Identifying Plausible Genetic Models Based on Association and Linkage Results: Application to Type 2 Diabetes

Weihua Guan,[1]* Michael Boehnke,[2] Anna Pluzhnikov,[3] Nancy J. Cox,[3] and Laura J. Scott[2]

[1] *Division of Biostatistics School of Public Health, University of Minnesota, Minneapolis, Minnesota*
[2] *Department of Biostatistics and Center for Statistical Genetics School of Public Health, University of Michigan, Ann Arbor, Michigan*
[3] *Section of Genetic Medicine Department of Medicine, University of Chicago, Chicago, Illinois*

When planning resequencing studies for complex diseases, previous association and linkage studies can constrain the range of plausible genetic models for a given locus. Here, we explore the combinations of causal risk allele frequency (RAF_C) and genotype relative risk (GRR_C) consistent with no or limited evidence for affected sibling pair (ASP) linkage and strong evidence for case-control association. We find that significant evidence for case-control association combined with no or moderate evidence for ASP linkage can define a lower bound for the plausible RAF_C. Using data from large type 2 diabetes (T2D) linkage and genome-wide association study meta-analyses, we find that under reasonable model assumptions, 23 of 36 autosomal T2D risk loci are unlikely to be due to causal variants with combined $RAF_C < 0.005$, and four of the 23 are unlikely to be due to causal variants with combined $RAF_C < 0.05$. *Genet. Epidemiol.* 00:1–9, 2012. © 2012 Wiley Periodicals, Inc.

Key words: gene mapping; genetics; genetic structure; complex diseases

Supporting Information is available in the online issue at wileyonlinelibrary.com.
Contract grant sponsor: National Institutes of Health; Contract grant numbers: HG000376; DK062370.

*Correspondence to: Weihua Guan, Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303, 420 Delaware St. S.E., Minneapolis, MN 55455. E-mail: wguan@umn.edu

Received 24 January 2012; Revised 19 June 2012; Accepted 20 June 2012
Published online in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21668

INTRODUCTION

Genome-wide association studies (GWAS) allow investigators to test for disease or trait (henceforward disease) association with common single nucleotide polymorphisms (SNPs) throughout the human genome. Today's commercial GWAS platforms, when combined with genotype imputation [e.g., Li et al., 2010; Marchini et al., 2007], typically cover 80–90% of known common genetic variants (minor allele frequency (MAF) > 0.05). In recent years, GWAS have been conducted for many diseases [see http://www.genome.gov/gwastudies/]. The combined effects of associated variants often explain only a small proportion of the disease genetic variation [Manolio et al., 2009]. Results to date suggest that most common variants associated with complex diseases have modest effect on disease risk. Less common ($0.005 < MAF < 0.05$) and rare ($MAF < 0.005$) variants have not yet been studied extensively and may (or may not) have larger effect sizes. With the recent advances in sequencing technology, it has become feasible to identify and genotype these variants. While multiple theoretical and data-driven approaches have examined genetic architecture of complex human diseases and traits [Anderson et al., 2011; Dickson et al., 2010; Lee et al., 2011; Pritchard and Cox, 2002; Purcell et al., 2009; Reich and Lander, 2001; Risch and Merikangus, 2001; Wray et al., 2011], our knowledge of their underlying architecture remains limited. Current and planned large-scale sequencing studies seek to address this issue.

Previous complex disease linkage studies generally reported limited evidence for linkage, and even in studies with strong linkage signals, most of the genome provides no evidence for linkage. These negative linkage results should limit the range of plausible effect sizes for disease risk variants and/or the cumulative frequency of risk variants. Similarly, evidence (or lack of evidence) for association in a region of interest should also limit the range of plausible models for these risk variant(s).

Existing association and/or linkage results together with simulations have been used by multiple groups to explore the likely genetic architectures underlying complex diseases. Purcell et al. [2009] showed that rare or less common causal variants are unlikely to be the sole explanation of schizophrenia genetic variation based on simulations to identify models that are consistent with GWAS results and heritability estimates. Similarly, the work of Wray et al. [2011] suggests that rare variants are unlikely to underlie a large proportion of GWAS associations as they would explain >100% of the heritability. Dickson et al. [2010] argued that many common variants identified in GWAS could reflect multiple less common ($0.005 < MAF < 0.02$) causal variants in high linkage disequilibrium (LD); using the same models, Anderson et al. [2010] concluded that rare variants were unlikely to underlie most GWAS-associated variants. Both studies provided graphical representations of the power of affected sibling pair (ASP) linkage (with Anderson et al. assuming a much larger linkage sample) and SNP disease association under limited number of models.

However, they did not provide a quantitative way to define plausible models (minimum RAF_C, maximum GRR_C) for specified disease loci given the results of existing linkage and association studies.

In this paper, we seek to identify the plausible range of genetic models, in terms of genotype relative risk (GRR_C) and risk allele frequency (RAF_C), consistent with rare or less common causal variant(s) underlying a given disease association. We consider scenarios in which no or modest evidence for ASP linkage is reported, and/or significant evidence for association is reported. To do so, we calculate the power to detect ASP linkage and/or case-control association and summarize the range of genetic models that appears plausible given results from available linkage and/or association studies. Our results show that for each risk allele frequency RAF_C, the effect sizes GRR_C of causal variants are constrained by ASP linkage or association results. When significant evidence for association is combined with no or modest ASP linkage evidence in the same chromosomal region, causal variants with small RAF_C can also be identified as implausible. In our calculations, we assume that a single causal variant underlies a common variant association, but our results can be extended to include multiple rare or less frequent tightly linked causal variants. Combining available T2D SNP association [Dupuis et al., 2010; Qi et al., 2010; Voight et al., 2010; Zeggini et al., 2008] and linkage [Guan et al., 2008; unpublished data] results suggest that at least 23 of 36 autosomal T2D loci are unlikely due to single or cumulatively rare disease variants.

METHODS

To understand the genetic architecture underlying a complex disease, we seek to identify a set of models that are plausible given prior results from ASP linkage and/or case-control association studies. We parameterize these models by the genotype relative risk (GRR_C) and risk allele frequency (RAF_C) of a causal variant C. We assume that the causal allele is the minor allele and is dominant, that the effects of the disease loci combine multiplicatively to determine disease risk, and that we have genotyped a sufficiently dense set of linkage markers that the identity by descent (IBD) relationship for the ASP can be observed. We assume a disease prevalence of 10%, and a standardized LD coefficient $D' = 0.6$, 0.8, or 1 between the causal variant C and a nearby genotyped variant M. We discuss the impact of these assumptions in the section Discussion.

