Title

Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes.

Journal

American Journal of Human Genetics, 93(6)

Authors

Lohmueller, Kirk
Sparsø, Thomas
Li, Qibin
et al.

Publication Date

2013-12-05

DOI

10.1016/j.ajhg.2013.11.005

Peer reviewed

ARTICLE

Whole-Exome Sequencing of 2,000 Danish Individuals and the Role of Rare Coding Variants in Type 2 Diabetes

Kirk E. Lohmueller,[1,18,19] Thomas Sparsø,[2,18] Qibin Li,[3] Ehm Andersson,[2] Thorfinn Korneliussen,[4] Anders Albrechtsen,[5] Karina Banasik,[2] Niels Grarup,[2] Ingileif Hallgrimsdottir,[6] Kristoffer Kiil,[2] Tuomas O. Kilpeläinen,[2] Nikolaj T. Krarup,[2] Tune H. Pers,[7,8,9] Gaston Sanchez,[6] Youna Hu,[1] Michael DeGiorgio,[1,20] Torben Jørgensen,[10,11,12] Annelli Sandbæk,[13] Torsten Lauritzen,[13] Søren Brunak,[7] Karsten Kristiansen,[3,5] Yingrui Li,[3] Torben Hansen,[2,14] Jun Wang,[2,3,5] Rasmus Nielsen,[1,5,15,*] and Oluf Pedersen[2,16,17,*]

It has been hypothesized that, in aggregate, rare variants in coding regions of genes explain a substantial fraction of the heritability of common diseases. We sequenced the exomes of 1,000 Danish cases with common forms of type 2 diabetes (including body mass index > 27.5 kg/m^2 and hypertension) and 1,000 healthy controls to an average depth of 56×. Our simulations suggest that our study had the statistical power to detect at least one causal gene (a gene containing causal mutations) if the heritability of these common diseases was explained by rare variants in the coding regions of a limited number of genes. We applied a series of gene-based tests to detect such susceptibility genes. However, no gene showed a significant association with disease risk after we corrected for the number of genes analyzed. Thus, we could reject a model for the genetic architecture of type 2 diabetes where rare nonsynonymous variants clustered in a modest number of genes (fewer than 20) are responsible for the majority of disease risk.

Introduction

Twin and segregation studies have suggested that complex diseases, such as common metabolic disorders, are determined, in part, by genetic factors.[1] As a result, over the last several decades there has been tremendous interest in identifying the genetic basis of common diseases.[2,3] Recently, researchers have used genome-wide association studies (GWASs) to identify common variants that increase risk of common disease. Hundreds of reproducible associations have been reported between common single SNPs and particular traits. Some of these associations have yielded novel biological insights that will be useful for biomedical research.[4]

However, it is now well documented that most of the identified loci have very small effect sizes.[5] Despite their relatively moderate frequency in the population, the common variants associated with complex traits, to date, can only account for a small amount of the heritability that has been estimated for these traits through twin and familial-aggregation studies.[6] This discrepancy between the heritability explained by the common SNPs identified in GWASs and familial studies has been termed the "missing heritability" problem.

Presently, researchers are searching for the missing heritability in a number of places.[6–8] One such location that is the subject of much current research is in low-frequency and rare (frequency $< 1\%$) genetic variants.[9,10] Population genetic theory suggests that if disease-causing variants are affected by purifying natural selection because they lead to a slight decrease in reproductive fitness in the individuals carrying them, then a greater proportion of the heritability will be explained by low-frequency and rare variants than by common variants.[11,12] Furthermore, low-frequency variants have probably eluded detection in currently used GWASs. There are two reasons for this. First, the currently used genotyping arrays are biased against the inclusion of low-frequency variants. Thus, many low-frequency variants are never directly tested for an association with the trait. Second, low-frequency variants are not tagged by the common variants genotyped in the GWAS. As a result, they would also escape indirect

[1]Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, USA; [2]The Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, 2100 Copenhagen, Denmark; [3]BGI-Shenzhen, Yantian District, 518083 Shenzhen, China; [4]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350 Copenhagen, Denmark; [5]Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark; [6]Center for Theoretical Evolutionary Genomics, University of California, Berkeley, Berkeley, CA 94720, USA; [7]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Lyngby, Denmark; [8]Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA; [9]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA 02115, USA; [10]Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark; [11]Faculty of Medicine, University of Aalborg, 9220 Aalborg, Denmark; [12]Research Center for Prevention and Health, Glostrup University Hospital, 2600 Glostrup, Denmark; [13]Department of Public Health, Section for General Practice, Aarhus University, 8000 Aarhus, Denmark; [14]Faculty of Health Sciences, University of Southern Denmark, 5230 Odense M, Denmark; [15]Department of Statistics, University of California, Berkeley, Berkeley, CA 94720, USA; [16]Institute of Biomedical Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark; [17]Faculty of Health Sciences, Aarhus University, 8000 Aarhus, Denmark
[18]These authors contributed equally to this work
[19]Present address: Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA 90095, USA
[20]Present address: Department of Biology, Pennsylvania State University, 502 Wartik Laboratory, University Park, PA 16802, USA
*Correspondence: rasmus_nielsen@berkeley.edu (R.N.), oluf@sund.ku.dk (O.P.)
http://dx.doi.org/10.1016/j.ajhg.2013.11.005.

detection because they are not well correlated with a typed common SNP.

Although there have been several reported associations between low-frequency variants and complex traits,[13–18] the hypothesis that rare variants account for a large proportion of the heritability of complex traits remains to be tested. With the advent of next-generation sequencing[19] and the affordability of exome sequencing, researchers are gaining the genomic tools with which to discover and test for associations between rare coding variants and complex disease and directly test the rare-variant common-disease hypothesis. Such approaches have recently been successfully applied to the identification of new mutations responsible for Mendelian diseases.[20–25] Additionally, over the past several years, advances have been made on the analysis side. The development of numerous statistical tests has allowed more efficient testing for associations between rare variants within a particular gene and a trait.[26–31] Such methods seek to combine the signal from multiple markers within a gene to provide greater statistical power than that for single-marker tests. However, these methods have yet to be widely applied to exome sequencing data from thousands of individuals. As such, their overall performance remains to be determined.

One common disease that has been subjected to intense genetic study is type 2 diabetes.[32] The heritability of type 2 diabetes has been estimated to be around 30%.[33–35] Through GWASs, 63 loci have been reproducibly associated with type 2 diabetes.[36] However, as for other complex traits, the associated SNPs can only account for <20% of the heritability estimated from family studies.[36]

Here, we seek to evaluate the role that rare coding variants play in the genetic basis of common forms of type 2 diabetes. We performed a deep whole-exome sequencing study of 2,000 Danish individuals. We applied both single-marker and gene-based association tests. Although we failed to detect any significant association after multiple test corrections, our simulations suggest that our results are informative about the genetic architecture of type 2 diabetes. In particular, our study suggests that when clustered in a small number of genes, rare coding variants of moderate to strong effect are unlikely to account for much of the missing heritability. Rather, if rare coding variants are an important factor in type 2 diabetes risk, they are most likely scattered across many genes. Our results have important implications for the design and interpretation of future medical resequencing studies.

