

**ARTICLE**

# ATRIUM: Testing Untyped SNPs in Case-Control Association Studies with Related Individuals

Zuoheng Wang<sup>1,3</sup> and Mary Sara McPeek<sup>1,2,\*</sup>

In genome-wide association studies, only a subset of all genomic variants are typed by current, high-throughput, SNP-genotyping platforms. However, many of the untyped variants can be well predicted from typed variants, with linkage disequilibrium (LD) information among typed and untyped variants available from an external reference panel such as HapMap. Incorporation of such external information can allow one to perform tests of association between untyped variants and phenotype, thereby making more efficient use of the available genotype data. When related individuals are included in case-control samples, the dependence among their genotypes must be properly addressed for valid association testing. In the context of testing untyped variants, an additional analytical challenge is that the dependence, across related individuals, of the partial information on untyped-SNP genotypes must also be assessed and incorporated into the analysis for valid inference. We address this challenge with ATRIUM, a method for case-control association testing with untyped SNPs, based on genome screen data in samples in which some individuals are related. ATRIUM uses LD information from an external reference panel to specify a one-degree-of-freedom test of association with an untyped SNP. It properly accounts for dependence in the partial information on untyped-SNP genotypes across related individuals. We demonstrate that ATRIUM is robust in that it maintains the nominal type I error rate even when the external reference panel is not well matched to the case-control sample. We apply the method to detect association between type 2 diabetes and variants on chromosome 10 in the Framingham SHARe data.

## Introduction

With the rapid advances in high-throughput genotyping technology, genome-wide association studies have become a viable approach to elucidating the genetic basis of human complex disease. It is now affordable to analyze on the order of  $10^5$  to  $10^6$  markers throughout the genome. However, because the set of single-nucleotide polymorphisms (SNPs) assayed by the current genotyping platforms covers only a fraction of the total variation in the human genome, it is likely that many disease-susceptibility alleles are not directly genotyped. Therefore, it is of great interest to develop powerful statistical methods to detect association with untyped causal variants. To do this, one can use the linkage disequilibrium (LD) structure of the genome, together with data on typed variants, to detect association between untyped variants and phenotype.

In the context of unrelated individuals, several approaches have been developed, including imputation approaches,<sup>1–8</sup> a likelihood-based method,<sup>9</sup> testing of tag SNPs,<sup>10</sup> and testing of haplotypes of tag SNPs,<sup>11</sup> as well as TUNA<sup>12,13</sup> and related approaches<sup>14</sup> that contrast estimated allele frequencies for cases and controls, where the estimates are based on a linear combination of haplotype frequencies of tag SNPs. A recent extension of BEAGLE<sup>8</sup> allows imputation of genotypes for parent-offspring pairs and trios as well as for unrelated individuals. This is an improvement over the approach of applying, to related individuals, imputation methods that were designed for unrelated individuals, because it allows one to use addi-

tional phase information from relatives and it avoids the introduction of Mendelian errors.

A key concern for case-control association testing with related individuals is that imputed genotypes are dependent among relatives, where the dependence among imputed genotypes differs from the ordinary dependence among genotypes and is affected by the type and amount of information available for each individual. This complex dependence among imputed genotypes would need to be taken into account in the analysis in order to construct a valid test. However, to our knowledge, the current generation of imputation methods gives information only on the marginal accuracy (e.g., marginal posterior probabilities and not joint posterior probabilities) of imputed genotypes, so these methods would not allow valid assessment of uncertainty in the general setting of case-control association testing with related individuals.

A few methods of case-control association testing that allow arbitrary combinations of related and unrelated individuals have been developed, including methods to detect association with a typed marker<sup>15–17</sup> and methods to detect haplotype association.<sup>18–20</sup> In this article, we propose the Association Test with Related Individuals for Untyped Markers, or ATRIUM. ATRIUM is a one-degree-of-freedom (1-df) association test based on genotype data from multiple typed SNPs that are in LD with the untyped SNP, where information on the joint distribution of typed and untyped SNPs is obtained from an external reference panel, and where the sample can include family members as well as unrelated individuals. ATRIUM properly accounts for dependence in the partial information on untyped

<sup>1</sup>Department of Statistics, <sup>2</sup>Department of Human Genetics, The University of Chicago, Chicago, IL 60637, USA

<sup>3</sup>Present address: Division of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA

\*Correspondence: [mcpeek@galton.uchicago.edu](mailto:mcpeek@galton.uchicago.edu)

DOI 10.1016/j.ajhg.2009.10.006. ©2009 by The American Society of Human Genetics. All rights reserved.

SNP genotypes across related individuals. Because we condition the analysis on the external LD information, ATRIUM still properly controls type I error even when the reference panel is not well matched to the case-control sample. Through simulation studies, we also investigate issues that may affect the power of ATRIUM, including limited size of the reference panel and mismatch between the reference panel and the case-control sample. We compare the power of ATRIUM with that of some existing approaches for testing in this context. We apply the new method to test for association of untyped SNPs with type 2 diabetes (MIM 125853) in the Framingham SHARe data set.

## Material and Methods

Suppose a particular untyped genetic variant,  $\mathcal{U}$ , plays an important role in the disease or binary trait of interest, and suppose that this leads to an association between  $\mathcal{U}$  and case-control status. The idea behind ATRIUM is that, if we consider an appropriate set of typed markers,  $\mathcal{M} = (\mathcal{M}_1, \dots, \mathcal{M}_t)$ , known to be in strong LD with  $\mathcal{U}$  based on external information, then the primary association of  $\mathcal{U}$  with case-control status will induce a secondary association of haplotypes of  $\mathcal{M}$  with case-control status. Rather than applying a generic test of association between haplotypes of  $\mathcal{M}$  and case-control status, we can improve power to detect association attributable to the untyped variant  $\mathcal{U}$  by using the information, available from an external reference such as HapMap, on the joint distribution of alleles of  $\mathcal{U}$  and haplotypes of  $\mathcal{M}$ , to construct a 1-df test that has power in the direction that corresponds to association of  $\mathcal{U}$  with the trait. By “has power in the direction that corresponds to association of  $\mathcal{U}$  with the trait,” we mean that we test for the specific pattern of change in frequencies of haplotypes of  $\mathcal{M}$  between cases and controls that would be expected if the haplotype frequency change were driven by association between the trait and  $\mathcal{U}$ . The major methodological challenges in developing ATRIUM are due to (1) allowing samples that contain related individuals and (2) the fact that genotypes, not haplotypes, are observed. The haplotype information used by ATRIUM for each sampled individual is based on the individual’s own genotypes as well as genotype information on the individual’s parents when available. The uncertainty in haplotype information and the dependence of haplotype information across related individuals are directly taken into account in the analysis through the use of the IQLS method.<sup>20</sup> The result is a test of association with an untyped SNP, based on genotype data on associated markers  $\mathcal{M}$  with missing data allowed, where the test is applicable in samples that contain related individuals, assuming that the individuals are outbred and that pedigree information is available.

We first briefly outline the method; more detailed development is given in the following paragraphs. To construct the ATRIUM test, we start with the MQLS association model,<sup>17</sup> which was developed for testing case-control association with a typed SNP in samples that contain related individuals. The MQLS model for the association of the untyped SNP  $\mathcal{U}$  with the trait implies a model for association of the haplotypes of  $\mathcal{M}$  with the trait. This allows us to form a 1-df, complete-data, quasi-likelihood score test for case-control association with the untyped SNP in related individuals, based on haplotype data from  $\mathcal{M}$ . To extend the test to the more realistic situation in which genotype, not haplotype, information is available, we use the IQLS framework,<sup>20</sup> which provides a general

approach to quasi-likelihood inference in the presence of both dependent and missing data. Application of the IQLS framework allows us to test for case-control association with  $\mathcal{U}$  by forming the 1-df, incomplete-data, ATRIUM test based on genotype data on associated markers  $\mathcal{M}$  in related individuals.

## Mean Model for Untyped SNP

Suppose we have a case-control sample of  $n$  outbred individuals, some of whom may be related with relationships specified by known pedigrees. We arbitrarily label the two alleles of the untyped SNP  $\mathcal{U}$  as “0” and “1,” and we let  $\mathbf{U}$  be the unobserved vector of genotypes at  $\mathcal{U}$ , where we define  $\mathbf{U} = (U_1, \dots, U_n)^T$  by  $U_i = 1/2\chi$ (the number of copies of allele 1 at  $\mathcal{U}$  held by individual  $i$ ),  $1 \leq i \leq n$ . We let  $\mathbf{A}$  be the observed phenotype vector, where we define  $\mathbf{A} = (A_1, \dots, A_n)^T$  by  $A_i = 1$  if individual  $i$  is affected,  $-K/(1 - K)$  if  $i$  is unaffected, and 0 if  $i$ ’s phenotype is unknown. Here  $0 < K < 1$  is a constant that represents an external estimate of the population prevalence of the trait. (The prevalence estimate is permitted to be very rough; the MQLS test is, in fact, valid for arbitrary fixed  $K$ .) The MQLS model specifies that

$$E(U_i | \mathbf{A}) = p + \gamma(\Phi\mathbf{A})_i = p + \gamma \sum_{j=1}^n 2\phi_{ij}A_j. \quad (\text{Equation 1})$$

Here  $p$  represents the population frequency of allele 1 at  $\mathcal{U}$ , which is treated as an unknown nuisance parameter;  $\phi_{ij}$  represents the kinship coefficient between individuals  $i$  and  $j$ , which is assumed known;  $\Phi$  is the  $n \times n$  kinship matrix having  $(i, j)$ th element  $2\phi_{ij}$ ; and  $\gamma$  is the unknown parameter of interest representing the strength and direction of association between the phenotype and the alleles of  $\mathcal{U}$ . This model incorporates the enrichment effect, which specifies, for example, that affected individuals with affected relatives are more likely to have alleles predisposing to the trait than are affected individuals without affected relatives. The MQLS score test based on this model was previously shown<sup>17</sup> to have high power to detect case-control association with a typed SNP, for a variety of multilocus trait models, in samples containing related individuals.