POWER TO DETECT LINKAGE IN AN ASP STUDY

Let N_i be the number of ASPs sharing $i = 0$, 1, or 2 alleles IBD at the causal locus. Although the specified penetrance model at the causal locus is dominant, we calculate the usual additive model based maximum LOD score (MLS) [Risch, 1990]

$$\mathrm{MLS} = \begin{cases} (N_1+2N_2)\log_{10}\left(\frac{N_1+2N_2}{N}\right) & \\ +(N_1+2N_0)\log_{10}\left(\frac{N_1+2N_0}{N}\right) & \text{if } \frac{1}{2}N_1+N_2>N \\ 0 & \text{if } \frac{1}{2}N_1+N_2\le N \end{cases}$$

Given our assumption of a multiplicative relationship *between* causal loci, power to detect linkage using ASPs depends only on the locus-specific relative risks [Risch, 1990].

We calculate power to detect linkage for studies of $N =$ 500, 1,000, and 5,000 ASPs and causal variant RAF_C from 0.001 to 0.05. We report results for MLS threshold values of 0 and 1, representing no or modest evidence for linkage, respectively. For a given RAF_C, we determine the value of GRR_C that results in 95% power to obtain MLS > 0 or 1, using the false position method [Press et al., 1992], an algorithm for root finding.

POWER TO DETECT ASSOCIATION IN A CASE-CONTROL STUDY

We assume a GWAS with n cases and n controls. Let C be a causal variant in LD with a genotyped marker M. Let RAF_M be the risk allele frequency at M and g_C and g_M be the genotypes at C and M, coded as the number of risk alleles (0, 1, or 2). Given RAF_C, RAF_M, and D', we calculate the conditional genotype probability $P(g_C \mid g_M)$. For a specified genetic model and disease prevalence, we can then compute the penetrance of g_M as

$$P(Y|g_M)=\sum_{g_C} P(Y|g_M, g_C)\, P(g_C|g_M)=\sum_{g_C} P(Y|g_C)\, P(g_C|g_M) \tag{1}$$

which will determine the power of association test at a given locus. Under the dominant model, $P(Y = 1 \mid g_C = 2) = P(Y = 1 \mid g_C = 1) = GRR_C \times P(Y = 1 \mid g_C = 0)$. We exclude models with large GRR_C for which $P(Y \mid g_C) > 1$ for any genotype g_C. Here, we assume a single causal variant C, but our results can easily be extended to multiple causal variants in the same region (see section Discussion).

Although the specified penetrance model, $P(Y \mid g_C)$, is dominant, we test for disease association at M using the additive-model version of the Cochran-Armitage trend test, as is typical in analysis of GWAS data. We calculate the power of the trend test by estimating the variance of the test statistic under the alternative hypothesis [Freidlin et al., 2002].

We calculate power to detect association assuming $n =$ 1,000, 10,000, and 50,000 cases and the same number of controls and causal variant C and genotyped variant M frequencies $0.001 \le RAF_C \le 0.05$ and $0.05 \le RAF_M \le 0.95$. Given RAF_C and RAF_M, we calculate the GRR_C value that results in 5% power to detect disease association at M at genome-wide significance level $\alpha = 5 \times 10^{-8}$ using the false position method.

LD IN 1000 GENOMES PROJECT DATA

To assess the plausibility of our assumption that there exists a GWAS marker M in strong LD with the causal variant C, we evaluate the range of LD values between less common ($0.005 < MAF < 0.05$) and common ($MAF > 0.05$) chromosome 1 variants identified in 283 European samples in the 1000 Genomes Project August 2010 release (http://www.1000genomes.org/). We first examine the distribution of maximum r^2 values for each less common variant between the less common variant-common variant pairs, and then examine the D' values between the less common variant and common variant with the maximum r^2. We limit our attention to common variants within a 1000 SNP (~250 kb) window of each less common variant.

TABLE I. T2D susceptibility loci detected with common variants. RAF_C is a lower bound at which there is 5% power to detect association at observed *P*-value at marker M in a GWAS of the given effective sample size, and 95% power to detect linkage at the observed MLS value given 4,200 ASPs