Subjects and Methods

Study Populations

We sequenced 2,000 Danish individuals, of which half (the cases) suffered from type 2 diabetes,[37] moderate adiposity (body mass index [BMI] > 27.5 kg/m^2), and hypertension (systolic/diastolic blood pressure [BP] > 140/90 mmHg or use of antihypertensive medication). The others were healthy individuals who all had fasting plasma glucose < 5.6 mmol/l, 2-h OGTT-based plasma glucose < 7.8 mmol/l, BMI < 27.5 kg/m^2, and BP < 140/90 mmHg (and no antihypertensive treatment). Clinical and biochemical characteristics of the 2,000 individuals involved are described in Table S1 in the Supplemental Data available with this article online. The 2,000 sequenced individuals were selected from three different Danish study populations (Inter99, Steno samples, and ADDITION [Anglo-Danish-Dutch Study of Intensive Treatment in People with Screen-Detected Diabetes in Primary Care]) as previously reported.[18]

Inter99

The Inter99 cohort is a randomized, nonpharmacological intervention study for the prevention of ischemic heart disease and was conducted on 6,784 randomly ascertained participants aged 30–60 years at the Research Centre for Prevention and Health in Glostrup (ClinicalTrials.gov ID NCT00289237). An oral glucose tolerance test (OGTT) measured plasma glucose and serum insulin at fasting and 30 and 120 min after glucose intake. Subsequently, 6,094 participants who were of Danish nationality and had available DNA were classified as having normal glucose tolerance (n = 4,525), impaired fasting glycaemia (n = 504), impaired glucose tolerance (n = 693), screen-detected type 2 diabetes (n = 253), or previously diagnosed type 2 diabetes (n = 119) according to the World Health Organization (WHO) 1999 criteria. Detailed characteristics of Inter99 have been published previously.[38,39]

Steno

A sample of individuals with clinical-onset type 2 diabetes and a group of nondiabetic control individuals were ascertained at the outpatient clinic at Steno Diabetes Center, Copenhagen. An OGTT was performed in all control individuals so that individuals with unknown diabetes or states of prediabetes according to WHO 1999 criteria could be excluded.[37]

ADDITION

The Danish ADDITION Study is a general-practice type 2 diabetes high-risk screening and intervention study sampled by the Department of General Practice at the University of Aarhus (ClinicalTrials.gov ID NCT00237548).[40] The 8,662 Danish participants with available DNA from the initial screening cohort included 1,626 participants with screen-detected and untreated type 2 diabetes and 7,036 nondiabetic subjects. Individuals with type 2 diabetes were diagnosed by two independent diabetic plasma glucose values at baseline investigation or at a 1-year follow-up investigation.

All study participants gave informed consent for use of their biological samples for genetic studies. The current research protocol was approved by The Danish National Ethical Committee on Health Research and is in accordance with the ethical scientific principles of the Helsinki Declaration II.

Exome Capture and Sequencing

Concentration and quantity of genomic DNAs (gDNAs) were measured by Qubit Fluorometer (Invitrogen). Agarose gel electrophoresis was employed for checking whether gDNA was degraded. Only samples without apparent degradation and quality > 3 μg were retained. DNA from each sample was broken into short fragments ranging from 150 to 200 bp. The resulting fragments were end repaired, ligated with adaptors, indexed (6 bp), and amplified by adaptor-mediated PCR (pre-PCR). After purification, the amplified fragments were hybridized to the SureSelect biotinylated RNA baits with the Agilent SureSelect All Exon Kit v.2 (with a 46 Mb target region). After a 24 hr hybridization, nonhybridized

fragments were washed away. The hybridized fragments were amplified for the production of a sequencing library, which was sequenced with an Illumina HiSeq 2000 machine. Case and control samples were randomized and sequenced in index tagged pools of four (two random cases and two random controls).

Alignment

From the fastq files generated from the Illumina pipeline, all samples were aligned to the GRCh37/hg19 human reference genome (UCSC Genome Browser) with the Burrows-Wheeler Aligner (BWA; v.0.5.8).[41] No alternative haplotypes or random chromosomes (chromosome parts without a known position) were used. All reads were mapped to the positive strand of hg19. Duplicate reads were removed.

Below are the BWA commands used for the alignments to the GRCH37/hg19 reference genome:

- bwa aln -n 3 -o 1 -e 10 -i 5 -l 32 -t 4 hg19.fa read1.fq -f read1.sai
- bwa aln -n 3 -o 1 -e 10 -i 5 -l 32 -t 4 hg19.fa read2.fq -f read2.sai
- bwa sampe -a 2000 hg19.fa read1.sai read2.sai read1.fq read2.fq | samtools view -C -S –b

Extended Target

Sites flanking the target region were also covered by sequencing reads. Thus, we decided to enlarge the target regions by 100 bp on each side (termed the "extended target region"). Compared to the original target, which covered only 46,205,397 bp of the hg19 reference genome, the extended target regions covered 82,207,242 bp.

SNP Discovery and Genotype Calling without Imputation

For all samples and for every position of the autosomes given by the extended target region (~82 Mb), we calculated genotype likelihoods for all ten possible genotypes by using SAMtools v.0.1.8.[42] From the raw likelihoods, we determined the major allele frequency, minor allele frequency (MAF), and maximum likelihood estimate.[43,44] The likelihood was then used for SNP detection. Specifically, for each site, we estimated the likelihood of the data under the alternative model by allowing for two alleles (our optimized likelihood) and under a null model in which only one allele was present. This gave us a likelihood ratio test statistic (LRT) that was χ^2 distributed with one degree of freedom and could be converted to a p value. We defined a p value of 10^{-6} (LRT $>$ 24) as our cutoff for putatively variable sites of interest. These analyses were done with ANGSD (Analysis of Next Generation Sequencing Data) software.[45]

On the basis of the genotype likelihoods calculated from SAMtools, we called genotypes for all 1,998 samples for 2,958,319 sites with a MAF $>$ 0.0001 (1,354,315 in the target regions and 1,604,004 in the extended target regions) by using the approach devised in Kim et al.,[43] where the genotype with the highest likelihood was assigned to each individual. Specifically,

$$G_{hwe} = \text{argmax}_{(g \in \{0,1,2\})} \left\{ \binom{2}{g} f^g (1-f)^{2-g} L(D \mid G = g) \right\},$$

where G_{hwe} is the called genotype for the individual, g is the number of copies of the minor allele, f is the allele frequency estimated at that site, and $L(D|G = g)$ is the genotype likelihood from SAMtools.

Filtering of Sites and Samples

After initial genotype calls, we applied a series of site filters to the 2,958,319 sites with a MAF $>$ 0.0001 to obtain a set of sites with high-quality genotype calls suitable for association analysis (see Appendix A). We also filtered individuals for data quality (see Appendix B).

Genotype Calling Using BEAGLE

We used BEAGLE[46] to impute genotypes from sequencing data on 2,958,319 sites with an estimated MAF $>$ 0.0001. We performed BEAGLE imputation with default parameters by using genotype likelihoods from SAMtools in the following stepwise manner:

1. From our sequencing data, we imputed genotypes for sites both in the target and extended regions and in the 1000 Genomes project[47] ($\pm$10 kb, at most 15 SNPs) by using the 1000 Genomes CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) haplotypes as a reference panel (~4.2 million sites).
2. We imputed genotypes from our exome sequencing data without using a reference panel (2,958,319 sites with an estimated MAF $>$ 0.0001).
3. We combined the two sets of calls. For sites present in the 1000 Genomes SNP set, we adopted imputed genotypes from step 1. For sites only present in the exome sequencing SNP set (not present in 1000 Genomes), we adopted the imputed genotypes from step 2.