## Mean Model for Haplotypes at Typed Markers in LD with Untyped SNP

For the untyped SNP  $\mathcal{U}$ , a set of typed SNPs  $\mathcal{M}$  is chosen based on some multilocus measure of association,<sup>14,21–23</sup> so that haplotypes of  $\mathcal{M}$  are highly informative about alleles of  $\mathcal{U}$ . Suppose there are  $H + 1$  possible haplotypes of  $\mathcal{M}$ , where for  $1 \leq j \leq H + 1$ ; we assume  $0 < h_j < 1$  is the population frequency of the  $j$ th haplotype, with  $\sum_{j=1}^{H+1} h_j = 1$ . The haplotype frequency vector  $\mathbf{h} = (h_1, \dots, h_H)^T$  is treated as an unknown nuisance parameter in the analysis. We let  $\mathbf{Y} = (Y_{11}, \dots, Y_{1H}, \dots, Y_{n1}, \dots, Y_{nH})^T$ , where  $Y_{ij} = 1/2\chi$ (the number of copies of haplotype  $j$  held by individual  $i$ ),  $1 \leq i \leq n$ ,  $1 \leq j \leq H + 1$ . If we assume that any association between haplotypes of  $\mathcal{M}$  and the trait is a secondary association that is attributable to the direct effect of  $\mathcal{U}$  on the trait, then we can derive a mean model for haplotypes of  $\mathcal{M}$  based on Equation 1. More precisely, we assume that given the allele at  $\mathcal{U}$  on a particular chromosome, the haplotype at  $\mathcal{M}$  on that chromosome is conditionally independent of the phenotype information. Then we find (Appendix A) that

$$E(Y_{ij} | \mathbf{A}) = h_j + \gamma \frac{h_j(p_{1|j} - p)}{p(1 - p)} (\Phi\mathbf{A})_i, \quad (\text{Equation 2})$$

where  $p_{1|j}$  is defined to be the conditional probability that a chromosome has allele 1 at  $\mathcal{U}$  given that it has the  $j$ th haplotype at  $\mathcal{M}$ . Thus, the association effect for the  $j$ th haplotype is  $\gamma h_j(p_{1|j} - p)/[p(1 - p)]$ , where  $\gamma$  is the association effect for allele 1 of  $\mathcal{U}$ , and  $h_j(p_{1|j} - p)/[p(1 - p)]$  is the slope of the regression line for the simple linear regression of the haplotype indicator  $\mathbf{Y}_j = (Y_{1j}, \dots, Y_{nj})^T$  on the allele indicator  $\mathbf{U}$ . If we reparameterize the model in terms of a new association parameter  $r = \gamma/[p(1 - p)]$  and apply the identity  $p = \sum_{k=1}^{H+1} h_k p_{1|k}$ , then we obtain the equivalent mean model

$$E(Y_{ij} | \mathbf{A}) = h_j + rh_j \left( p_{1|j} - \sum_{k=1}^{H+1} h_k p_{1|k} \right) (\Phi \mathbf{A})_i. \quad (\text{Equation 3})$$

The genotype data for the case-control sample give direct information on  $\mathbf{h}$  but give no information at all on the  $p_{1|j}$ 's. Therefore we replace each  $p_{1|j}$  by an estimate,  $\hat{p}_{1|j}^R$ , which is obtained from the reference panel and is taken to be

$$\hat{p}_{1|j}^R = \frac{\hat{h}_{j,1}^R}{\hat{h}_j^R}, \quad (\text{Equation 4})$$

provided  $\hat{h}_j^R > 0$ , where  $\hat{h}_j^R$  denotes an estimate of haplotype frequency  $h_j$  in the reference panel, and  $\hat{h}_{j,1}^R$  denotes an estimate, from the reference panel, of the frequency of the haplotype having allele 1 at  $\mathcal{U}$  and haplotype  $j$  at  $\mathcal{M}$ . Then we define  $\eta_j$  by

$$\eta_j = h_j \left( \hat{p}_{1|j}^R - \sum_{k=1}^{H+1} h_k \hat{p}_{1|k}^R \right), \quad (\text{Equation 5})$$

which can be treated in the analysis as a fixed function of the unknown haplotype frequency  $\mathbf{h}$ . (When  $\hat{h}_j^R = 0$ , we set  $\hat{p}_{1|j}^R = (\sum_{k=1}^{H+1} h_k \hat{p}_{1|k}^R 1_{\hat{h}_k^R > 0}) / (\sum_{l=1}^{H+1} h_l 1_{\hat{h}_l^R > 0})$ , which, when plugged into Equation (5), leads to  $\eta_j = 0$ .) Combining Equations 5 and 3, we finally obtain

$$E(Y_{ij} | \mathbf{A}) = h_j + \eta_j r(\Phi \mathbf{A})_i, \quad (\text{Equation 6})$$

which is the mean model for haplotypes of  $\mathcal{M}$ , based on the information, from the reference panel, on LD between alleles of  $\mathcal{U}$  and haplotypes of  $\mathcal{M}$ . Alternatively, we could use a logistic version (Appendix B) of the model in Equation 6, and we would still obtain the same quasi-likelihood score test as in the next subsection. We note that the role of the reference panel information is solely to determine the direction in which to perform a 1-df test, so the accuracy of the reference panel information will affect only the power, not the validity, of the score tests we propose in the next two subsections.

### ATRIUM When Haplotypes Are Observed

When the haplotypes at  $\mathcal{M}$  are observed, ATRIUM is a 1-df, quasi-likelihood score test based on the mean model in Equation 6. Our null hypothesis is  $H_0: r = 0$ , and our alternative hypothesis is  $H_A: r \neq 0$ . The null hypothesis represents no association of haplotypes of  $\mathcal{M}$  with the trait, whereas the alternative hypothesis is formulated to detect the specific kind of association that would be expected if the haplotype association were driven by the effects of the untyped SNP  $\mathcal{U}$ .

To form ATRIUM in the case of observed haplotypes, we require the null conditional covariance matrix of  $\mathbf{Y}$ , which can be written as

$$\text{Var}_0(\mathbf{Y} | \mathbf{A}) = \text{Var}_0(\mathbf{Y}) = \Phi \otimes \mathbf{B}, \quad (\text{Equation 7})$$

where  $\mathbf{B}$  is the  $H \times H$  matrix having  $(j, k)$ th element  $B_{jk} = h_j(1 - h_j)/2$  if  $j = k$  and  $-h_j h_k/2$  if  $j \neq k$ . We can also think of  $\mathbf{B}$

as being the correlation matrix of the  $H$  vector  $(Y_{11}, \dots, Y_{iH})$  for any individual  $i$ . Similarly, note that  $\Phi$  (defined in the previous subsection) is the correlation matrix of the  $n$  vector  $\mathbf{Y}_j = (Y_{1j}, \dots, Y_{nj})^T$  for any haplotype  $j$ . The Kronecker product  $\otimes$  is defined in Appendix C. ATRIUM for the case when haplotypes are observed is the quasi-likelihood score test based on the model of Equations 6 and 7, with the resulting ATRIUM test statistic given by

$$T = \frac{2 \left[ \sum_{i=1}^n A_i \sum_{j=1}^{H+1} \hat{p}_{1|j}^R (Y_{ij} - \hat{h}_j) \right]^2}{\left[ \mathbf{A}^T \Phi \mathbf{A} - (\mathbf{A}^T \mathbf{1})^2 / (\mathbf{1}^T \Phi^{-1} \mathbf{1}) \right] \left[ \sum_{k=1}^{H+1} \hat{h}_k (\hat{p}_{1|k}^R)^2 - (\sum_{l=1}^{H+1} \hat{h}_l \hat{p}_{1|l}^R)^2 \right]}, \quad (\text{Equation 8})$$

where  $\mathbf{1}$  is a vector with every entry equal to 1, and  $\hat{h}_j = (\mathbf{1}^T \Phi^{-1} \mathbf{1})^{-1} \mathbf{1}^T \Phi^{-1} \mathbf{Y}_j$  is the BLUE<sup>24</sup> of  $h_j$  under the null hypothesis of no association. In the special case when the individuals are unrelated, the term  $[\mathbf{A}^T \Phi \mathbf{A} - (\mathbf{A}^T \mathbf{1})^2 / (\mathbf{1}^T \Phi^{-1} \mathbf{1})]$  in the denominator reduces to  $\sum_{i=1}^n (A_i - \bar{A})^2$ , where  $\bar{A} = n^{-1} \sum_{i=1}^n A_i$  is the sample average of  $\mathbf{A}$ . Under regularity conditions, the ATRIUM test statistic  $T$  is asymptotically  $\chi^2_1$  distributed under the null hypothesis of no association. This asymptotic null distribution holds regardless of whether the reference panel provides biased estimates of the  $p_{1|j}$ , as might happen, for instance, when the reference panel and case-control sample are drawn from different populations. Thus, validity of our test does not depend on choice of reference panel.