SNP	Nearby gene(s)	OR	*P*-value	MLS	Effective sample size	RAF_M in controls	Minimum RAF_C	Maximum GRR_C	Reference[a]
rs1531343[b]	*HMGA2*	1.08	1.1×10^{-4}	1.00	68,314	0.10	<0.001	>4.1	Voight et al. [2010]
rs4607517[c]	*GCK*	1.07	5.0×10^{-8}	0.88	94,370	0.16	<0.001	>7.3	Dupuis et al. [2010]
rs8042680[b]	*PRC1*	1.06	1.6×10^{-6}	0.82	79,246	0.22	<0.001	>8.1	Voight et al. [2010]
rs10923931[b]	*NOTCH2*	1.11	1.9×10^{-3}	0.00	32,514	0.11	<0.001	>8.5	Zeggini et al. [2008]
rs7961581[b]	*TSPAN8,LGR5*	1.09	4.3×10^{-5}	1.22	31,364	0.27	<0.001	>9.8	Zeggini et al. [2008]
rs12779790[b]	*CDC123, CAMK1D*	1.09	1.5×10^{-4}	0.00	31,364	0.18	0.001	6.2	Zeggini et al. [2008]
rs4457053[b]	*ZBED3*	1.07	2.7×10^{-7}	0.06	67,214	0.26	0.002	6.0	Voight et al. [2010]
rs896854[b]	*TP53INP1*	1.05	2.2×10^{-5}	0.05	67,012	0.48	0.003	5.1	Voight et al. [2010]
rs11634397[b]	*ZFAND6*	1.05	1.2×10^{-5}	0.69	79,246	0.60	0.003	6.5	Voight et al. [2010]
rs10830963[c]	*MTNR1B*	1.09	8.0×10^{-13}	0.14	94,370	0.23	0.004	4.9	Dupuis et al. [2010]
rs4607103[b]	*ADAMTS9*	1.06	3.5×10^{-3}	0.24	31,364	0.76	0.004	5.0	Zeggini et al. [2008]
rs972283[b]	*KLF14*	1.06	6.4×10^{-6}	0.35	56,763	0.55	0.004	5.2	Voight et al. [2010]
rs2191349[c]	*DGKB/TMEM195*	1.06	1.1×10^{-8}	0.74	94,370	0.52	0.004	5.9	Dupuis et al. [2010]
rs864745[b]	*JAZF1*	1.10	1.3×10^{-7}	1.09	31,364	0.50	0.007	4.8	Zeggini et al. [2008]
rs7957197[b]	*HNF1A*	1.05	4.6×10^{-4}	0.05	67,751	0.85	0.009	3.4	Voight et al. [2010]
rs780094[c]	*GCKR*	1.06	1.3×10^{-9}	0.27	94,370	0.62	0.009	3.9	Dupuis et al. [2010]
rs340874[c]	*PROX1*	1.07	7.2×10^{-10}	0.00	94,370	0.52	0.010	3.0	Dupuis et al. [2010]
rs5215[d]	*KCNJ11*	1.09	1.6×10^{-5}	0.01	22,044	0.45	0.010	3.2	Voight et al. [2010]
rs243021[b]	*BCL11A*	1.08	6.2×10^{-11}	0.09	64,343	0.46	0.010	3.5	Voight et al. [2010]
rs1470579[d]	*IGF2BP2*	1.14	2.2×10^{-9}	0.20	22,044	0.30	0.010	3.6	Voight et al. [2010]
rs231362[b]	*KCNQ1*	1.07	3.2×10^{-9}	0.00	73,750	0.52	0.011	3.1	Voight et al. [2010]
rs9939609[d]	*FTO*	1.12	8.7×10^{-8}	0.14	22,044	0.38	0.011	3.4	Voight et al. [2010]
rs4430796[b]	*HNF1B*	1.12	1.6×10^{-4}	0.00	13,930	0.51	0.014	2.8	Voight et al. [2010]
rs7593730[b]	*RBMS1*	1.09	9.1×10^{-5}	0.01	32,172	0.77	0.021	2.5	Qi et al. [2010]
rs10010131[b]	*WFS1*	1.11	4.6×10^{-7}	0.05	22,044	0.60	0.026	2.5	Voight et al. [2010]
rs13292136[b]	*CHCHD9*	1.08	2.4×10^{-4}	0.00	79,246	0.93	0.029	2.3	Voight et al. [2010]
rs7578597[b]	*THADA*	1.12	9.2×10^{-5}	0.18	32,514	0.90	0.038	2.5	Zeggini et al. [2008]
rs7754840[d]	*CDKAL1*	1.18	3.1×10^{-15}	0.00	22,044	0.36	0.043	2.1	Voight et al. [2010]
rs7578326[b]	*IRS1*	1.10	2.2×10^{-15}	0.00	67,701	0.64	0.043	2.1	Voight et al. [2010]
rs13266634[b]	*SLC30A8*	1.15	1.5×10^{-8}	0.11	20,675	0.68	0.048	2.3	Voight et al. [2010]
rs1801282[d]	*PPARG*	1.15	8.0×10^{-6}	0.11	22,044	0.82	0.048	2.3	Voight et al. [2010]
rs1111875[d]	*HHEX*	1.17	9.1×10^{-15}	0.18	22,044	0.52	0.049	2.3	Voight et al. [2010]
rs7903146[d]	*TCF7L2*	1.40	2.2×10^{-51}	0.08	22,044	0.18	0.059	2.1	Voight et al. [2010]
rs11708067[c]	*ADCY5*	1.12	9.9×10^{-21}	0.00	94,370	0.78	0.090	1.8	Dupuis et al. [2010]
rs1552224[b]	*CENTD2*	1.14	3.2×10^{-18}	0.09	79,246	0.88	0.130	1.9	Voight et al. [2010]
rs10811661[d]	*CDKN2A/B*	1.19	1.4×10^{-10}	0.00	22,044	0.85	0.217	1.7	Voight et al. [2010]

[a] Each locus had a single or multistage *P*-value $< 5\times10^{-8}$. Some T2D-associated loci were reported in multiple references. We use results from the largest available follow-up cohort when possible (28 variants), or if not, from the largest available GWAS (which also contain the initial discovery samples) and list the estimated OR, *P*-values, and the effective sample sizes correspondingly.

[b] Follow-up sample.

[c] Top loci from a GWAS for fasting glucose were tested for association with T2D, equivalent to a candidate gene study.

[d] GWAS meta-analysis including the discovery samples.

APPLICATION TO TYPE 2 DIABETES (T2D)

We illustrate how existing ASP linkage and case-control GWAS results provide information on plausible models for variants underlying complex diseases using results for T2D. We carried out a joint analysis of data from 23 linkage studies as part of the International Type 2 Diabetes Linkage Analysis Consortium [Guan et al., 2008; unpublished data]. Here, we restrict our attention to an ASP linkage analysis of 6,552 individuals in 2,315 families of European ancestry, equivalent to ~4,200 ASPs, using the approximation that m-affected siblings correspond approximately to m-1-independent ASPs [Hodge 1984]. In this study, the largest MLS was approximately 2.2, and for approximately 54% of the genome, MLS = 0. For T2D linkage results, we calculate power as above based on 4,200 ASPs, using the observed MLS from the linkage study at that location as the MLS threshold for power calculations (unpublished data).

Published European ancestry association studies of T2D have identified 36 autosomal T2D loci using standard case-control analysis (Table I), most from GWAS. To place the T2D linkage and association results on the same map, we linearly interpolate positions for the 36 T2D-associated variants onto a genetic map of 2,164 microsatellite markers from our linkage analysis based on their physical positions in NCBI build 36.1. We identify the plausible genetic models at these 36 loci given the observed linkage and association results. At each locus, we calculate power to detect linkage and association as above. To minimize possible overestimation of genetic effect owing to the “winner’s curse” (for example Zöllner and Pritchard, 2007), we use results from the largest available follow-up cohort when possible (28 variants), or alternatively from the largest available GWAS (eight variants where discovery samples are ~40% of the total sample). For T2D association results, we calculate power as described above, using the sum of the effective numbers of genotyped cases and controls in each study as sample size, the observed RAF_M in controls as the population allele frequency, and the observed association *P*-value as the significance threshold.

RESULTS

Here, we address the range of plausible model parameters (RAF_C and GRR_C) for rare or less common causal variants ($RAF_C < 0.05$), assuming a dominant genetic model for a genomic region given results from prior linkage and/or (common variant) association studies. To do so, we compute the power to detect linkage and/or association as a function of genetic model.