Variable sites were retained if they were successfully imputed ($r^2 > 0.2$) and if they passed all previously applied filters. Imputed genotypes were used for association analyses throughout the paper.

Annotation of Putative Variable Sites

The SeattleSeq Annotation 137 server was used for annotation of all sites. If a variable site was given two or more annotations because there were different isoforms, the most functional annotation was used. In other words, annotations were ranked as nonsense $>$ splice $>$ nonsynonymous $>$ synonymous $>$ outside coding.

Validation of Variants

For validation purposes, we performed in-house Sanger sequencing of all singletons (n = 31) and doubletons (n = 15) identified via our exome sequencing study in seven genes for maturity-onset diabetes of the young (MIM 606391) and monogenic obesity. All gene segments were amplified by standard PCR and directly sequenced by Sanger sequencing. All primers were designed with Primer3. The sequences were analyzed on a 3130XL Genetic Analyzer (Applied Biosystems), and mutations were detected with SeqScape v.2.5 (Applied Biosystems).

Metabolic Gene Sets, Pathways, and Networks

We defined gene sets for monogenic diabetes, obesity, and hypertension, as well as sets of genes identified by GWASs of type 2 diabetes, BMI, and hypertension (see Table S2 for lists of genes in each set).

We next constructed protein-protein interaction networks. To do this, we selected 76 genes known from monogenic forms of diabetes, obesity, and hypertension or GWAS hits (type 2 diabetes, obesity, and hypertension) for which the lead association lies within the protein-coding part of the gene (Table S3).

Protein subnetworks were constructed with the InWeb protein-protein interaction database v.3.[48] InWeb comprises more than 960,000 experimentally derived interactions. We discarded low-confidence interactions and focused our analysis on a network consisting of 170,000 high-confidence interactions (defined as interactions identified in multiple independent studies and often reported in small-scale experiments rather than in high-throughput studies). For each gene of interest, we defined protein subnetworks by the proteins reported to directly interact with the given gene's product. To further restrict the analysis to interactions likely to underlie type 2 diabetes, obesity, or hypertension, we constructed tissue-specific protein subnetworks by pruning away proteins whose encoding genes' mRNA expression levels are below the median gene expression level of the assigned tissue (Table S3). We obtained tissue-specific gene expression data from the BioGPS database[49] and manually assigned a single most likely tissue to each gene of interest.

Association Tests

For single-marker association tests, we applied an allele-based χ^2 test (implemented in PLINK[50]) to test for an association between case-control status and each of the identified putatively functional variants (annotated as nonsynonymous, splice-site, or nonsense SNPs).

Because single-marker tests have suboptimal power to detect genes with many rare variants affecting the trait,[30,51] we also employed several gene-based tests that combine information across multiple variants within each gene. Specifically, we applied the SKAT (Sequence Kernel Association Test),[29] KBAC (Kernel-Based Adaptive Cluster),[27] WSS (Weighted Sum Statistic),[28] VT (Variable Threshold),[52] Score,[31] SSU (Sum of Squared Score),[31] SSUw (Weighted Sum of Squared Score),[31] and Sum[31] tests. SKAT was implemented with the R package described in Wu et al.[29] To model the relationship between genetic variants and disease status, we ran SKAT by using both the linear kernel (defined as $SKAT_1$) and the identity-by-state (IBS) kernel (defined as $SKAT_2$). Additionally, we ran SKAT with two different weighting schemes. First, we gave all SNPs equal weight ($SKAT_1$). Second, we used the default weights, which give extra weight to SNPs with low MAF ($SKAT_2$). In the first case, SKAT has been shown to be equivalent to the C-alpha test.[29] We ran KBAC by using the KBAC R package. To assess significance for each test, we used 100,000 permutations and the adaptive approach to increase speed. We ran the WSS method[28] by using the AssotesteR package with 500 permutations to assess significance. Only SNPs with a MAF < 1% were used for this test. For genes with initial p values ≤ 0.002, we ran an additional 50,000 permutations. One gene still had $p = 0$, so we ran 500,000 permutations for that gene. We ran the VT method[52] by using the AssotesteR package with 1,000 permutations to assess significance. Only SNPs with a MAF < 5% were used for this test. For genes with initial p values ≤ 0.001, we ran an additional 50,000 permutations. We ran the other four gene-based tests by using the R code described in Pan et al.[31] We assessed significance by using 1,000 permutations. For the genes with initial p values equal to 0, we ran an additional 100,000 permutations. We focus on the SKAT test throughout much of this paper because it has been shown to have high statistical power under a variety of conditions.[29]

Additionally, all analyses using the gene-based tests were restricted to include only putatively functional variants, which we defined to be nonsynonymous, splice-site, or nonsense SNPs. Three different frequency thresholds for including SNPs were used: (1) all SNPs regardless of frequency, (2) MAF < 5%, and (3) MAF < 1%. All analyses described in the paper used the SNPs with MAF < 5% unless otherwise noted.

It has been suggested that some genes might have too few SNPs to enable detection of any statistically significant associations.[53] Including such genes in the analyses could reduce power by increasing the number of tests performed and, consequently, the stringency of the correction for multiple tests. To mitigate this effect, we only analyzed genes with at least two SNPs meeting the inclusion criteria described above. For the 5% MAF threshold, 15,133 genes met this criterion, suggesting a Bonferroni-corrected threshold of 3×10^{-6} for a genome-wide 5% significance level. We also restricted some of the gene-based analyses to only include genes containing at least five or ten SNPs under a 5% MAF threshold. A total of 11,347 or 6,105 genes (containing at least five or ten SNPs, respectively) met these criteria, giving 5% Bonferroni significance thresholds of 4×10^{-6} and 8×10^{-6}, respectively.

Statistical Power Simulations

To assess the power of our exome sequencing study, we performed power simulations. We explored different values of the total heritability of diabetes risk (on the liability threshold scale) that could be explained by rare nonsynonymous variants in a number of genes. We conditioned these simulations on the observed patterns of genetic variation within our exome sequencing data and then assigned them effects on the trait to generate the desired heritability. Although this approach differs from traditional power simulations, which fix the effect sizes and allele frequencies without regard to the heritability, our simulation approach allows a more direct investigation of the genetic architecture that is compatible with our empirical results (also see Long and Langley[54] for a similar simulation approach).

We assumed that the total narrow-sense heritability of type 2 diabetes risk (h^2) is 0.3 (which is likely to be an underestimate[33–35,55]; Table S4). We then assumed that this heritability can be divided among coding variants in $n \in \{5, 10, 15, 20, 50, 100, 150, 200, 500\}$ different risk genes. Under the assumption of n different genes, the amount of heritability contributed by each gene was $h_n^2 = 0.3/n$. We also varied the causal proportion of SNPs within each gene as $c \in \{0.25, 0.5, 1.0\}$. Within each simulation replicate, each nonsynonymous SNP with a MAF < 5% had a probability c of being retained as a causal SNP. If a gene had fewer than $2/c$ total SNPs (both causal and noncausal), it was discarded before the simulation started. This step increased the efficiency of our simulations by retaining genes that were expected to carry at least two causal SNPs.