While accuracy of  $\hat{p}_{1|j}^R$  does not affect validity of the test, it can affect power. When  $\hat{p}_{1|j}^R$  is exactly equal to the true  $p_{1|j}$ , which one could think of as the case of an infinitely large, perfectly matched reference panel, there is an optimality result for our test, namely, that it is asymptotically locally most powerful among tests based on  $\mathbf{Y}$  (under regularity conditions and assuming that Equations 3 and 7 hold). Thus, compared to other association tests between haplotypes of  $\mathcal{M}$  and the trait, ATRIUM should have increased power to detect case-control association with the untyped SNP  $\mathcal{U}$ .

### ATRIUM When Unphased Genotypes Are Observed

For the  $i$ th sampled individual,  $1 \leq i \leq n$ , let  $\mathbf{G}_i$  denote the observed, unphased genotype data on the associated markers  $\mathcal{M}$ , where missing genotypes are allowed, and let  $\mathbf{G} = (\mathbf{G}_1, \dots, \mathbf{G}_n)$ . Note that  $\mathbf{G}$  provides only partial information about the haplotype indicator vector  $\mathbf{Y}$ . To obtain the ATRIUM test statistic in this setting, we use the IQLS method,<sup>20</sup> which provides a quasi-likelihood score test that can be used with missing and dependent data. Instead of being based on  $\mathbf{Y}$ , which is now only partially observed, ATRIUM is the quasi-likelihood score test based on a vector,  $\mathbf{Z}$ , of conditional expectations of elements of  $\mathbf{Y}$ , which incorporates partial haplotype information. We define  $\mathbf{Z} = (Z_{11}, \dots, Z_{1H}, \dots, Z_{n1}, \dots, Z_{nH})^T$  with  $Z_{ij} = E(Y_{ij} | \mathbf{G}_i, \mathbf{G}_{mi}, \mathbf{G}_{fi}, \mathbf{A})$ , where  $\mathbf{G}_{mi}$  and  $\mathbf{G}_{fi}$  denote the observed, unphased genotype data for the mother and father of individual  $i$ , respectively, where these may be missing. Throughout the analysis, we condition on the observed pattern of missing genotypes, and we assume that the pattern of missingness is not informative about the underlying haplotypes.

We have  $E(\mathbf{Z} | \mathbf{A}) = E(\mathbf{Y} | \mathbf{A})$ , so the mean model of Equation 6 still applies to  $\mathbf{Z}$ . As before, we are interested in testing the null hypothesis of no association between the trait and the untyped SNP  $\mathcal{U}$ ,  $H_0: r = 0$  versus  $H_A: r \neq 0$ . Note that explicit computation of  $\mathbf{Z}$  under the alternative hypothesis would require additional assumptions about the genetic model. Fortunately, the IQLS

method allows one to perform the quasi-likelihood score test without having to actually compute  $\mathbf{Z}$  under the alternative model. We define  $\Omega = \text{Var}_0(\mathbf{Z}|\mathbf{A}) = \text{Var}_0(\mathbf{Z})$  to be the conditional covariance matrix of  $\mathbf{Z}$  under the null hypothesis, and we let  $\mathbf{F}_r = -E_0(\partial(\mathbf{Z} - \mu)/\partial r|\mathbf{A})$  and  $\mathbf{F}_h = -E_0(\partial(\mathbf{Z} - \mu)/\partial \mathbf{h}|\mathbf{A})$ , where  $\mu = E(\mathbf{Z}|\mathbf{A})$  with  $\mu_{ij}$  given by Equation 6 and where  $E_0(\cdot)$  denotes expectation under the null hypothesis. Additional details on how  $\Omega$ ,  $\mathbf{F}_r$  and  $\mathbf{F}_h$  are obtained can be found in Appendix D. Then the ATRIUM test statistic has the form

$$T = \left\{ \frac{[\mathbf{F}_r^T \Omega^{-1} (\mathbf{Z} - \mu)]^2}{\mathbf{F}_r^T \Omega^{-1} \mathbf{F}_r - \mathbf{F}_r^T \Omega^{-1} \mathbf{F}_h (\mathbf{F}_h^T \Omega^{-1} \mathbf{F}_h)^{-1} \mathbf{F}_h^T \Omega^{-1} \mathbf{F}_r} \right\}_{(r, \mathbf{h})=(0, \hat{\mathbf{h}})}, \quad (\text{Equation 9})$$

where the entire right-hand side is evaluated at  $(r, \mathbf{h}) = (0, \hat{\mathbf{h}})$ . Here  $\hat{\mathbf{h}}$  is the IQL estimator of  $\mathbf{h}$  when  $r = 0$ , which is the solution to the IQLS equation  $\mathbf{F}_h^T \Omega^{-1} (\mathbf{Z} - \mu) = 0$ , when  $r = 0$ . This equation is easily solved numerically by an iterative algorithm. The ATRIUM test statistic asymptotically follows a  $\chi^2_1$  distribution under the null hypothesis of no association, under regularity conditions.<sup>20</sup>

### Connections with Previous Methods for Unrelated Individuals

The WHAP method<sup>14</sup> has been developed for the special case of complete haplotype information on the set of markers  $\mathcal{M}$  in a sample consisting of equal numbers of unrelated cases and controls, where no distinction is made between unaffected controls and unphenotyped (i.e., general population) controls. The optimal-weight WHAP test statistic is obtained as a special case of our complete-data ATRIUM test statistic in Equation 8, under the assumptions that the sampled individuals are unrelated and that the controls are all of one type, either all unaffected or all unphenotyped.

The TUNA method<sup>12,13</sup> uses a similar idea to test for association with an untyped SNP in a sample of unrelated cases and controls, when unphased genotype data are available on the set of markers  $\mathcal{M}$ . Where ATRIUM uses a score test, the TUNA software<sup>13</sup> uses a Wald test based on different null and alternative hypotheses than those of ATRIUM. Nonetheless, in the special case of complete haplotype data on unrelated individuals, where all controls are of the same type, the TUNA and ATRIUM test statistics are identical except for the choice of variance estimator.

### Comparison to Other Approaches for Related Individuals

In the next section, we perform simulation studies to compare the power of ATRIUM to that of (1) the single-SNP MQLS association test<sup>17</sup> with the SNP among  $\mathcal{M}_1, \dots, \mathcal{M}_t$  that has highest  $r^2$  with the untyped SNP; (2) the full-degree-of-freedom IQLS haplotype association test<sup>20</sup> applied to haplotypes of  $\mathcal{M}$ ; and (3) a 1-df haplotype association test for deviation in the direction of the single haplotype of  $\mathcal{M}$  that has highest  $r^2$  with the untyped SNP, where test (3) is novel, to our knowledge. We now give a brief overview of tests (2) and (3).

The full-degree-of-freedom IQLS haplotype association test<sup>20</sup> is similar to ATRIUM in that it is a quasi-likelihood score test based on the vector  $\mathbf{Z}$ , but it uses a different mean model given by

$$E(Z_{ij}|\mathbf{A}) = E(Y_{ij}|\mathbf{A}) = h_j + \gamma_j(\Phi\mathbf{A})_{ij}, \quad (\text{Equation 10})$$

where  $\gamma = (\gamma_1, \dots, \gamma_H)^T$  is the parameter of interest, which measures trait-haplotype association. The full-degree-of-freedom IQLS test allows one to test  $H_0 : \gamma = 0$  versus  $H_A : \gamma \neq 0$ . Under regularity conditions, the asymptotic null distribution of the IQLS test is  $\chi^2_H$ .