RANGE OF PLAUSIBLE MODELS GIVEN NO OR MODEST EVIDENCE FOR LINKAGE

Complex disease linkage studies generally reveal no ($MLS = 0$) or modest ($MLS \leq 1$) evidence for linkage for most of the genome. We explore the range of genetic model parameters consistent with these observations. Figure 1 displays values for GRR_C that result in 95% power to observe $MLS > 0$ or $MLS > 1$ given analysis of $N = 500–5{,}000$ ASPs as a function of risk allele frequency RAF_C.

Assuming a causal variant exists, models (RAF_C and GRR_C) above the power curves in Figure 1A have ≥ 95% probability of showing at least some evidence for linkage ($MLS > 0$), and therefore such variants are unlikely to be present in a region with no evidence of linkage ($MLS = 0$). For example, given $N = 5{,}000$ ASPs, a causal variant with $RAF_C = 0.01$ and $GRR_C > 2.9$ or with $RAF_C = 0.05$ and $GRR_C > 1.9$ has ≥ 95% power to achieve $MLS > 0$, suggesting these models are unlikely at a locus with no evidence for linkage. Similarly, given $N = 5{,}000$ ASPs and $MLS = 1$, a causal variant with $RAF_C = 0.01$ is unlikely to have $GRR_C > 3.9$ (Figure 1B). As expected, all else being equal, the larger the linkage study sample, the more restricted the set of plausible models.

RANGE OF PLAUSIBLE MODELS GIVEN SIGNIFICANT EVIDENCE FOR ASSOCIATION

Genome-wide significant associations with common SNPs have been reported for many common diseases (http://www.genome.gov/gwastudies/). We explore the range of models (RAF_C and GRR_C) for which a disease association could be explained by a rare or less common causal variant(s). Figure 2 shows values of GRR_C that lead to 5% power to detect association ($P < 5\times10^{-8}$) at SNP M with $RAF_M = 0.05–0.95$, assuming a study of *n* cases

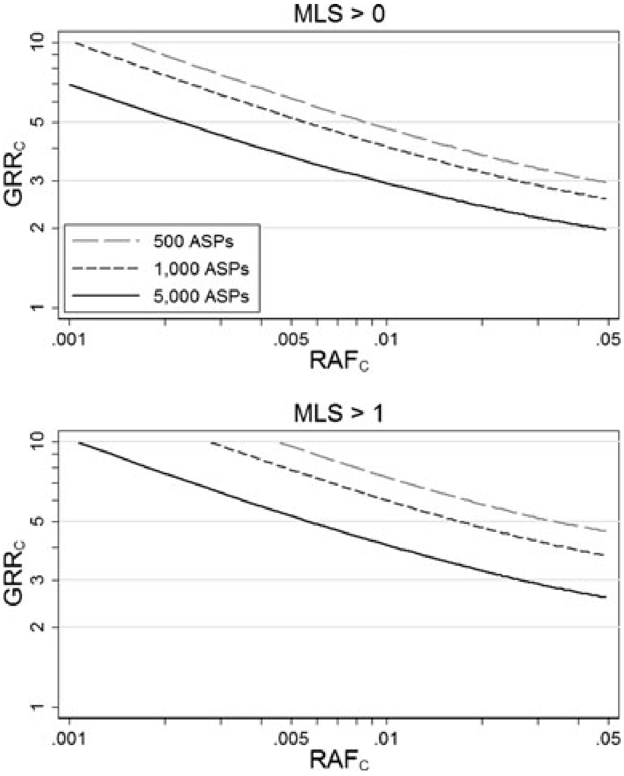

Fig. 1. Genotype relative risks at causal variant C (GRR_C) that result in 95% power to detect some evidence for linkage at MLS > 0 and MLS > 1.

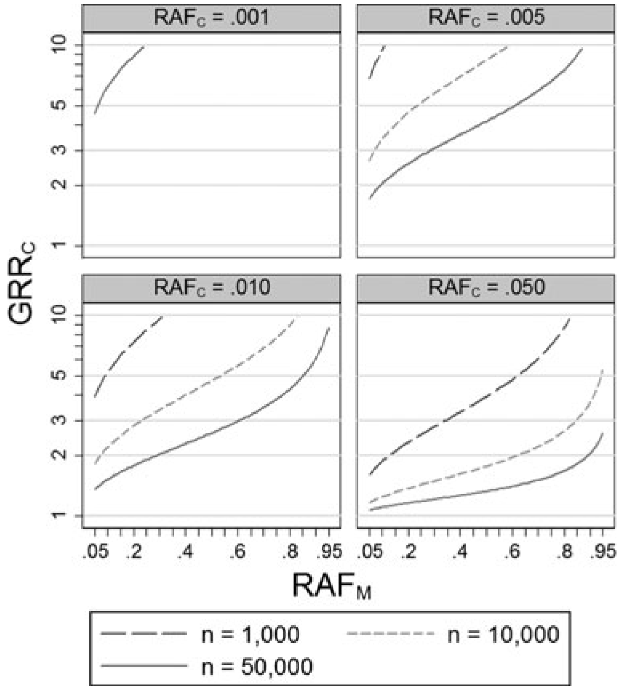

Fig. 2. Genotype relative risks at causal variant C (GRR_C) that result in 5% power to detect association ($P < 5\times10^{-8}$) at genotyped variant M using *n* cases and *n* controls. We assume disease prevalence 10% and D′ = 1 between M and C.

and n controls (n = 1,000–50,000) and D′ = 1 between the genotyped variant M and causal variant C. Models below the power curves have < 5% probability of achieving such evidence for association for a genotyped variant M in high LD (D′ = 1) with the causal variant C. We have chosen 5% power so that a causal variant C with a small chance of underlying a common variant M association (based on $\mathrm{GRR_C}$ and $\mathrm{RAF_C}$) will be considered as plausible, given the current 10s to 100s of associated loci for common diseases.

For example, given n = 10,000 cases and 10,000 controls, a causal variant with $\mathrm{RAF_C} = 0.01$ and $\mathrm{GRR_C} < 3.4$ has < 5% power to achieve genome-wide significance ($P < 5\times10^{-8}$) at a genotyped variant M (D′ = 1) with $\mathrm{RAF_M} = 0.3$, suggesting these models are unlikely to explain the corresponding association at M. Holding the significance level and sample size constant, a marker with larger $\mathrm{GRR_M}$ will yield a more limited set of plausible genetic models (Figure 2). We also estimate the plausible range of models assuming D′ = 0.8 or 0.6 between the causal variant C and genotyped variant M (Supporting information Figure 1). For a given significance level and $\mathrm{RAF_C}$, a causal variant with D′ < 1 requires larger $\mathrm{GRR_C}$ to reach the same power as a causal variant with D′ = 1, resulting in a more limited set of plausible models.