For a given gene sampled from our data set and values of h_n^2 and c, we simulated cases and controls by conditioning on the genotypes in our data. We did this by using the –simu-cc function of the Genome-wide Complex Trait Analysis (GCTA) package.[56] Specifically, for each causal SNP i, we drew the SNP effect, α_i, from a standard normal distribution. Let $\mathbf{W}$ denote the normalized genotype matrix for all 1,965 individuals in the study at the causal SNPs. Each entry in this matrix is thus $w_{ij} = x_{ij} - 2p_i/\sqrt{2p_i(1-p_i)}$, where $x_{ij} \in \{0, 1, 2\}$ is the genotype for the j^{th} individual at the i^{th} SNP and p_i is the allele frequency in our 1,965 individuals.[56]

Individual quantitative phenotypes were assigned by a standard linear model,

$$y_j = \sum_{\text{all causal SNPs}} w_{ij}\alpha_i + \varepsilon_j,$$

where y_j is the (quantitative) phenotype of the j^{th} individual, w_{ij} is the genotype of the j^{th} individual at the i^{th} SNP, α_i is the effect of the i^{th} SNP, and ε_j is the environmental effect (see below). In matrix notation, this can be expressed as $\boldsymbol{y} = \mathbf{Wu} + \varepsilon$. The environmental variance was assigned such that the proportion of the phenotypic variance attributable to genetic variation was equal to h_n^2. Let $\sigma_{G,n}$ be the empirical variance of $\mathbf{Wu}$. Then, the environmental variance for individual j is drawn from a normal distribution with a mean of 0 and a variance of

$$\sigma_{G,n}\left[\frac{1}{h_n^2} - 1\right].$$

Intuitively, this takes the genetic variance actually found in the sample $\sigma_{G,n}$ and then sets the environmental variance such that

$$h_n^2 = \frac{\sigma_{G,n}}{\sigma_{G,n} + \sigma_E}.$$

Note that the GCTA approach used the empirical variance of $\mathbf{Wu}$. Essentially equivalent results can be found with the classic quantitative genetic equation

$$\sigma_{G,n} = \sum_{\text{all causal SNPs}} 2p_i(1-p_i)\alpha_i^2.$$

The 981 individuals with the highest values of y were considered to be cases, and the other 984 were the controls.

Our simulations assigned SNP effect sizes such that the observed patterns of genetic variation explained the desired heritability. One potential drawback to conditioning on the observed patterns of variation is that the effect sizes assigned can be unrealistically large, especially when a given gene (that might only contain a small number of rare variants) accounts for much of the heritability. Thus, we implemented an extra rejection step into our power simulations. In order to simplify the discussion of the rejection step, we rescaled the variances so that the total phenotypic variance equaled 1. We did this by finding the value of C such that $C[\sigma_{G,n} + \sigma_E] = 1$. Then, we let $\kappa_i = C\alpha_i$. In other words, κ_i was the normalized SNP effect (still on the liability scale). If $\kappa_i > 3$ for any SNP within a gene, we rejected that gene and selected a different one. This procedure increased the probability of including genes containing either common SNPs (still with a MAF $< 5\%$) or more SNPs.

The threshold of rejecting genes containing a SNP with $\kappa_i > 3$ was chosen for the following reason. Assume a liability threshold model (liability follows a normal distribution) in which an individual whose liability is >1 has the disease. Further assume that there is a single causal variant. The liability of individuals who do not carry the causal variant follows a standard normal distribution. Thus, roughly 16% of individuals not carrying any risk variants would have the disease. Under the assumption that a single causal SNP i has an effect $\kappa_i = 3$ (the cutoff we used) per allele copy on the liability scale, the liability of heterozygous individuals would be normally distributed with a mean equal to 3. Then, 97.8% of individuals who carry the causal SNP as a heterozygote would have the disease (i.e., have a risk score > 1). Thus, the risk variant is almost fully penetrant, and further increasing κ_i would not substantially increase the penetrance of the risk allele. Thus, requiring SNPs to have $\kappa_i < 3$ is biologically reasonable and was the upper bound on the effect size used in our power simulation.

We then ran SKAT and KBAC on the simulated phenotypes and the actual genotypes and recorded the p values. This process was repeated until we had 1,000 simulation replicates for each combination of values of n and c. The proportion of simulation replicates with p values less than the specified significance level was the power of the test.

The simulations described above evaluate the SKAT test's power to detect a single association. But, under the polygenic models of disease risk, there are many genes (n of them) that contribute to disease risk. Although the effect of an individual gene under this model becomes smaller, making the gene harder to detect, there are more opportunities to detect a truly associated gene. Therefore, we also evaluated the power to detect at least one associated gene at our Bonferroni significance threshold. The power to detect at least one gene was calculated as $P(\text{detect} \geq \text{one gene}) = 1 - \beta^n$, where β is the type II error, or the proportion of simulation replicates that were not significant under our Bonferroni threshold, and n is the number of risk genes in the particular model.

Results

Description of Variants Found and Data Quality

The fraction of the target region covered by various depths is shown in Figure S1. The median and mean depth of coverage were 46× and 56.3×, respectively. The exome coverage (the percent of targeted bases that were covered by at least one read) ranged from 94.11% to 98.76%. The average exome coverage per sample was 97.27% (SD = 0.00595; Figure S2). No sample had less than 90% of its exome covered.

For general quality assessment of the sequencing, we used nonimputed genotype calls for the 729,538 variants identified and retained after site filtering and application of the LRT score > 24 (Appendix A). The majority of variants present in the data set had a low minor allele count, as expected. There were no general differences between cases and controls in the frequency patterns (Figure S3). Additionally, we found no bias in the transition-to-transversion (Ti/Tv) ratios for different bins of minor allele counts (Figure S4), total depth (Figure S5), or LRT statistics (Figure S6). Furthermore, we found few differences in the Ti/Tv ratios, average total depths, and average LRT scores between rare SNPs (minor allele count < 10) absent from and those already present in dbSNP v.134 (Table S5).

As expected, the minor allele count differed across annotation classes in that functional variants had lower MAFs (Figure S7). We also stratified the Ti/Tv ratios into different annotation classes, both for all individuals and for cases and controls separately (Table S6). In some annotations, we detected a slight decrease in the Ti/Tv ratio in cases compared to in controls. However, the Ti/Tv ratios for the different annotation categories were generally in proximity to what others have reported.[53] These findings suggest that the data are of sufficient quality.