In addition, we propose a 1-df haplotype association test that, to our knowledge, is novel. Let  $\alpha$  denote the haplotype of  $\mathcal{M}$  that has highest  $r^2$  with the untyped SNP. We modify the model of Equation 10 to incorporate the constraint  $\gamma_j = -\gamma_\alpha h_j / (\sum_{k=1, k \neq \alpha}^{H+1} h_k)$ , for  $j \neq \alpha$ . The resulting model has a one-dimensional parameter of interest,  $\gamma_\alpha$ , which represents association between the trait and haplotype  $\alpha$ . The constraint specifies that conditional on a haplotype being not of type  $\alpha$ , its type is independent of the phenotype information. Then we perform the quasi-likelihood score test of the null hypothesis  $H_0 : \gamma_\alpha = 0$  versus  $H_A : \gamma_\alpha \neq 0$ , which is a test of association between the trait and haplotype  $\alpha$ . This test is equivalent to the ATRIUM test with the setting  $\hat{p}_{1|\alpha}^R = 1$ , and  $\hat{p}_{1|j}^R = 0$ , for  $j \neq \alpha$ . In other words, we are effectively assuming that the untyped SNP is in perfect LD with haplotype  $\alpha$ .

## Results

### Simulation Studies

We perform simulation studies to explore the validity and power of ATRIUM. We consider a scenario in which a case-control sample is genotyped for the Illumina 300K SNP set. The European (CEU) HapMap sample is taken to be a well-matched reference panel for the population from which the case-control sample is drawn. (Note that although we use the CEU HapMap sample to choose tag SNPs, we actually simulate a reference panel for each replicate of our simulation studies, with the simulated reference panel used for calculation of  $\hat{p}_{1|j}^R$ .) We refer to any SNP in HapMap that is not in the Illumina 300K set as an “untyped” SNP. For each untyped SNP on chromosome 1, we use the TUNA software to find a set of four Illumina 300K tag SNPs that maximizes the  $M_D$  information measure<sup>23</sup> within a 400 kb window, based on the 60 HapMap CEU parents’ data. ( $M_D$  can be viewed as a multi-locus extension of  $r^2$ . It can be interpreted as the asymptotic ratio of sample sizes needed to obtain the same power when the SNP is typed versus when it is untyped and a particular set of tag SNPs is used.<sup>23</sup>) From each set of four tag SNPs, we then remove SNPs that do not provide significant information on the untyped SNP, based on the adjusted  $M_D$  measure.<sup>13</sup> We randomly choose five of the untyped SNPs on chromosome 1, from among those with  $M_D \geq 0.4$ , to be used in the simulations. Tables 1–5 list the joint distributions, estimated from the CEU HapMap sample, for each of these five untyped SNPs with their Illumina 300K tag SNPs. In the simulations, we assume that the distributions in Tables 1–5 represent the true joint distributions, of the corresponding SNP sets, in the population from which the case-control sample is drawn.

We simulate the case-control data based on a trait model with two unlinked causal SNPs, both acting dominantly, with epistasis between them. The minor allele frequencies

**Table 1. Haplotype Frequencies, Estimated from the CEU HapMap Sample, for SNP rs10797373, Denoted by  $\mathcal{U}$ , and Three Tag SNPs on the Illumina 300K Set**

Haplotype	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{U}$	Frequency
$H_1$	0	0	1	1	0.158
$H_2$	0	1	1	1	0.042
$H_3$	1	0	0	1	0.092
$H_4$	1	0	1	0	0.075
$H_5$	1	1	0	1	0.067
$H_6$	1	1	1	0	0.567

(MAFs) of SNP 1 and SNP 2 are denoted by  $p_1$  and  $p_2$ , respectively. In addition to the two allele frequencies, there are two penetrance parameters,  $f_1$  and  $f_2$  ( $f_1 > f_2$ ), with penetrance  $f_1$  for individuals who have at least one copy of the minor allele at SNP 1 and at least one copy of the minor allele at SNP 2 and penetrance  $f_2$  for all other individuals. We consider five different parameter settings, which are listed as Models a–e in Table 6. For each model, we set SNP 2 to be one of the untyped SNPs in Tables 1–5, and its MAF,  $p_2$ , is determined accordingly. Table 6 gives the trait model parameters; the resulting population prevalence,  $K$ ; and the sibling risk ratio,  $\lambda_s = K_s/K$ , where  $K_s$  is the prevalence conditioned on having an affected sibling. We sample 90 outbred, three-generation, 16-person pedigrees, of which 30 pedigrees have four affected individuals, 30 have five affected individuals, and 30 have six affected individuals. In each sampled pedigree, phenotypes for all 16 individuals are observed. The individual's genotypes are observed if and only if at least 30% of the individual's siblings, parents, and offspring are affected. This is similar to the study design in a previous report.<sup>17</sup>

An important feature of the ATRIUM analysis is that it is conditional on the value of  $\hat{p}_{1j}^R$  obtained from the reference panel. This results in a theoretical robustness property for type I error of the test, namely, ATRIUM will be valid, in the sense of having the correct type I error, even when  $\hat{p}_{1j}^R$  is a biased or inaccurate estimate of  $p_{1j}$ . We verify this in our simulations, and we also use simulations to assess the impact of the reference panel on power. Specifically, we

**Table 2. Haplotype Frequencies, Estimated from the CEU HapMap Sample, for SNP rs10907174, Denoted by  $\mathcal{U}$ , and Three Tag SNPs on the Illumina 300K Set**

Haplotype	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{U}$	Frequency
$H_1$	0	0	1	0	0.017
$H_2$	0	1	0	0	0.708
$H_3$	1	0	1	1	0.083
$H_4$	1	1	0	0	0.017
$H_5$	1	1	0	1	0.125
$H_6$	1	1	1	0	0.033
$H_7$	1	1	1	1	0.017

**Table 3. Haplotype Frequencies, Estimated from the CEU HapMap Sample, for SNP rs2794347, Denoted by  $\mathcal{U}$ , and Three Tag SNPs on the Illumina 300K Set**

Haplotype	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{U}$	Frequency
$H_1$	0	0	0	0	0.050
$H_2$	0	0	0	1	0.008
$H_3$	0	1	0	0	0.675
$H_4$	0	1	0	1	0.067
$H_5$	0	1	1	0	0.017
$H_6$	0	1	1	1	0.042
$H_7$	1	0	0	0	0.008
$H_8$	1	0	0	1	0.075
$H_9$	1	1	0	0	0.050
$H_{10}$	1	1	0	1	0.008

consider the effect on power of (1) bias in  $\hat{p}_{1j}^R$  introduced by a mismatched reference panel and (2) variability in  $\hat{p}_{1j}^R$  because of small sample size of the reference panel. To assess (1), we compare results from simulations based on three different types of reference panel: well-matched, in which the reference panel consists of phased genotype data on 60 unrelated individuals simulated based on the CEU HapMap sample; mismatched, in which the reference panel consists of phased genotype data on 90 unrelated individuals simulated based on the Asian (JPT+CHB) HapMap sample; and extremely mismatched, in which the reference panel consists of phased genotype data on 60 unrelated individuals simulated based on the African (YRI) HapMap sample. To assess (2), we compare results from simulations in which the true value of  $p_{1j}$  is plugged in for  $\hat{p}_{1j}^R$  (“perfect panel”) to those in which  $\hat{p}_{1j}^R$  is estimated from the well-matched reference panel of 60 unrelated individuals described above. In our simulations, every replicate of the case-control sample has its own simulated reference panel (except “perfect panel” in which true values are used instead of a reference panel).

For the assessment of type I error, association is tested with an untyped SNP that is unlinked and unassociated

**Table 4. Haplotype Frequencies, Estimated from the CEU HapMap Sample, for SNP rs12031614, Denoted by  $\mathcal{U}$ , and Three Tag SNPs on the Illumina 300K Set**

Haplotype	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{U}$	Frequency
$H_1$	0	0	0	1	0.025
$H_2$	0	0	1	1	0.208
$H_3$	0	1	0	0	0.108
$H_4$	0	1	0	1	0.133
$H_5$	1	0	0	0	0.008
$H_6$	1	0	0	1	0.033
$H_7$	1	0	1	0	0.050
$H_8$	1	1	0	0	0.433

**Table 5. Haplotype Frequencies, Estimated from the CEU HapMap Sample, for SNP rs10910097, Denoted by  $U$ , and Four Tag SNPs on the Illumina 300K Set**

Haplotype	$M_1$	$M_2$	$M_3$	$M_4$	$U$	Frequency
$H_1$	0	0	0	0	0	0.025
$H_2$	0	0	0	0	1	0.050
$H_3$	0	0	0	1	1	0.050
$H_4$	0	0	1	0	0	0.250
$H_5$	0	0	1	1	0	0.075
$H_6$	0	1	0	0	0	0.067
$H_7$	0	1	0	1	0	0.383
$H_8$	1	0	0	0	1	0.083
$H_9$	1	1	0	0	1	0.017

with any causal variant, where phenotype is simulated according to Model b. We compare the proportion of simulations in which the statistic exceeds the  $(1 - \alpha)$ th quantile of the  $\chi^2$  distribution to the nominal type I error level  $\alpha$ , for  $\alpha = 0.01$  and  $0.05$ . Table 7 gives the empirical type I error of the ATRIUM test, based on 10,000 replicates, where  $\hat{p}_{1|j}^R$  is obtained from a reference panel that mimics the HapMap CEU, YRI, or JPT+CHB sample. We find that even for an extremely mismatched (YRI) reference panel, type I error is not significantly different from the nominal level when  $\alpha = 0.05$  or  $0.01$ . These results verify that the type I error of the ATRIUM test for an untyped SNP is robust to the choice of reference sample.

