk of genotype Aa
# rAA = relative risk of genotype AA
# n_case = number of cases
# n_control = number of controls
#
cc_gpc <- function( fA, k, rAa, rAA, n_case, n_control ) {
    # frequency of non-risk allele
    fa <- 1 - fA

    # calculate genotype frequencies
    fAA <- fA * fA
    fAa <- 2 * fA * fa
    faa <- fa * fa

    # baseline risk of genotype aa
```

```
    raa <- k / (fAA*rAA + fAa*rAa + faa)

    # risk of genotypes
    rrAA <- rAA * raa
    rrAa <- rAa * raa
    rraa <- raa

    nrrAA <- 1 - rrAA
    nrrAa <- 1 - rrAa
    nrraa <- 1 - rraa

    # compute odds ratio
    orAa <- ( rrAa / nrrAa ) / ( rraa / nrraa )
    orAA <- ( rrAA / nrrAA ) / ( rraa / nrraa )

    # genotype frequencies in cases
    case_AA <- fAA * rrAA
    case_Aa <- fAa * rrAa
    case_aa <- faa * rraa
    case_sum <- case_AA + case_Aa + case_aa

    case_AA <- case_AA / case_sum
    case_Aa <- case_Aa / case_sum
    case_aa <- case_aa / case_sum

    # genotype frequencies in controls
    control_AA <- fAA * nrrAA
    control_Aa <- fAa * nrrAa
    control_aa <- faa * nrraa
    control_sum <- control_AA + control_Aa + control_aa

    control_AA <- control_AA / control_sum
    control_Aa <- control_Aa / control_sum
    control_aa <- control_aa / control_sum

    # allele frequencies in cases and controls
    case_A <- case_AA + case_Aa / 2
    control_A <- control_AA + control_Aa / 2

    # turn into case-control counts
    n_case_A <- 2 * n_case * case_A
    n_case_a <- 2 * n_case * (1 - case_A)
    n_control_A <- 2 * n_control * control_A
    n_control_a <- 2 * n_control * (1 - control_A)

    # compute 2x2 chi-square test for association
    x2 <- chisq.test(matrix(c(n_case_A, n_control_A,
                              n_case_a, n_control_a),
                            nrow=2, ncol=2),
                     correct=F)
    # NCP is the chi-square statistic
    ncp <- x2$statistic

    # return frequency of A allele in cases and controls, and NCP
    c(case_A, control_A, ncp)
}
```

REFERENCES

Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., et al. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40:** 955–962.

Browning, B.L. and Browning, S.R. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84:** 210–223.

de Bakker, P.I., Ferreira, M.A., Jai, X., Neale, B.M., Raychauduri, S., and Voight, B.F. 2008. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17:** R122–R128.

Ferreira, M.A., O'Donovan, M.C., Meng, Y.A., Jones, I.R,. Ruderfer, D.M., et al. 2008. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat. Genet*. **40:** 1056–1058.

Guan, Y. and Stephens, M. 2008. Practical issues in imputation-based association mapping. *PLoS Genet.* **4:** e1000279.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851–861.

Kathiresan, S.,Willer, C.J., Peloso, G.M., Demissie, S., Musunuru, K., et al. 2009. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41:** 56–65.

Li, Y. and Abecasis, G.R. 2006. MACH 1.0: Rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet.* **S79:** 2290.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39:** 906–913.

Newton-Cheh, C., Eijgelsheim, M., Rice, K.M., de Bakker, P.I., Yin, X., et al. 2009. Common variants at ten loci influence QT interval duration in the QTGEN Study. *Nat. Genet.* **41:** 399–406.

Pe'er, I., de Bakker, P.I.W., Maller, J., Yelensjy, R., Altschuler, D., and Daly, M.J. 2006. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.* **38:** 663–667.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81:** 559–575.

Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I., et al. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316:** 1331–1336.

Servin, B. and Stephens, M. 2007. Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet.* **3:** e114.

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., et al. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40:** 638–645.

WEB RESOURCES

http://www.1000genomes.org 1000 Genomes Project.
www.broad.mit.edu/diabetes Diabetes Genetics Initiative.
http://debakker.med.harvard.edu/ de Bakker lab page.
http://www.hapmap.org HapMap Project.
http://pngu.mgh.harvard.edu/~purcell/plink/gplink.shtml Purcell et al. 2007. PLINK.
http://quartus.uchicago.edu/~yguan/bimbam/index. html Servin and Stephens 2007. BIMBAM.
http://www.sph.umich.edu/csg/abecasis/MACH/download/ Li and Abecasis 2006. MACH.
http://www.stat.auckland.ac.nz/~bbrowning/beagle/beagle.html Browning and Browning 2009. BEAGLE.
http://www.stats.ox.ac.uk/~marchini/software/gwas/impute. html Marchini et al. 2007. IMPUTE.