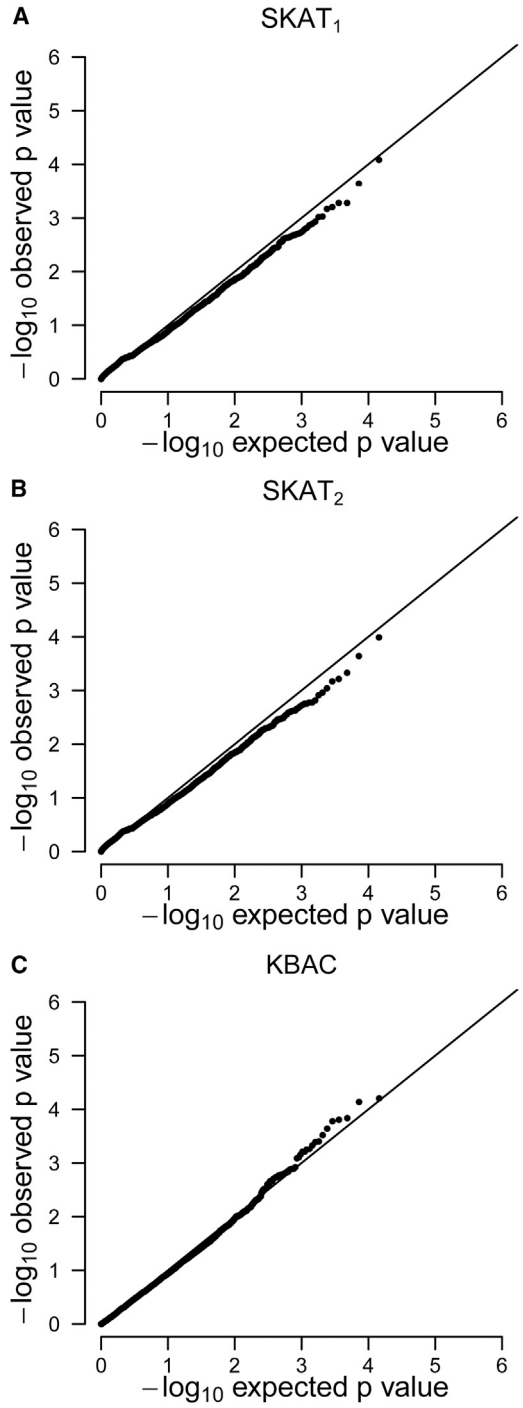

Figure 1. Q-Q Plots Showing the p Values from the SKAT and KBAC Association Tests
(A) $SKAT_1$ test (linear kernel, all variants have equal weight).
(B) $SKAT_2$ test (weighted IBS kernel, extra weight to rare variants).
(C) KBAC test.
Solid lines denote the diagonal.

As discussed in the Subjects and Methods, we extended the target regions by 100 bp. The extended target regions showed a lower median depth of coverage than did the actual target regions (Figure S8). Also, the extended target regions showed a lower Ti/Tv ratio than did the target regions. This was because variants outside of the coding region were included in the extended region (Table S7). However, the ratio was still within the expected threshold.[57] The majority of variants in the extended region were noncoding, but we also found approximately 2,500 additional variants annotated to the exonic regions (Table S8).

To test for potential bias in the BEAGLE data set, we calculated the Ti/Tv ratio for each sample (Figure S9). We observed no outlying samples. We also plotted the Ti/Tv ratio as a function of r^2 (Figure S10) and found that as r^2 increased, the Ti/Tv ratio stabilized around 2.5. Finally, as seen in the unimputed data, putatively functional variants were present at a lower frequency than were other variants (Figure S11).

Figure S12 shows the distribution of the number of sites across the exome where each individual carries at least one nonreference allele for different types of coding SNPs. Figure S13 shows the number of nonreference alleles carried by each exome for different types of coding SNPs. These counts are broadly in line with what has been previously reported from exome sequencing data.[53,58,59]

We also attempted to validate some of the singletons and doubletons detected in our data set. All 31 singletons were validated by traditional Sanger sequencing. For one of the 15 doubletons, we could only validate one carrier of the variant; the other 14 doubletons were validated for both carriers.

After genotype imputation and quality-control analyses, we were left with 1.6 million autosomal variants (of which 286,083 were in the exons). See Table S9 for the number of SNPs per gene.

Results of the Association Tests

We next applied the gene-based association tests to our exome sequencing data. The quantile-quantile (Q-Q) plots of the p values from the SKAT and KBAC association tests showed good agreement with the expected distribution, suggesting few biases (e.g., population stratification, differences in technical artifacts, etc.) between cases and controls (Figure 1). However, none of the genes under either test passed the Bonferroni-correction threshold (Tables S10–S12). We obtained similar results when we only included SNPs with a MAF $< 1\%$, when we included all SNPs regardless of frequency, and for the additional six gene-based association tests (Figures S14–S18 and Tables S10–S18). Additionally, we found no genes passing the Bonferroni-correction threshold when we reduced the number of genes analyzed to include only those with at least five or at least ten SNPs (Figures S19 and S20).

We also applied a single-marker association test to each of the putative functional variants. However, no SNP passed the Bonferroni-correction threshold of $p < 10^{-7}$ (Figure S21 and Tables S19–S21). Interestingly, there was a substantial overlap in the top genes identified in the single-marker analysis and the SKAT association tests. For example, 9 of the top 20 genes with the lowest p values

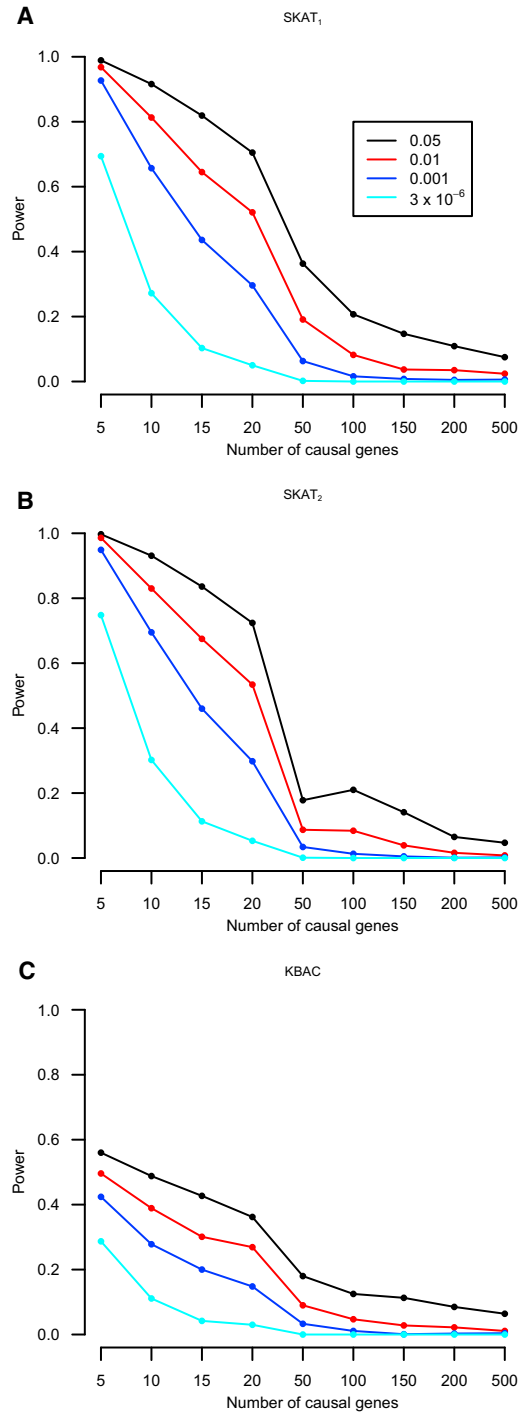

Figure 2. Power to Detect an Association with SKAT or KBAC for Different Numbers of Causal Genes

Different colored curves denote different significance levels. The Bonferroni threshold was 3×10^{-6}. Note that power was low when there were many causal genes (genes containing causal variants) such that the heritability explained by a given gene was very low.
(A) $SKAT_1$ test (linear kernel, all variants have equal weight).
(B) $SKAT_2$ test (weighted IBS kernel, extra weight to rare variants).
(C) KBAC test.

in the $SKAT_1$ test were also among the top 20 genes with the lowest single-marker p values. This finding suggests that in our data, the effects detected with the SKAT test could also be captured by the single-marker analysis. However, this pattern is not universal to all association tests, or data sets, given that only three genes with the lowest p values in the KBAC test in our data were among the top 20 genes with the lowest single-marker p values. Further investigation of the relationship between single-marker and gene-based association tests is warranted.