To assess power for each model, we mask the genotypes at causal SNPs 1 and 2 and perform the tests based only on the tag SNPs for SNP 2, which are given in Tables 1–5 for Models a–e, respectively. These tag SNPs form the set  $M$  described in Material and Methods. Table 7 illustrates the power of ATRIUM for the three different types of reference panel. Under Models a, b, and e, power remains high even when the mismatched JPT+CHB reference panel is used, and power is reasonable even for the extremely mismatched YRI reference panel. However, in Models c and d, power is compromised with a mismatched reference panel. This results from the fact that the joint haplotype distribution of SNP 2 with its tag SNPs differs dramatically across

the three populations (where this distribution for the CEU sample is given in Tables 3 and 4).

Table 8 compares the power of ATRIUM, with a well-matched reference panel of 60 unrelated individuals, to the power of three other tests that are valid in samples containing related individuals: (1) the single-SNP MQLS association test with the SNP in  $M$  that has the highest  $r^2$  with the untyped SNP; (2) the full-degree-of-freedom haplotype test applied to the haplotypes of  $M$ ; and (3) the 1-df haplotype association test for deviation in the direction of the single haplotype of  $M$  that has the highest  $r^2$  with the untyped SNP. In every setting, the ATRIUM test outperforms the other three, verifying that in samples containing related individuals, the strategy of using reference panel information to select an optimal direction for testing association with an untyped SNP improves power over other approaches. To assess the possible effects on power of variability in  $\hat{p}_{1|j}^R$  as a result of small sample size of the reference panel, we also compare the power of ATRIUM with the well-matched reference panel of 60 unrelated individuals to ATRIUM with the true value of  $p_{1|j}$  plugged in for  $\hat{p}_{1|j}^R$  ("perfect panel"). Based on these simulations, there appears to be little loss of power between the unattainable perfect panel and the well-matched reference panel.

#### Analysis of Type 2 Diabetes in the Framingham SHARe Data

The Framingham Heart Study (FHS)<sup>25</sup> is a multicohort, longitudinal study of risk factors for cardiovascular disease. The FHS sample consists of unrelated individuals as well as individuals from multigenerational pedigrees. For individuals in Cohort 1 (original Framingham cohort), we use the data from exams 1–27 to determine type 2 diabetes status, which is coded as follows: individuals with at least one exam with (nonfasting) blood glucose (BG) level  $\geq 200$  mg/dl or who were under treatment for diabetes, where the measurement or treatment occurred between the ages of 35 and 75 years, are classified as affected. We classify as unaffected those who satisfy all of the following conditions: (1)  $\geq 70$  years at the time of the last exam for which BG is available, (2) BG  $< 200$  mg/dl for all exams for which it is available, and (3) not taking any treatment by the time of the last exam. We classify as unknown phenotype those who were  $< 70$  years at the time of the last exam for which BG is available and who satisfy both conditions (2) and (3). For individuals in Cohort 2 (offspring cohort), we use the data from exams 1–7, and for individuals in Cohort 3 (generation three cohort), we use the data from exam 1 to determine type 2 diabetes status, which is coded as follows for Cohorts 2 and 3: individuals with at least one exam with fasting plasma glucose (FPG)  $\geq 126$  mg/dl or who were under treatment for diabetes, where the measurement or treatment occurred between the ages of 35 and 75 years, are classified as affected. We classify as unaffected those who satisfy all of the following conditions: (1)  $\geq 70$  years at the time of the

**Table 6. Parameter Settings for Simulation Models**

Model	SNP 2	Minor Allele Frequencies		Penetrance Parameters			
		$p_1$	$p_2$	$f_1$	$f_2$	$K$	$\lambda_s$
a	rs10797373	0.15	0.358	0.15	0.05	0.066	1.118
b	rs10907174	0.30	0.225	0.22	0.09	0.116	1.075
c	rs2794347	0.40	0.200	0.40	0.20	0.246	1.045
d	rs12031614	0.50	0.400	0.18	0.07	0.123	1.077
e	rs10910097	0.35	0.200	0.25	0.10	0.131	1.081

**Table 7. Power and Type I Error of ATRIUM, When the Reference Panel Is Well Matched to the Sample or Mismatched, Based on 5,000 Simulated Replicates for Power or 10,000 Simulated Replicates for Type I Error**

Reference Panel	Empirical Type I Error (SE)		Estimated Power (SE) with Significance Level of 0.05				
	$\alpha = 0.05$	$\alpha = 0.01$	Model a	Model b	Model c	Model d	Model e
CEU (match)	0.050 (0.002)	0.009 (0.0010)	0.915 (0.004)	0.957 (0.003)	0.868 (0.005)	0.859 (0.005)	0.985 (0.002)
JPT+CHB (mismatch)	0.049 (0.002)	0.011 (0.0010)	0.911 (0.004)	0.944 (0.003)	0.447 (0.007)	0.782 (0.006)	0.932 (0.004)
YRI (extreme mismatch)	0.051 (0.002)	0.010 (0.0010)	0.838 (0.005)	0.750 (0.006)	0.355 (0.007)	0.432 (0.007)	0.930 (0.004)

last exam for which FPG is available, (2) FPG < 126 mg/dl for all exams for which it is available, and (3) not taking any treatment by the time of the last exam. We classify as unknown phenotype those who were < 70 years at the time of the last exam for which FPG is available and who satisfy both conditions (2) and (3). Note that for exams 1 and 2 for Cohort 2, instead of the FPG  $\geq 126$  mg/dl criterion, we use the hand-curated diabetes mellitus (DM) status, which is based on detailed chart review and is available for exams 1 and 2 of Cohort 2 in Framingham SHARe. This study had approval from dbGaP and from the University of Chicago Institutional Review Board.

Among the FHS individuals who are genotyped on the Affymetrix 500K array, and who are coded as affected, unaffected, or unknown phenotype based on the above criteria, we include only those who satisfy the following quality-control conditions: (1) completeness  $> 96\%$ , where completeness is the proportion of all markers on the Affymetrix 500K array for which a given individual has genotypes called, and (2) diagonal of empirical kinship matrix  $< 1.05$ . We also use the off-diagonals of the empirical kinship matrix to exclude an additional 298 individuals with kinship values that are not consistent with the pedigree information. Based on the above criteria, a total of 7,678 individuals are retained in the analysis, with 576 affected, 1,254 unaffected, and 5,848 unknown phenotype, of which 793 are original cohort, 3,142 are offspring cohort, and 3,743 are third generation.

We perform case-control association testing of type 2 diabetes with both typed and untyped SNPs on chromosome 10 in the Framingham SHARe data. The typed SNPs in the analysis consist of the 21,777 SNPs on chromosome 10 that are on the Affymetrix 500K array, pass the Affyme-

trix 500K quality-control tests, and meet the following additional criteria: (1) call rate  $\geq 96\%$ , (2) Mendelian error rate  $\leq 0.02$ , and (3) MAF  $\geq 0.01$ . The untyped SNPs in the analysis are the 60,344 SNPs on chromosome 10 in HapMap that meet the following criteria: (1) are not among the 21,777 typed SNPs and (2) are tagged by at least two typed SNPs, where the tagging is done by applying TUNA to the CEU HapMap samples, in a similar fashion to what is described in detail in subsection *Simulation Studies* (with the Illumina 300K array replaced by the Affymetrix 500K array and with chromosome 1 replaced by chromosome 10). We use MQLS to test for association with each typed SNP and ATRIUM to test for association with each untyped SNP. The prevalence of diagnosed and undiagnosed diabetes among people aged 60 years or older in the United States is reported to be  $\sim 23\%$ , with type 2 diabetes accounting for about 90%–95% of all diagnosed cases of diabetes.<sup>26</sup> Therefore, we set  $K = 0.2$  in the MQLS and ATRIUM analyses.

Table 9 gives the results for SNPs for which the corresponding ATRIUM or MQLS test has a p value  $< 8.0 \times 10^{-5}$ . The first SNP in Table 9 (rs2904802) is in an intron of the gene encoding aldo-keto reductase family 1, member C1 (*AKR1C1* [MIM 600449]), located at 10p15.1. The next four SNPs in Table 9 (rs7901695, rs4506565, rs7903146, and rs4132670) are intronic SNPs for the gene encoding transcription factor 7-like 2 (*TCF7L2* [MIM 602228]), located at 10q25.2. Recently, five separate type 2 diabetes genome-wide association studies<sup>27–31</sup> have identified association between *TCF7L2* and type 2 diabetes. One study<sup>32</sup> has shown that the rs7903146 T allele is associated with hepatic insulin resistance and diminished glucose-stimulated plasma insulin secretion. The sixth SNP in Table 9

**Table 8. Power of ATRIUM Compared to Single-SNP and Haplotype Association Tests, Based on 5,000 Simulated Replicates,  $\alpha = 0.05$ , CEU Reference Panel**

Test	Estimated Power (SE)				
	Model a	Model b	Model c	Model d	Model e
SNP	0.555 (0.007)	0.946 (0.003)	0.530 (0.007)	0.749 (0.006)	0.793 (0.006)
HAP (full degree of freedom)	0.723 (0.006)	0.856 (0.005)	0.713 (0.006)	0.650 (0.007)	0.899 (0.004)
HAP (1 degree of freedom)	0.849 (0.005)	0.935 (0.003)	0.723 (0.006)	0.832 (0.005)	0.741 (0.006)
ATRIUM	0.915 (0.004)	0.957 (0.003)	0.868 (0.005)	0.859 (0.005)	0.985 (0.002)
ATRIUM (perfect panel) <sup>a</sup>	0.915 (0.004)	0.960 (0.003)	0.887 (0.005)	0.867 (0.005)	0.987 (0.002)

<sup>a</sup> ATRIUM (perfect panel) is ATRIUM with true  $p_{1|j}$  plugged in for  $\hat{p}_{1|j}^R$ .