RANGE OF PLAUSIBLE MODELS GIVEN RESULTS FROM ASSOCIATION AND LINKAGE STUDIES

For complex diseases for which both linkage and association scans have been carried out, we observe no evidence for linkage (MLS = 0) in most regions of the genome, and some of these regions may contain genome-wide significant association results ($P < 5\times10^{-8}$). Figure 3A shows values of $\mathrm{GRR_C}$ that result in 5% power to detect association ($P < 5\times10^{-8}$) at genotyped variant M given n = 10,000 cases and n = 10,000 controls, and 95% power to detect at least some evidence for linkage (MLS > 0) given 1,000 ASPs, as a function of $\mathrm{RAF_C}$, assuming D′ = 1 between the causal and common GWAS variant. The models above the 5% power curve for association but below the 95% power curve for linkage (shaded area in Figure 3) are consistent with strong evidence for association ($P < 5\times10^{-8}$) and no evidence for linkage (MLS = 0) at the corresponding position. Here, significant evidence for association ($P = 5\times10^{-8}$) and no evidence for linkage (MLS = 0) suggest $\mathrm{RAF_C} > 0.014$.

Figure 3B and C shows values of $\mathrm{GRR_C}$ for D′ = 0.8 or 0.6 as a function of $\mathrm{RAF_C}$ that result in 5% power to detect association ($P < 5 \times 10^{-8}$) at genotyped variant M given n = 10,000 cases and n = 10,000 controls, and 95% power to detect some evidence for linkage (MLS > 0) given 1,000 ASPs. Again the shaded areas in the figures are consistent with strong evidence for association ($P < 5 \times 10^{-8}$) and no evidence for linkage (MLS = 0) at the corresponding position. In these scenarios, the range of plausible $\mathrm{RAF_C}$ values is more extensive than those for D′ = 1. For example, significant evidence for association ($P = 5 \times 10^{-8}$) and no evidence for linkage (MLS = 0) suggest $\mathrm{RAF_C} > 0.043$ for D′ = 0.6 and $\mathrm{RAF_C} > 0.022$ for D′ = 0.8 compared to $\mathrm{RAF_C} > 0.014$ for D′ = 1.

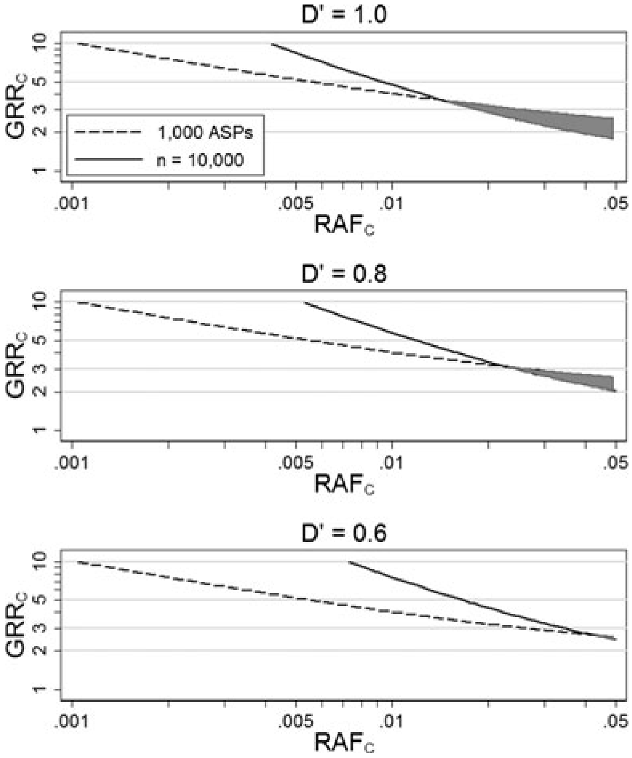

Fig. 3. Genotype relative risks at causal variant C ($\mathrm{GRR_C}$) that result in 95% power to detect some evidence for linkage (MLS > 0) using 1,000 ASPs and 5% power to detect association ($P < 5\times10^{-8}$) at genotyped variant M with $\mathrm{RAF_M} = 0.5$ using n = 10,000 cases and n = 10,000 controls. The shaded area is the estimated range of plausible models. We assume disease prevalence 10% and D′ = 1, 0.8, and 0.6 between M and C.

OBSERVED D′ AND r^2 IN 1000 GENOMES DATA

In practice, the LD between an unidentified causal variant C and a common associated variant M is unknown. To explore the LD between common and less common variants, we examine sequence data on 283 European subjects from the 1000 Genomes Project (http://www.1000genomes.org/). We calculate D′ and r^2 between 268,287 less common SNPs (0.005 < MAF < 0.05) and 423,648 common SNPs (MAF > 0.05) on chromosome 1, limiting our attention to pairs of SNPs within approximately 250 kb of each other.

Given a causal variant in the region, the most strongly associated GWAS variant is expected to be the common GWAS variant in highest r^2 with the causal variant. For every less common 1000 Genomes Project SNP, we identify the common 1000 Genomes SNP in highest r^2. We find that the best common pairing SNPs usually have $\mathrm{RAF_M} < 0.3$ (Figure 4). We also find that 49% of the maximum r^2 SNP pairs have D′ = 1, 67% have D′ ≥ 0.8, and 88% have D′ ≥ 0.6, which suggests that an assumption of D′ = 1 between the common associated variant and the causal SNP would cause the bounds on $\mathrm{RAF_C}$ to be too wide about half the time.