Gene-Set Analyses

Sets of genes related to diabetes or metabolic traits might be enriched with lower p values from the gene-based association tests, even though none of the individual p values passed the multiple test correction. We examined several sets of genes related to metabolism, as well as the corresponding interactomes of a subset of these genes (Tables S2 and S3). We found that genes previously associated with obesity through GWASs were enriched with lower $SKAT_1$ p values ($p < 0.006$, Table S22). Additionally, a few of the interactomes, including those for two genes in which mutations are known to cause monogenic forms of diabetes[60] (*HNF1A* [MIM 142410] and *HNF4A* [MIM 600281]), were marginally enriched with putatively functional variants relative to synonymous variants. However, these results were not significant after correction for multiple tests (Table S23). A comprehensive overview of the gene-set analysis is found in Tables S23 and S24.

Finally, monogenic forms of diabetes might be hidden among individuals diagnosed with type 2 diabetes. We therefore examined whether any of the individuals in our study carried mutations previously demonstrated to cause diabetes in the most common monogenic-diabetes-associated genes (*HNF1A* [MIM 142410], *GCK* [MIM 138079], and *HNF4A* [MIM 600281]). We identified three carriers of *HNF1A* mutations (two cases and one control) and two carriers of *GCK* mutations (two cases).[60,61]

Statistical Power of the Gene-Based Association Tests

To evaluate the statistical power of these tests in the context of our study under a variety of genetic models, we performed a series of power simulations. These simulations conditioned on the number of SNPs per gene and the genotypes of the individuals sequenced in our study. As such, they should closely reflect the power of our study. These models assumed that the total heritability of type 2 diabetes (approximately 30%[33–35,55]) is equally divided among a number (n) of different genes, each accounting for $1/n^{th}$ of the heritability. As expected, the power to detect an association depended on the amount of heritability explained and the number of causal variants (Figure 2). For example, if all of the heritability of the trait is explained by functional variation at five genes, we would detect a given causal gene (a gene containing causal mutations) by using the $SKAT_1$ test at our Bonferroni threshold 70% of the time. As the amount of the heritability explained by a given gene in our simulations decreased, so did the power to detect that association. When there were more than 15 genes contributing risk, the power to

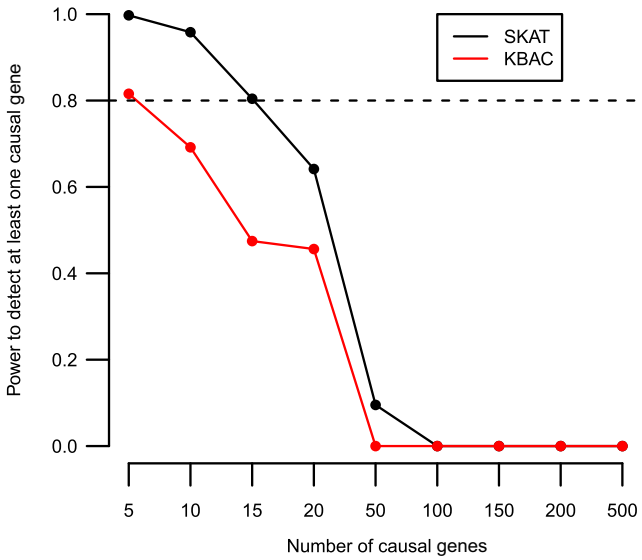

Figure 3. Power to Detect at Least One of *n* Causal Genes at a Bonferroni Significance Threshold for Different Numbers of Causal Genes

The power to detect at least one causal gene is calculated as $1 - (1 - \text{power})^n$, where n is the number of causal genes and power is estimated from the simulations shown in Figure 2.

detect an association at a given gene was extremely low (<10% with the Bonferroni threshold), suggesting that we did not have power to detect genes of weak effect with the present sample size. Similar results were obtained for the $SKAT_2$ and KBAC tests, although KBAC showed lower power than the SKAT tests for the parameters simulated here (Figure 2). We also varied the proportion of SNPs assumed to be causal (c) within a gene. As c decreased, power also slightly decreased (Figures S22–S24). Our power simulations conditioned on including genes for which the individual normalized SNP effects (κ_i) were <3. The simulations not including this conditioning step suggest that our study had lower power than described (Figure S25). This apparent decrease in power was a result of including genes with very little genetic variation and very large κ_i values. As discussed in the Subjects and Methods section, including genes with SNPs whose $\kappa_i > 3$ is not biologically reasonable. Thus, further results only include genes in which all SNPs have $\kappa_i < 3$.

The simulations described above estimated the power to detect a single associated gene. However, applying them to the entire exome provides the possibility of detecting an association at any of the n different risk loci. We next tabulated the probability of detecting at least one significantly associated gene (at the Bonferroni threshold) for each of the genetic models analyzed (see Subjects and Methods). Models with many risk genes have smaller effect sizes per gene, making it harder to detect each gene. However, as the number of risk genes increases, there are more opportunities to detect a causal gene, increasing the probability of detecting at least one true association (Figure 3). The power to detect at least one significant gene was highest (>80%) for the SKAT tests when there were a limited number of risk genes (fewer than 15). At 20 risk genes, we had >60% power to detect at least one risk gene by using the SKAT test (Figure 3). As the number of risk loci increased beyond 20, the amount of the heritability explained by any one gene was so low that we had limited power to find even one such gene. The overall trend held regardless of the proportion of nonsynonymous SNPs assumed to be causal (c) within each gene. Although the power to detect at least one significant gene decreased as c decreased, we still had >50% power to detect at least one gene even when $c = 0.25$ and there were $\leq$20 causal genes (Figure S26).

In sum, the statistical power analyses suggest that we had limited power to detect a particular association unless the gene in question explained a substantial proportion of the heritability of disease risk; however, when we applied the simulations to the whole exome, we had substantial power to detect at least one significant association at our Bonferroni threshold if rare variation in a modest number of genes (fewer than 20) was responsible for the majority of disease risk. Because we did not detect such an association, our results suggest that low-frequency variants in a modest number of genes do not explain a substantial amount of the heritability of type 2 diabetes.

Discussion

It has been hypothesized that rare genetic variants with moderate effects on disease risk could account for much of the missing heritability of complex traits.[6,9,10,62] We have taken a first step toward testing this hypothesis for type 2 diabetes. We did not detect any significant associations between rare coding variants and common forms of diabetes. Our study was underpowered to detect weak genetic effects, but if much of the heritability of type 2 diabetes is explained by variants in a modest number of genes, we should have detected at least one associated locus at our Bonferroni significance threshold. Thus, our empirical results, combined with the statistical power simulations, suggest that when clustered in fewer than 20 genes, coding variants of moderate effect do not account for much of the missing heritability of a common polygenic disorder such as type 2 diabetes.

Importantly, although GWASs have identified more than 60 common SNPs associated with type 2 diabetes risk,[36] these data alone are insufficient to reject a model where fewer than 20 genes containing variants of strong effect can account for much of the heritability of the disorder. The reason that GWAS data alone cannot be used for rejecting this model is that all the loci implicated through GWASs have very weak genetic effects and only explain a small fraction of the total heritability of the trait.[36] Thus, there is still a tremendous amount of heritability to be explained. GWASs of common variants do not address the question of whether that heritability

could be accounted for by low-frequency and rare variants of moderate effect in a small number of genes. Our whole-exome sequencing study has explicitly addressed this question. Additionally, we did not examine whether there are fewer than 20 genes involved in type 2 diabetes but rather looked at whether rare coding variants in fewer than 20 genes account for much of the heritability. In such a model, any number of other genes that do not account for much of the heritability can be involved. The previously identified GWAS loci would fall into this category.