**Table 9. Type 2 Diabetes Association Results in Framingham SHARe, for Typed and Untyped SNPs on Chromosome 10**

Region	Gene	SNP	Position (nucleotides)	Typed	# of Tag SNPs	$M_D$	Test	p Value
10p15.1	<i>AKR1C1</i>	rs2904802	4,999,364	no	4	0.44	ATRIUM	4.8e-5
10q25.2	<i>TCF7L2</i>	rs7901695	114,744,078	yes	-	-	MQLS	4.9e-6
10q25.2	<i>TCF7L2</i>	rs4506565	114,746,031	yes	-	-	MQLS	3.1e-6
10q25.2	<i>TCF7L2</i>	rs7903146	114,748,339	no	3	0.94	ATRIUM	1.9e-5
10q25.2	<i>TCF7L2</i>	rs4132670	114,757,761	yes	-	-	MQLS	2.4e-5
10q26.13	<i>GPR26<sup>a</sup></i>	rs859510	125,337,377	no	4	1.00	ATRIUM	7.8e-5
10q26.2	<i>DOCK1</i>	rs4615933	128,716,002	no	4	0.65	ATRIUM	3.4e-5
10q26.2	<i>DOCK1</i>	rs6482989	128,753,959	no	3	0.68	ATRIUM	1.6e-5
10q26.2	<i>DOCK1</i>	rs9418739	128,758,976	yes	-	-	MQLS	3.4e-5

<sup>a</sup> rs859510 is 78.5 kb upstream of the gene *GPR26*.

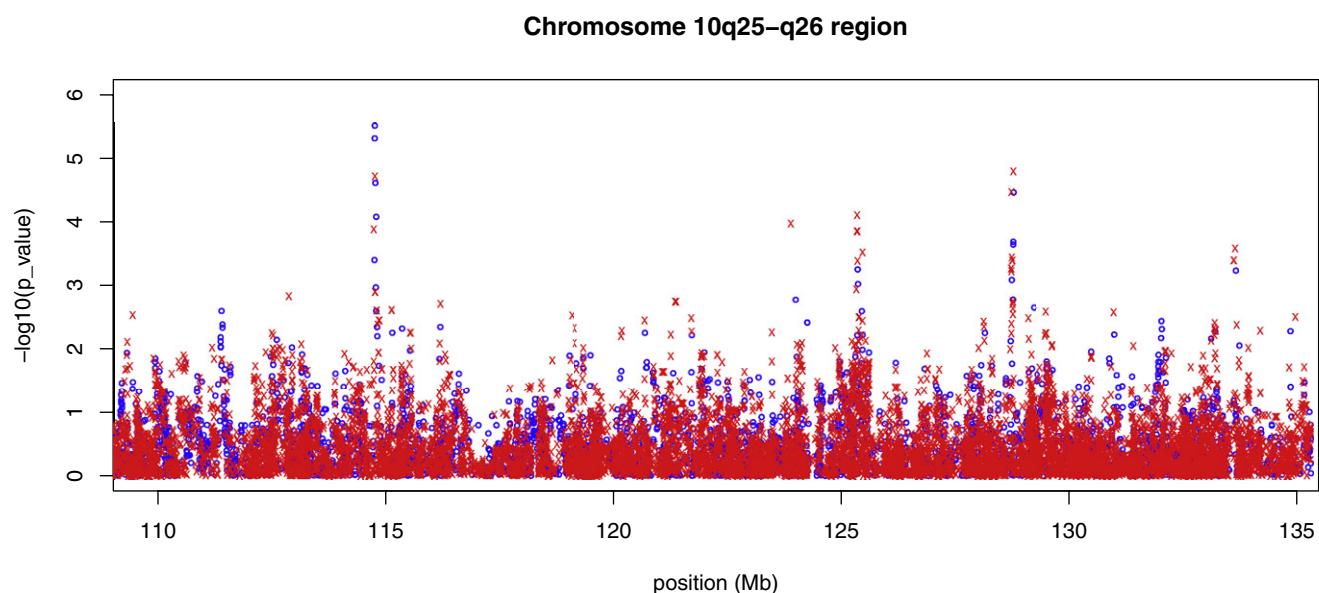
(rs859510) is 78.5 kb from the gene encoding G protein-coupled receptor 26 (*GPR26* [MIM 604847]), located at 10q26.13. This SNP lies in a previously identified,<sup>33</sup> 6.3 Mb linkage region, for hemoglobin A1c (HbA1c), a diabetes-related quantitative glucose trait, in the FHS based on the Affymetrix 100K array. The last three SNPs in Table 9 (rs4615933, rs6482989, and rs9418739) are intronic SNPs for the gene encoding dedicator of cytokinesis 1 (*DOCK1* [MIM 601403]), located at 10q26.2.

Previous studies have found the region 10q25-q26 to be syntenic to several quantitative trait loci for both weight and type 2 diabetes in rats.<sup>34-36</sup> 10q25-q26 also contains a 6.3 Mb region showing evidence for linkage to HbA1c in FHS.<sup>33</sup> Figure 1 shows our FHS type 2 diabetes association results for the 10q25-q26 region, where we plot  $-\log_{10}(p\text{ value})$  for each of 12,653 untyped HapMap SNPs, based on ATRIUM, and 4,550 typed Affymetrix 500K

SNPs, based on MQLS. Overall, the untyped SNPs are well represented among those SNPs having small p values, taking into account the fact that although they represent about 75% of the tested SNPs, the tests for untyped SNPs are generally expected to have lower power because of lower information content. In the previously identified linkage region (119.5–125.8 Mb), the untyped SNPs provide stronger evidence of association than do the typed SNPs.

#### Assessment of Computation Time

The computational burden of untyped SNP analysis is much greater in samples containing related individuals than it is in samples of unrelated individuals. This is because the dependence, across related individuals, of the partial information on untyped-SNP genotypes must be assessed and incorporated into the analysis. We note



**Figure 1. Association Results for Type 2 Diabetes in Framingham SHARe for the Region 10q25-q26**

$-\log_{10}(p\text{ value})$  is plotted against chromosomal location for 4,550 typed SNPs (blue circles) and 12,653 untyped SNPs (red x).

that this dependence differs from the ordinary dependence among genotypes and is affected by the type and amount of information available, on the untyped-SNP genotype, for each individual. Thus, the computationally costly part of ATRIUM is the calculation, at each iteration, of the covariance matrix  $\Omega$ .

The Framingham SHARe data we analyze contain some extremely large pedigrees. For instance, there is one family with 317 genotyped individuals who are included in our analysis, and there are four other families that each have more than 90 genotyped individuals included in our analysis. Accordingly, we did two different assessments of the computation time needed to test the 60,344 untyped SNPs of chromosome 10: (1) computation with extremely large pedigrees, where we include 7,678 individuals from 1,147 families with the number of genotyped individuals per family ranging from 1 to 317, with five families having 90 or more genotyped individuals; and (2) computation with moderate-size pedigrees, where we include 4,926 individuals from 1,084 families in which there are no more than 20 genotyped individuals per family. With extremely large pedigrees, the analysis took 17.5 hr on an Intel 2.6 GHz Mac laptop with 4 GB RAM when the haplotype information for an individual is based only on that individual's genotype (i.e., when  $Z_{ij}$  is taken to be  $Z_{ij} = E(Y_{ij}|\mathbf{G}_i, \mathbf{A})$ ), and it took 45 hr when we included parental genotype data when considering an individual's haplotype information (i.e., when  $Z_{ij}$  is taken to be  $Z_{ij} = E(Y_{ij}|\mathbf{G}_i, \mathbf{G}_{mi}, \mathbf{G}_{fi}, \mathbf{A})$ ). With moderate-size pedigrees, it took 4.5 hr to do the analysis when haplotype information for an individual is based only on that individual's genotype, and it took 12 hr to do the analysis when we included parental genotype data when considering an individual's haplotype information. We have not optimized the code, so it is likely that these times could be greatly improved.