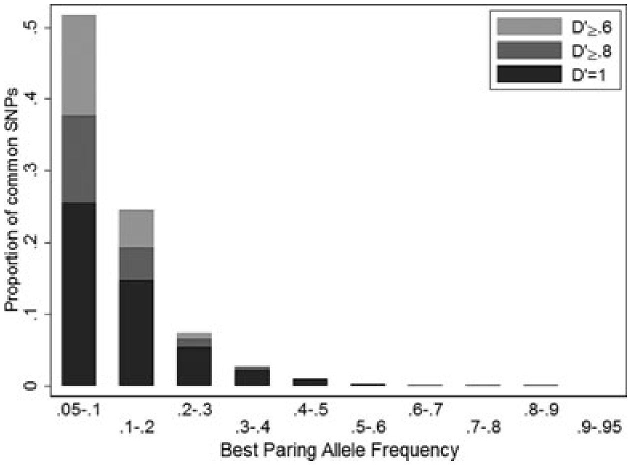

Fig. 4. Frequency distribution of best pairing alleles for less common ($0.005 < \mathrm{MAF} < 0.05$) variants in 1000 Genome Project sequence data (August 2010 release). We define the best pairing alleles as having the highest r^2 but lowest minor allele frequency.

EXAMPLE: T2D

Many linkage and association studies have been carried out for T2D. Perhaps the largest single linkage study was one based on the equivalent of 4,200 ASPs with European ancestry carried out by the International Type 2 Diabetes Linkage Analysis Consortium [Guan et al. 2008; unpublished data]. This study found no genome-wide significant evidence for linkage, and a maximum MLS genome-wide of approximately 2.2. In contrast, published GWAS and candidate gene association studies in European ancestry samples (through October 2011) have reported genome-wide significant association ($P < 5{\times}10^{-8}$) at 36 autosomal loci using the standard case-control test (Table I, Figure 5). For these 36 T2D loci, we observe that higher MLS are modestly correlated with lower $\mathrm{RAF_M}$ ($r = -0.31$, $P = 0.06$), suggesting that at least some of the linkage peaks may be detecting rare or less common underlying causal variants (Supporting information Figure 2). Likewise, 29 of the 36 T2D-associated SNPs are at positions with MLS > 0 (P-value = 0.0002 compared to an expectation of 50%, or P-value = 0.0009 compared to the observed proportion of 54%).

Using the observed T2D linkage and association results, we estimate the range of plausible $\mathrm{RAF_C}$ (Table I) assuming $D' = 1$ between the causal variant and a common GWAS variant. Thirteen association signals could plausibly be explained by a very wide range of risk allele frequencies $\mathrm{RAF_C}$ including $\mathrm{RAF_C} < 0.005$. Four of the five loci with the smallest plausible combined $\mathrm{RAF_C}$ (< 0.001) have modest evidence for linkage (0.82 < MLS < 1.22) and association $\mathrm{RAF_M}$ < 0.30. For the 23 other association signals, the associations are unlikely to be explained by one or more causal variants with combined $\mathrm{RAF_C} < 0.005$, and for four of these (*TCF7L2, ADCY5, CENTD2, CDKN2A/B*), combined $\mathrm{RAF_C}$ was less than 0.05 is unlikely. In these regions, a GWAS study with a dense marker set with good coverage for variants with MAF > 0.01 might well result in the causal variant being genotyped or tagged by genotyped markers.

DISCUSSION

We have sought to determine the range of disease models consistent with existing linkage and/or association results. Specifically, we have focused on determining the minimum plausible risk allele frequency $\mathrm{RAF_C}$ and corresponding genotype relative risk $\mathrm{GRR_C}$ for variants at a given locus assuming a single causal variant underlies an association signal. Our results show that a linkage study alone or an association study alone can restrict the plausible magnitude of $\mathrm{GRR_C}$, while all $\mathrm{RAF_C}$ in the range we consider remain possible. Joint consideration of linkage and association results can further reduce the set of plausible models. In particular, at loci with significant evidence for association and no evidence for linkage, one or more causal variant(s) with a low summed risk allele frequency may be implausible.

To calculate the power for linkage and association tests, we have made several assumptions. First, we assume that only a single causal variant C exists in the region of interest in our power calculation. If multiple causal variants are present within the region, the linkage signal will reflect the combined effects of all causal variants. Using linkage results alone, our estimates of the causal allele frequency would approximately correspond to the sum of the risk allele frequencies; individual causal variants could be much rarer. In contrast, the impact of multiple variants on a given common association signal is more complex as the observed signal will only reflect the causal variants in LD with the tested common allele. Wang et al. [2010] and Dickson et al. [2010] have described scenarios where multiple rare causal variants could contribute to an apparent common variant association, a phenomenon they termed “synthetic association” (Wang et al. [2010]). If all causal alleles occur on haplotypes with the associated common allele, the synthetic causal marker will have a $D' = 1$ with the common associated marker. In contrast, if the causal alleles occur on haplotypes with and without the associated allele, the synthetic causal marker will have a $D' < 1$ with the common associated marker. These two scenarios described above are analogous to the ones shown in Figure 3A and B. If P-value $= 5 \times 10^{-8}$ is observed for a common marker and the underlying synthetic causal marker has a $D' = 0.8$ with the common marker, analysis assuming a $D' = 1$ (Figure 3A)

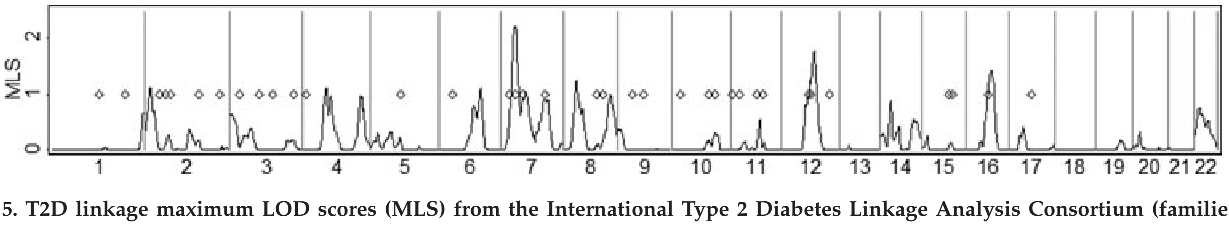

Fig. 5. T2D linkage maximum LOD scores (MLS) from the International Type 2 Diabetes Linkage Analysis Consortium (families of European origin) (solid line) and significant T2D associations from various sources (Table I) (diamonds).

will yield a lower estimate of the minimum plausible cumulative RAF_C than analysis under the true model of $D' = 0.8$ (Figure 3B) (RAF_C of 0.014 vs. 0.022 in this scenario). Thus, as in the case of a single causal marker, we will underestimate the lower bound of RAF_C assuming $D' = 1$. Estimates of minimum plausible cumulative RAF_C for multiple causal variants under different assumptions of D′ can be used to construct more realistic simulations of multiple rare variants RAF_C and GRR_C. This will aid in the estimation of the power of burden tests [e.g. Li and Leal, 2008] for given regions.