In our statistical power simulations, we assumed that each of the n risk genes account for the same proportion of the heritability. Both theory and data from common variants suggest that this is unlikely to apply in practice.[5,63–65] However, the assumption that n distinct genes contribute to the heritability equally is actually conservative and means that we can reject additional models that we did not explicitly simulate. For example, imagine a type 2 diabetes architecture in which 10 of 100 risk loci account for almost all of the heritability. Although we did not directly simulate this scenario, it would be very similar to our simulation in which ten genes explained all of the heritability of type 2 diabetes. The reason for this is that the remaining 90 genes of weak effect would most likely have failed to be detected in our study. If they collectively do not account for much of the heritability, then they can be discounted in our simulations, and this scenario becomes equivalent to one that we simulated (i.e., the scenario with ten causal genes).

Our power simulations included several important assumptions. First, we assumed that causal variants act in an additive manner. Although such a model is predicted from theory and data,[66] it is not clear how well the additive model will hold for rare coding mutations. If many rare coding mutations are weakly deleterious and are affected by purifying natural selection, they might be slightly recessive.[67,68] Second, we assumed that simulated causal variants can either increase or decrease diabetes risk. Third, we assumed a heritability of 30% for the trait. If the true heritability of common forms of type 2 diabetes is lower, then our power to detect an association in our study similarly would have been lower. However, published estimates of the heritability of type 2 diabetes and related metabolic traits suggest that 30% is toward the low end of the heritability estimates for these phenotypes (Table S4). For this reason, we did not further decrease the heritability used in our power simulations to account for the small (5.7%) variance in disease risk that can be accounted for by the 63 previously identified associations with common variants.

Our empirical and simulation results are compatible with a variety of different genetic architectures for type 2 diabetes. First, if rare coding variants are responsible for the majority of the heritability of the trait, the variants are most likely scattered across many (>20) different genes. Thus, genetic variants in no one gene can account for much of the heritability of the trait. Biologically, such a model would postulate that there are a large number of genes that can be mutated to cause type 2 diabetes in a given individual. Each individual would then carry a subset of genetic variants located in several of the many causal genes. Our finding that genes previously implicated in obesity risk through GWASs showed unusually low SKAT p values in our study supports a scenario in which low-frequency and rare variants in multiple genes could be responsible for risk of common metabolic diseases. It also suggests that genes carrying common variants associated with a trait could also carry additional low-frequency and rare coding variants that increase disease risk.

Yet another model for the genetic architecture of diabetes that could explain our results is that low-frequency and rare coding variants do not account for much of the heritability of type 2 diabetes. Under this scenario, the missing heritability could be located in common or low-frequency and rare variants in noncoding regions of the genome. Recent studies that jointly modeled diabetes or obesity risk as a function of genetic relatedness across all of the GWAS SNPs have suggested that much of the heritability of these traits can be explained by common variants with effects that are too small to reach genome-wide significance in currently used GWASs.[69–71] Under this model, low-frequency and rare coding variants do not account for a substantial amount of the heritability of complex traits. Our results are consistent with such a model. Alternatively, the heritability of type 2 diabetes could have been overestimated in family studies as a result of environmental factors or gene-gene interactions.[72]

Recently, is has been suggested that when clustered within the same gene, rare variants of strong effect could explain some of the associations between common variants found in GWASs and complex traits.[73,74] Such signals have been termed "synthetic associations." We found little evidence for such synthetic associations within our data. In particular, we did not detect an excess of functional variants within genes containing common variant(s) implicated in diabetes, obesity, or hypertension through GWASs. Such an excess could be expected under a model where the original GWAS signal could be explained by rare functional variants. Additionally, we found that genes containing GWAS hits for obesity had significantly lower $SKAT_1$ p values than did random genes. Although this pattern could be driven by synthetic associations, we found that the $SKAT_1$ p values were weakly correlated (Spearman's $\rho = 0.34$, $p = 0.067$) with the best single-marker p value within each gene. The fact that the single-marker p values were even slightly informative regarding the multimarker $SKAT_1$ p values suggests that the skew in $SKAT_1$ p values was not entirely driven by very rare variants (e.g., singletons in cases), as predicted by the synthetic-association hypothesis. However, further investigations using larger numbers of cases and controls

will be required for convincingly supporting or rejecting the synthetic association hypothesis.

The fact that we did not detect any significant association in our data has implications for designing and analyzing further sequencing studies for elucidating the genetic basis of complex traits. If there is substantial locus heterogeneity, then gene-based tests of association are likely to be severely underpowered. The reason for this is that a particular causal gene is likely to carry causal variants in only a small number of the affected cases. Other cases carry risk variants in different genes. The gene-based association tests currently used are substantially underpowered in this model, even with sample sizes of thousands of cases and controls. If the many distinct genes that, if mutated, could give rise to the trait of interest all cluster within a small number of pathways, power could be gained through the implementation of the gene-based tests at a pathway level. Another possibility would be to use family-based association studies of rare variants on extended pedigrees.[75,76] This approach could be particularly effective if the same causal genes are responsible for the phenotype for most members of the pedigree. Further methodological work is required in this area. Conversely, if rare variation in coding regions contributes little to complex disease risk, then this would argue for alternative study designs. For example, whole-genome sequencing, rather than exome sequencing, would allow for the detection of rare variants outside of the coding regions. This would be an effective strategy if the missing heritability could be accounted for by rare noncoding regulatory variants.

Although our results argue that low-frequency and rare coding variants in a modest number of genes do not account for the majority of the heritability of common forms of type 2 diabetes, it is not clear how generalizable this result is to other complex traits. Several other exome sequencing studies have failed to detect any significant associations between low-frequency variants and schizophrenia,[77] epilepsy,[78] autism,[79] or autoimmune diseases.[80] However, recent studies have associated rare variants with age-related macular degeneration.[81–83] Thus, the genetic architecture and the role of low-frequency and rare variants are likely to be trait dependent and will need to be addressed empirically.

Appendix A: Filtering Sites

After generating initial genotype calls from SAMtools, we applied a series of site filters to the 2,958,319 sites with a MAF > 0.0001 to obtain a set of sites with high-quality genotype calls suitable for association analysis. We describe those filters here:

Depth Filter

The average depth per site across all 1,998 samples was calculated from pile-up files generated by SAMtools. Sites with an average depth less than 4 or greater than 150 were removed.

Base-Quality-Score Filter

For each site, we tested whether base quality scores of the minor allele were significantly smaller than those from the major allele. Base quality scores were collected for the major and minor alleles. Because the combined quality scores were always very large, even sites that only had a small difference between the median quality scores for both alleles gave very significant Wilcoxon rank-sum p values. To set up a proper filter, we defined a quantity $Q_{\text{diff}} = |m_1 - m_2|/0.5(m_1 + m_2)$, which measured the absolute difference in read quality scores between the major allele and the minor allele. Here, m_1 is the median read quality score for the major allele, and m_2 is the median read quality score for minor allele. Sites with a Wilcoxon rank-sum p value $< 10^{-7}$ and $Q_{\text{diff}} > 0.1$ were removed.