## Discussion

We propose the ATRIUM method for testing association with untyped genetic variants in samples containing general combinations of related and unrelated individuals. An important feature of ATRIUM is that it properly accounts for dependence in the partial information on untyped-SNP genotypes across related individuals, which is crucial for construction of a valid test. ATRIUM is potentially useful in a wide range of study designs, including extremely large pedigrees as well as samples that combine families and unrelated individuals. ATRIUM uses information from an external reference panel, such as HapMap, to select an optimal direction for testing association with an untyped SNP, based on genotype data from typed SNPs. ATRIUM allows both phased and unphased genotype data for both the case-control sample and the reference panel. We demonstrate, both theoretically and through simulation, that the validity of ATRIUM is robust to mismatch between the reference panel and the case-control sample, though power is highest when the refer-

ence panel is reasonably well-matched to the case-control sample. We also find that small sample size of the reference panel results in little loss of power. We further demonstrate, both theoretically and through simulation, that ATRIUM provides higher power to detect association with untyped SNPs than do other single-SNP and haplotype tests based on typed SNPs.

We apply ATRIUM to the Framingham SHARe data to test for association between type 2 diabetes and SNPs on chromosome 10 that are in HapMap but are untyped on the Affymetrix 500K array in the Framingham sample. We replicate association between type 2 diabetes and intronic SNPs of the *TCF7L2* gene, where we obtain p values for association in the range of 3.1e-6 to 2.4e-5 for three typed and one untyped intronic SNP of *TCF7L2*. We also obtain p values < 8e-5 for SNPs in or near three other genes on chromosome 10, including one typed and two untyped intronic SNPs for the *DOCK1* gene. In a previously identified linkage region for a diabetes-related phenotype,<sup>33</sup> the untyped SNPs, analyzed with ATRIUM, provide stronger evidence for association than do the typed SNPs.

A key challenge that arises in association analysis with samples of related individuals is that specification of an alternative model is vastly more problematic with related than with unrelated individuals. This is because the background effects of environmental factors and multiple loci, other than the particular variant being tested, can create very different patterns of dependence of phenotype among related individuals under different modeling assumptions. This makes it particularly challenging to develop appropriate likelihood-based or Bayesian analyses in this context. The use of a score-function or quasi-score-function approach, as in ATRIUM, avoids this problem. A close connection between the MQS and IQLS tests and the retrospective likelihood score test has previously been shown.<sup>20</sup> A key point is that the retrospective likelihood score, MQS, IQLS, and ATRIUM tests can be formed without specifying the joint distribution of phenotypes among related individuals under the null or alternative hypothesis, whereas likelihood-based and Bayesian approaches would, in principle, require this to be specified.

Ideally, one should be able to improve the power of ATRIUM by making use of one or more of the available hidden Markov model (HMM) imputation approaches<sup>1–8</sup> for the modeling and computation of the conditional probabilities  $\hat{p}_{1|j}^R$  that are needed from the reference database. A difficulty is that the calculation of the required covariance matrix  $\Omega$  becomes increasingly onerous as the number of possible haplotypes of typed SNPs used to predict the untyped SNP increases. An analogous problem arises if imputation methods are used directly on the case-control sample with related individuals, because it would be necessary to compute and take into account the joint posterior distribution, across related individuals, of the untyped SNP genotypes. To our knowledge, current imputation methods do not provide these joint posterior probabilities for related individuals.

When comparing p values across SNPs, particularly when there are both typed and untyped SNPs being tested, it is useful to keep in mind that the p value does not make any adjustment for the differing power of the tests. This is particularly relevant when typed and untyped SNP p values are compared, because power to detect association with an untyped SNP will be reduced to the extent that the untyped SNP is not well characterized by haplotypes of typed SNPs. The extent of information on the untyped SNP, relative to the information that would be available if the SNP were typed, can be assessed by the  $M_D$  measure,<sup>23</sup> which can be helpful in interpreting the resulting p values.

When the sample size of the reference panel is small, it can arise that the case-control sample contains a haplotype (call it haplotype  $j$ ), for the typed SNPs, that has an estimated frequency of zero in the reference panel. Thus, the reference panel lacks information on LD between haplotype  $j$  and the untyped SNP. As described in subsection, Mean Model for Haplotypes at Typed Markers in LD with Untyped SNP, we solve this problem by treating haplotype  $j$  as independent of the untyped SNP. An alternative approach<sup>12</sup> would be to group haplotype  $j$  with the closest haplotype (call it haplotype  $k$ ), among those having nonzero estimated frequency in the reference panel, and then set  $\hat{p}_{1|j}^R = \hat{p}_{1|k}^R$ . This approach assumes that the untyped SNP has the same conditional distribution given haplotype  $j$  as given haplotype  $k$ . In most cases, one would expect little difference between the two approaches, but if haplotype  $j$  were greatly enriched in the case-control sample, the results of the two approaches might be somewhat different. This problem could be resolved with larger reference panels.

## Appendix A

We derive the mean model of Equation 2 for haplotypes of  $\mathcal{M}$  from the mean model of Equation 1 for alleles of  $\mathcal{U}$ . Let  $\mathbf{X}_i = (a_i, x_i)$  be a random variable representing a randomly chosen combined haplotype from individual  $i$ , where  $a_i$  is the allele at  $\mathcal{U}$  and  $x_i$  is the haplotype at  $\mathcal{M}$ . Then our assumption is that  $P(x_i = j|a_i = m, \mathbf{A}) = p_{j|m}$  for  $m = 0, 1, 2, \dots, H + 1$ , i.e., that given the allele at  $\mathcal{U}$ , the haplotype at  $\mathcal{M}$  is conditionally independent of the phenotype information. Note that  $E(U_i|\mathbf{A}) = P(a_i = 1|\mathbf{A})$  and  $E(Y_{ij}|\mathbf{A}) = P(x_i = j|\mathbf{A})$ . Then we have  $E(Y_{ij}|\mathbf{A}) = E(U_i|\mathbf{A})P(x_i = j|a_i = 1, \mathbf{A}) + [1 - E(U_i|\mathbf{A})]P(x_i = j|a_i = 0, \mathbf{A}) = E(U_i|\mathbf{A})p_{j|1} + [1 - E(U_i|\mathbf{A})]p_{j|0}$ , and the result follows, where we use the fact that  $p_{j|1}p + p_{j|0}(1 - p) = h_j$  and that  $p_{j|1} - p_{j|0} = h_j(p_{1|j} - p)/[p(1 - p)]$ .

## Appendix B

Instead of Equation 6, we could use the logistic model

$$\text{logit}[E(Y_{ij}|\mathbf{A})] = \text{logit}(h_j) + \frac{\eta_j}{h_j(1 - h_j)} r(\Phi\mathbf{A})_i, \quad (\text{Equation 11})$$

where all quantities have the same definitions as in Equation 6. The quasi-likelihood score test for  $H_0:r = 0$  versus  $H_A:r \neq 0$  based on Equation 11 is identical to that based on Equation 6, where this test statistic is given in Equation (8). Conceptually, the advantage of the logistic model (11) over the linear model (6) is that in the logistic model,  $r$  can be any real number, whereas in the linear model,  $r$  is constrained in a rather complicated way relative to  $\mathbf{h} = (h_1, \dots, h_H)^T$  to ensure that  $0 \leq E(Y_{ij}|\mathbf{A}) \leq 1$ .

## Appendix C

Given the  $n \times m$  matrix  $\mathbf{A}$  with  $(i, j)$ th element  $a_{ij}$  and the  $p \times q$  matrix  $\mathbf{B}$  with  $(i, j)$ th element  $b_{ij}$ , their Kronecker product, denoted by  $\mathbf{A} \otimes \mathbf{B}$ , is the  $np \times mq$  matrix with block structure

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1m}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & \cdots & a_{nm}\mathbf{B} \end{pmatrix}_{np \times mq}.$$