Second, we assume a dominant model at each disease locus. Since we focus on models with uncommon or rare risk variants, risk allele homozygotes are rare, and so dominant, additive, and multiplicative models are essentially equivalent. A recessive model would result in very rare homozygotes and is not considered. Third, we assume the minor allele of variant C is causal. If instead the minor allele is protective, associations detected with high-frequency RAF_M might be inconsistent with rare risk causal variants but consistent with rare protective variants. Fourth, for linkage, we assume fully informative markers. If the markers are not fully informative, our estimation results would give a smaller range of plausible models (higher RAF_C and lower GRR_C). All our assumptions, except for the assumptions of known IBD in linkage studies and no winner's curse for the association results (see below), are conservative in the sense that they should result in less strict bounds on our model parameters: the minimum plausible RAF_C would be higher and/or the maximum plausible GRR_C would be lower if the assumptions are violated.

We explore plausible models for 36 T2D variants identified by large-scale association studies in European ancestry populations in combination with results from a T2D linkage study based on the equivalent of approximately 4,200 ASPs. Our results suggest that 23 of the 36 association signals are unlikely to have been caused by causal variants with combined $RAF_C < 0.005$, and four of these are unlikely to have been caused by causal variants with combined $RAF_C < 0.05$. Multiple assumptions underlie these results. We assume that $D' = 1$ between the causal variant and the associated variant. This assumption will yield the widest range of plausible models, as it assumes that all of the causal alleles are on the same haplotype as the common risk allele. In the 1000 Genomes Project data, 51% of the maximum r^2 variant pairs have $D' < 1$ so our estimates for the plausible range of models may be too wide for these loci. For example, for the *CDC123/CAMK1D* locus the minimum plausible RAF_C is 0.001 with $D' = 1$ and 0.004 for $D' = 0.6$. The majority of the significance thresholds used in our calculations are based on results from follow-up samples. However, for eight variants the discovery samples make up approximately 40% of the effective association sample size and our results could be impacted by the "winner's curse." This could cause overestimation of the strength of the association, and thus our estimate of the minimum plausible RAF_C may be too high. This concern is balanced by our use of a fairly conservative 5% power to detect the observed association that may have caused us to underestimate the RAF_C and overestimate GRR_C for some loci. To explore the sensitivity of the minimum RAF_C and maximum GRR_C estimates to the set power thresholds, we repeated our analysis using 50% power for association (i.e. assuming the observed OR is the true effect size rather an overestimate of the true effect size) and 80% power for linkage (Supporting information Table 1). As expected, we found a greater number of loci that were inconsistent with the cumulatively rare causal variants. Specifically, we found that 30 (compared to 23) of the 36 association signals are unlikely to have been caused by causal variants with combined $RAF_C < 0.005$, and 14 (compared to four) of these are unlikely to have been caused by causal variants with combined $RAF_C < 0.05$. For one association signal (*CDKN2A/B*), no plausible model could be found under this assumption due to the large value of RAF_M and strong evidence for association but no evidence for linkage. Under either set of power assumptions, our results suggest that the causal variant(s) for many T2D loci may already have been detected by the 1000 Genomes and other sequencing projects. However, even for these loci, resequencing may be useful to identify other independent disease variants. For the other loci for which the summed frequency of causal variants may be < 0.005, sequencing studies may be particularly important for variant detection, since such uncommon variants may not have been identified in existing catalogues.

In summary, we estimate ranges of plausible genetic models based on results from association and/or linkage studies for complex diseases. Given no or modest evidence for linkage in a region of interest, we can estimate an upper bound on the GRR_C of potential rare or less common variants. Similarly, in the presence of association with a common genotyped variant, we can estimate a lower bound on the GRR for the causal variant. Taken together, significant evidence for association and no or modest evidence for linkage allow a joint estimate of a lower bound for RAF_C and upper bound for GRR_C. Our approach provides a useful starting point for modeling genetic architecture of complex diseases and has allowed us to identify T2D loci more likely to be caused by common variants. The knowledge of plausible genetic models for a given region will aid in estimating the power of burden tests [for example Li and Leal, 2010] for a given sample size and sequencing depth, and will allow more efficient design and interpretation of sequencing studies.

WEB RESOURCES

Software to carry out this sort of analysis is available (in Stata code) at http://www.biostat.umn.edu/∼wguan/software/.

ACKNOWLEDGMENTS

This research was supported by grants HG000376 (MB) and DK062370 (MB) from the National Institutes of Health. The authors have no conflict of interest.

REFERENCES

Anderson CA, Soranzo N, Zeggini E, Barrett JC. 2011. Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS Biol 9:e1000580.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create synthetic genome-wide associations. PLoS Biol 8:e1000294.

Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, Wheeler E, Glazer NL, Bouatia-Naji N, Gloyn AL, Lindgren