## Appendix D

We assume that the markers in the set  $\mathcal{M}$  are tightly linked, and that, under the null hypothesis, both Hardy-Weinberg equilibrium and Mendelian inheritance hold. Let  $\mathbf{Z}^0$  denote  $\mathbf{Z}$  evaluated under the null hypothesis,  $r = 0$ . Then we have  $Z_{ij}^0 = E_0(Y_{ij}|\mathbf{G}_i, \mathbf{G}_{mi}, \mathbf{G}_{fi})$ , which can be explicitly computed, as a function of  $\mathbf{h}$  and  $(\mathbf{G}_i, \mathbf{G}_{mi}, \mathbf{G}_{fi})$ , for an outbred, parent-offspring trio, where we allow some genotypes to be missing. When we plug in  $\hat{\mathbf{h}}$  for  $\mathbf{h}$ , we obtain  $\mathbf{Z}$  evaluated at  $(r, \mathbf{h}) = (0, \hat{\mathbf{h}})$ , which is needed for Equation 9. We can obtain  $\Omega = \text{Var}_0(\mathbf{Z}) = \text{Var}_0(\mathbf{Z}^0)$  by finding the joint conditional distribution of  $(\mathbf{G}_i, \mathbf{G}_{mi}, \mathbf{G}_{fi}, \mathbf{G}_j, \mathbf{G}_{mj}, \mathbf{G}_{fj})$  for each pair of sampled individuals  $(i, j)$ , given the pattern of missing genotypes, where this distribution is explicitly computed as a function of  $\mathbf{h}$ , the pedigree information, and the pattern of missingness. To obtain  $\mathbf{F}_h$ , note that  $\mathbf{F}_h = -E_0(\partial(\mathbf{Z} - \mu)/\partial\mathbf{h}|\mathbf{A}) = -E_0(\partial\mathbf{Z}^0/\partial\mathbf{h}) + \mathbf{1}_n \otimes \mathbf{I}_H$ , where  $\mathbf{1}_n$  is an  $n$ -vector with all entries equal to 1 and  $\mathbf{I}_H$  is the  $H \times H$  identity matrix. We can explicitly obtain  $\partial\mathbf{Z}^0/\partial\mathbf{h}$  from  $\mathbf{Z}^0$ , and the null expectation is obtained from the joint conditional distribution of  $(\mathbf{G}_i, \mathbf{G}_{mi}, \mathbf{G}_{fi})$ , given the pattern of missing genotypes, for each sampled individual  $i$ . Finally, consider  $\mathbf{F}_r = -E_0(\partial(\mathbf{Z} - \mu)/\partial r|\mathbf{A}) = -E_0(\partial\mathbf{Z}^0/\partial r|\mathbf{A}) + (\Phi\mathbf{A}) \otimes \boldsymbol{\eta}$ , where  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_H)^T$ . Note that knowledge of  $\mathbf{Z}^0$  is not sufficient to obtain  $E_0(\partial\mathbf{Z}/\partial r|\mathbf{A})$ . At the same time, it is not necessary to fully specify  $\mathbf{Z}$  under the alternative model either. Instead, we need the first-order term of the power series expansion for  $\mathbf{Z}$  around  $r = 0$ . This first-order term is the same for any two-allele disease model for the untyped SNP, and we use it to obtain  $\mathbf{F}_r$ . Note that this is the same assumption used by Thornton and McPeek (2007)<sup>17</sup> to obtain the MQS mean model, so we do not need to impose any additional assumptions to obtain  $\mathbf{F}_r$ .

## Acknowledgments

We thank Mark Abney and William Wen for discussion, critical comments, and help with software implementation; Dan Nicolae, Matthew Stephens, and Peter McCullagh for discussion and critical comments; and two anonymous reviewers for critical comments. This study was supported by National Institutes of Health grant R01 HG001645. The Framingham Heart Study (FHS) and the Framingham SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. The Framingham SHARe data used for the analyses described in this manuscript were obtained through dbGaP (phs000007.v6.p3). This work was not prepared in collaboration with investigators of the FHS and does not necessarily reflect the opinions or views of the FHS, Boston University, or NHLBI.

Received: June 18, 2009

Revised: October 6, 2009

Accepted: October 9, 2009

Published online: November 12, 2009

## Web Resources

The URLs for data presented herein are as follows:

EntrezGene, <http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>  
ATRIUM source code, <http://www.stat.uchicago.edu/~mcpeek/software/index.html>  
Online Mendelian Inheritance in Man (OMIM), <http://www.ncbi.nlm.nih.gov/Omim/>

## References

- Stephens, M., and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76**, 449–462.
- Scheet, P., and Stephens, M. (2006). A fast and flexible statistical method for large-scale population genotype data: Application to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644.
- 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.
- Servin, B., and Stephens, M. (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet.* **3**, e114.
- Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097.
- Willer, C.J., Sanna, S., Jackson, A.U., Scuteri, A., Bonnycastle, L.L., Clarke, R., Heath, S.C., Timpson, N.J., Najjar, S.S., Stringham, H.M., et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* **40**, 161–169.
- Guan, Y., and Stephens, M. (2008). Practical issues in imputation-based association mapping. *PLoS Genet.* **4**, e1000279.
- 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.
- Lin, D.Y., Hu, Y., and Huang, B.E. (2008). Simple and efficient analysis of disease association with missing genotype data. *Am. J. Hum. Genet.* **82**, 444–452.
- Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., DiGenova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F., et al. (2001). Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**, 233–237.
- de Bakker, P.I.W., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nat. Genet.* **37**, 1217–1223.
- Nicolae, D.L. (2006). Testing untyped alleles (TUNA)—applications to genome-wide association studies. *Genet. Epidemiol.* **30**, 718–727.
- Wen, X., and Nicolae, D.L. (2008). Association studies for untyped markers with TUNA. *Bioinformatics* **24**, 435–437.
- Zaitlen, N., Kang, H.M., Eskin, E., and Halperin, E. (2007). Leveraging the HapMap correlation structure in association studies. *Am. J. Hum. Genet.* **80**, 683–691.
- Slager, S.L., and Schaid, D.J. (2001). Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects. *Am. J. Hum. Genet.* **68**, 1457–1462.
- Bourgain, C., Hoffjan, S., Nicolae, R., Newman, D., Steiner, L., Walker, K., Reynolds, R., Ober, C., and McPeek, M.S. (2003). Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. *Am. J. Hum. Genet.* **73**, 612–626.
- Thornton, T., and McPeek, M.S. (2007). Case-control association testing with related individuals: A more powerful quasi-likelihood score test. *Am. J. Hum. Genet.* **81**, 321–337.
- Browning, S.R., Briley, J.D., Briley, L.P., Chandra, G., Charnecki, J.H., Ehm, M.G., Johansson, K.A., Jones, B.J., Karter, A.J., Yarnall, D.P., and Wagner, M.J. (2005). Case-control single-marker and haplotypic association analysis of pedigree data. *Genet. Epidemiol.* **28**, 110–122.
- Zhang, J., Schneider, D., Ober, C., and McPeek, M.S. (2005). Multilocus linkage disequilibrium mapping by the decay of haplotype sharing with samples of related individuals. *Genet. Epidemiol.* **29**, 128–140.
- Wang, Z., and McPeek, M.S. (2009). An incomplete-data quasi-likelihood approach to haplotype-based genetic association studies on related individuals. *J. Am. Stat. Assoc.* **104**, 1251–1260.
- Chapman, J.M., Cooper, J.D., Todd, J.A., and Clayton, D.G. (2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power. *Hum. Hered.* **56**, 18–31.
- Stram, D.O. (2004). Tag SNP selection for association studies. *Genet. Epidemiol.* **27**, 365–374.
- Nicolae, D.L. (2006). Quantifying the amount of missing information in genetic association studies. *Genet. Epidemiol.* **30**, 703–717.
- McPeek, M.S., Wu, X., and Ober, C. (2004). Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics* **60**, 359–367.
- Splansky, G.L., Corey, D., Yang, Q., Atwood, L.D., Cupples, L.A., Benjamin, E.J., D'Agostino, R.B. Sr., Fox, C.S., Larson, M.G., Murabito, J.M., et al. (2007). The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: Design, recruitment, and initial examination. *Am. J. Epidemiol.* **165**, 1328–1335.

26. National Institute of Diabetes and Digestive and Kidney Diseases. (2008). National Diabetes Statistics, 2007 Fact Sheet (Bethesda, MD: U.S. Department of Health and Human Services, National Institutes of Health).
27. Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885.
28. Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I.W., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N., Daly, M.J., et al. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336.
29. Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R.B., Rayner, N.W., Freathy, R.M., et al. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341.
30. Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345.
31. Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G.B., Styrkarsdottir, U., Gretarsdottir, S., Emilsson, V., Ghosh, S., et al. (2007). A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat. Genet.* **39**, 770–775.
32. Wegner, L., Hussain, M.S., Pilgaard, K., Hansen, T., Pedersen, O., Vaag, A., and Poulsen, P. (2008). Impact of TCF7L2 rs7903146 on insulin secretion and action in young and elderly Danish twins. *J. Clin. Endocrinol. Metab.* **93**, 4013–4019.
33. Meigs, J.B., Manning, A.K., Fox, C.S., Florez, J.C., Liu, C., Cupples, L.A., and Dupuis, J. (2007). Genome-wide association with diabetes-related traits in the Framingham Heart Study. *BMC Med. Genet.* **8** (*Suppl 1*), S16.
34. Kato, N., Hyne, G., Bihoreau, M.T., Gauguier, D., Lathrop, G.M., and Rapp, J.P. (1999). Complete genome searches for quantitative trait loci controlling blood pressure and related traits in four segregating populations derived from Dahl hypertensive rats. *Mamm. Genome* **10**, 259–265.
35. Galli, J., Li, L.S., Glaser, A., Ostenson, C.G., Jiao, H., Fakhrai-Rad, H., Jacob, H.J., Lander, E.S., and Luthman, H. (1996). Genetic analysis of non-insulin dependent diabetes mellitus in the GK rat. *Nat. Genet.* **12**, 31–37.
36. Klöting, I., Kovács, P., and van den Brandt, J. (2001). Sex-specific and sex independent quantitative trait loci for facets of the metabolic syndrome in WOKW rats. *Biochem. Biophys. Res. Commun.* **284**, 150–156.