CM, Mägi R, Morris AP, Randall J, Johnson T, Elliott P, Rybin D, Thorleifsson G, Steinthorsdottir V, Henneman P, Grallert H, Dehghan A, Hottenga JJ, Franklin CS, Navarro P, Song K, Goel A, Perry JR, Egan JM, Lajunen T, Grarup N, Sparsø T, Doney A, Voight BF, Stringham HM, Li M, Kanoni S, Shrader P, Cavalcanti-Proença C, Kumari M, Qi L, Timpson NJ, Gieger C, Zabena C, Rocheleau G, Ingelsson E, An P, O'Connell J, Luan J, Elliott A, McCarroll SA, Payne F, Roccasecca RM, Pattou F, Sethupathy P, Ardlie K, Ariyurek Y, Balkau B, Barter P, Beilby JP, Ben-Shlomo Y, Benediktsson R, Bennett AJ, Bergmann S, Bochud M, Boerwinkle E, Bonnefond A, Bonnycastle LL, Borch-Johnsen K, Böttcher Y, Brunner E, Bumpstead SJ, Charpentier G, Chen YD, Chines P, Clarke R, Coin LJ, Cooper MN, Cornelis M, Crawford G, Crisponi L, Day IN, de Geus EJ, Delplanque J, Dina C, Erdos MR, Fedson AC, Fischer-Rosinsky A, Forouhi NG, Fox CS, Frants R, Franzosi MG, Galan P, Goodarzi MO, Graessler J, Groves CJ, Grundy S, Gwilliam R, Gyllensten U, Hadjadj S, Hallmans G, Hammond N, Han X, Hartikainen AL, Hassanali N, Hayward C, Heath SC, Hercberg S, Herder C, Hicks AA, Hillman DR, Hingorani AD, Hofman A, Hui J, Hung J, Isomaa B, Johnson PR, Jørgensen T, Jula A, Kaakinen M, Kaprio J, Kesaniemi YA, Kivimaki M, Knight B, Koskinen S, Kovacs P, Kyvik KO, Lathrop GM, Lawlor DA, Le Bacquer O, Lecoeur C, Li Y, Lyssenko V, Mahley R, Mangino M, Manning AK, Martínez-Larrad MT, McAteer JB, McCulloch LJ, McPherson R, Meisinger C, Melzer D, Meyre D, Mitchell BD, Morken MA, Mukherjee S, Naitza S, Narisu N, Neville MJ, Oostra BA, Orrù M, Pakyz R, Palmer CN, Paolisso G, Pattaro C, Pearson D, Peden JF, Pedersen NL, Perola M, Pfeiffer AF, Pichler I, Polasek O, Posthuma D, Potter SC, Pouta A, Province MA, Psaty BM, Rathmann W, Rayner NW, Rice K, Ripatti S, Rivadeneira F, Roden M, Rolandsson O, Sandbaek A, Sandhu M, Sanna S, Sayer AA, Scheet P, Scott LJ, Seedorf U, Sharp SJ, Shields B, Sigurethsson G, Sijbrands EJ, Silveira A, Simpson L, Singleton A, Smith NL, Sovio U, Swift A, Syddall H, Syvänen AC, Tanaka T, Thorand B, Tichet J, Tönjes A, Tuomi T, Uitterlinden AG, van Dijk KW, van Hoek M, Varma D, Visvikis-Siest S, Vitart V, Vogelzangs N, Waeber G, Wagner PJ, Walley A, Walters GB, Ward KL, Watkins H, Weedon MN, Wild SH, Willemsen G, Witteman JC, Yarnell JW, Zeggini E, Zelenika D, Zethelius B, Zhai G, Zhao JH, Zillikens MC; DIAGRAM Consortium; GIANT Consortium; Global BPgen Consortium, Borecki IB, Loos RJ, Meneton P, Magnusson PK, Nathan DM, Williams GH, Hattersley AT, Silander K, Salomaa V, Smith GD, Bornstein SR, Schwarz P, Spranger J, Karpe F, Shuldiner AR, Cooper C, Dedoussis GV, Serrano-Ríos M, Morris AD, Lind L, Palmer LJ, Hu FB, Franks PW, Ebrahim S, Marmot M, Kao WH, Pankow JS, Sampson MJ, Kuusisto J, Laakso M, Hansen T, Pedersen O, Pramstaller PP, Wichmann HE, Illig T, Rudan I, Wright AF, Stumvoll M, Campbell H, Wilson JF; Anders Hamsten on behalf of Procardis Consortium; MAGIC investigators, Bergman RN, Buchanan TA, Collins FS, Mohlke KL, Tuomilehto J, Valle TT, Altshuler D, Rotter JI, Siscovick DS, Penninx BW, Boomsma DI, Deloukas P, Spector TD, Frayling TM, Ferrucci L, Kong A, Thorsteinsdottir U, Stefansson K, van Duijn CM, Aulchenko YS, Cao A, Scuteri A, Schlessinger D, Uda M, Ruokonen A, Jarvelin MR, Waterworth DM, Vollenweider P, Peltonen L, Mooser V, Abecasis GR, Wareham NJ, Sladek R, Froguel P, Watanabe RM, Meigs JB, Groop L, Boehnke M, McCarthy MI, Florez JC, Barroso I. 2010. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 42:105–116.

Freidlin B, Zheng G, Li Z, Gastwirth JL. 2002. Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered 53:146–152.

Guan W, Pluzhnikov A, Cox NJ, Boehnke M, for the International Type 2 Diabetes Linkage Analysis Consortium. 2008. Meta-analysis of 23 type 2 diabetes linkage studies from the international type 2 diabetes linkage analysis consortium. Hum Hered 66:35–49.

Hodge SE. 1984. The information contained in multiple sibling pairs. Genet Epidemiol 1:109–122.

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

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

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311–321.

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

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753.

Press W, Teukolsky S, Vetterling W, Flannery B. 1992. Numerical recipes for C (2nd ed.). Cambridge: Cambridge University Press.

Pritchard JK, Cox NJ. 2002. The allelic architecture of human disease genes: common disease-common variant . . . or not? Hum Mol Genet 11:2417–2423.

Qi L, Cornelis MC, Kraft P, Stanya KJ, Linda Kao WH, Pankow JS, Dupuis J, Florez JC, Fox CS, Paré G, Sun Q, Girman CJ, Laurie CC, Mirel DB, Manolio TA, Chasman DI, Boerwinkle E, Ridker PM, Hunter DJ, Meigs JB, Lee CH, Hu FB, van Dam RM. Meta-Analysis of Glucose and Insulin-related traits Consortium (MAGIC); Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium. 2010. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum Mol Genet 19:2706–2715.

Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. Trends Genet 17:502–510.

Risch N. 1990. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241.

Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segrè AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Bengtsson Boström K, Bravenboer B, Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PR, Jørgensen T, Kao WH, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieverse A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JR, Petersen AK, Platou C, Proença C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shrader P, Sigurdsson G, Sparsø T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van Herpt T, van Vliet-Ostaptchouk JV, Walters GB, Weedon MN, Wijmenga C, Witteman J, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllensten U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Mohlke KL, Morris AD, Palmer CN, Pramstaller PP, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Wareham NJ, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Hu FB, Meigs JB, Pankow JS, Pedersen O, Wichmann HE, Barroso I, Florez JC, Frayling TM, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI; MAGIC investigators; GIANT Consortium. 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42:579–589.

Wang K, Dickson SP, Stolle CA, Krantz ID, Goldstein DB, Hakonarson H. 2010. Interpretation of association signals and identification of causal variants from genome-wide association studies. Am J Hum Genet 86:730–742.

Wray NR, Purcell SM, Visscher PM. 2011. Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol 9:e1000579.

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Boström KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jørgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjögren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ; Wellcome Trust Case Control Consortium, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D. 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.

Zöllner S, Pritchard JK. 2007. Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet 80:605–615.