Strand-Bias Filter

Because the capture experiment shows biases with respect to strand, it is inappropriate to test the hypothesis that half of the reads were derived from the forward strand and the other half were derived from the reverse strand. From the pile-up files generated by SAMtools, we tested the homogeneity of the distribution of reads along the two strands for the major and minor alleles. We first made a 2×2 table whose rows contained the major (M) and minor (m) alleles and whose columns contained the number of reads that came from the forward strand (+) and the number of reads from the reverse strand (−). The odds ratio (OR), defined as (M+ / M−) / (m+ / m−), was used for measuring the difference of the strand distribution for reads from the major allele and reads from the minor allele. The p values were calculated with a Fisher's exact test. Sites with p values $< 10^{-7}$ and $\log_2$ (OR) less than −3 or greater than 3 were removed.

Mappability Filter

We computed a mappability score to assess the accessibility of each base in the human genome under typical Illumina sequencing conditions. The mappability score represents the probability that a read comes from the hg19 genomic position to which it mapped. As a probability, the mappability score ranges from 0 to 1. Sites with mappability scores < 0.5 were removed.

Homopolymer Filter

The homopolymer run is the length of the homopolymer surrounding a SNP site. For example, on hg19, the 10-base sequence on either side of site chr1: 14,673 is 5′-CTGGGTCTGG[G]GGGGAAGGTG-3′. The maximum homopolymer run of the SNP site (denoted by [G]) is 7. Sites with homopolymer runs greater than 6 were removed.

Allele-Balance Filter

At well-behaved heterozygous sites, within a given individual, the number of reads for the major allele should equal the number for the minor allele. We tested this by using a binomial test with $p = 0.5$. Specifically, from the SAMtools pile-up files, we calculated the total number of reads for the major allele and the minor allele at all heterozygous genotype calls. We calculated the absolute difference in the number of reads for the major and minor alleles as $B_{diff} = |r_1 - r_2|/|r_1 + r_2|$, where r_1 is the number of reads for the major allele and r_2 is the number of reads for the minor allele. Sites with p values $< 10^{-6}$ and $B_{diff} > 0.5$ were removed.

Hardy-Weinberg Filter

The Hardy-Weinberg filter was applied after genotype calling (estimated from the 1,000 control samples). The exact test[11] was applied with the software PLINK.[12]

SNPs with p values $< 10^{-6}$ were removed. After application of the above quality threshold, 729,538 (713,122 autosomal) variants (of which 282,823 were in the exonic region) were retained for further analyses.

Appendix B: Filtering Individuals

Low Sequencing Depth

In total, 1,998 samples were successfully sequenced. Of these, 999 were male and the remaining 999 were female. After removal of duplicated reads, the average depth of each sample was 56× (SD = 8.71). With the exception of one sample with an average depth of 25×, all samples were sequenced to a depth ≥ 30×, and 1,522 (76.2%) samples were sequenced to a depth ≥ 50×. The sample with a depth of 25× was removed.

Contamination

In the process of library construction and DNA sequencing, a sample could become contaminated. For a heavily contaminated sample, more variants will be identified and more heterozygous genotypes will be called. But for a sample that is only slightly contaminated, we might not observe such a deviation. For two samples (A and B), let A_i and B_i be the genotypes at a variable site i. A_i and B_i can take values *MM*, *Mm*, or *mm*, corresponding to a homozygote for the reference allele, a heterozygote, and a homozygote for the nonreference allele, respectively. If sample A is contaminated by B and A_i is *MM* and B_i is *mm* or *Mm*, for a sequencing read, we have a higher probability to observe an *m* than in the situation when A is not contaminated. Given a sample that has been genotyped at thousands or more variable sites across the autosomes, we can calculate the fraction of reads that disagree with known genotypes at homozygous sites. Samples that show a large deviation are most likely contaminated. In practice, we applied the above method to 1,963 samples with Iselect data.[13] We did calculations at 7,219 variable sites with a MAF > 0.05 in unique regions of 22 autosomes. Reads that disagreed with known homozygous genotypes were counted. SNP sites with sequencing depth < 8× were skipped. We found six samples showing obvious deviations, indicating that they were contaminated. These samples were excluded from further analyses.

Sex Check, Inbreeding, Relatedness, and Concordance with Previous Genotype Projects

We used PLINK to check for correct sex, inbreeding, and relatedness. Five samples were removed because of F-values (inbreeding coefficients) lower than −0.12 (indicating contamination, admixture, or genotyping errors) or higher than 0.12 (indicating inbreeding). Eight samples had disagreements between the genotypic and phenotypic sex or had undetermined genotypic sex.

We assessed the concordance with previous genotyping projects by comparing genotype data from previous projects to our exome sequencing data. The number of overlapping SNPs from previous genotyping and exome sequencing was 65. Eighteen samples had missing genotypes for all sites and could thus not be compared. Three samples had a genotype concordance < 0.9. These samples were excluded.

We found two pairs of duplicate samples. One was part of a parent-offspring pair. We also found seven pairs of full siblings and three pairs with second-degree relationships (half siblings). To extract the half siblings, we first removed IDs with too high or low inbreeding coefficient (five individuals, described above). When two samples were found to be related, we removed one of them from our analyses.

After application of the above filters, 1,965 samples (983 controls and 982 cases) were retained for further analyses.

Supplemental Data

Supplemental Data include 26 figures and 24 tables and can be found with this article online at http://www.cell.com/AJHG.

Acknowledgments

We thank Melissa Wilson Sayres and Vincent Plagnol for helpful discussions and comments on the manuscript. We also thank A. Forman, T. Lorentzen, B. Andreasen, and G.J. Klavsen for technical assistance and A.L. Nielsen, G. Lademann, and M.M.H. Kristensen for management assistance. K.E.L. was supported by a Miller Research Fellowship from the Miller Research Institute at the University of California, Berkeley. T.S. was supported by the Danish Council for Independent Research. T.H.P. was supported by The Danish Council for Independent Research Medical Sciences. M.D. was supported by National Science Foundation grant DBI-1103639. This project was funded by the Lundbeck Foundation and produced by The Lundbeck Foundation Centre for Applied Medical Genomics in Personalised Disease Prediction, Prevention, and Care (www.lucamp.org). The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center at the University of Copenhagen and is partially funded by

an unrestricted donation from the Novo Nordisk Foundation (http://metabol.ku.dk/). Further funding came from the Danish Council for Independent Research Medical Sciences. The Inter99 was initiated by Torben Jørgensen (principal invesitigator [PI]), Knut Borch-Johnsen (co-PI), Hans Ibsen, and Troels F. Thomsen. The steering committee comprises the former two and Charlotta Pisinger. The study was financially supported by research grants from the Danish Research Council, the Danish Centre for Health Technology Assessment, Novo Nordisk, the Research Foundation of Copenhagen County, the Ministry of Internal Affairs and Health, the Danish Heart Foundation, the Danish Pharmaceutical Association, the Augustinus Foundation, the Ib Henriksen Foundation, the Becket Foundation, and the Danish Diabetes Association.

Received: June 7, 2013
Revised: October 16, 2013
Accepted: November 4, 2013
Published: November 27, 2013

Web Resources

The URLs for data presented herein are as follows:

2,000 exomes sequenced in this study, http://metabol.ku.dk/scientific_sections/metabolic_genetics/exome-sequencing-of-2000-danish-individuals/
Analysis of Next Generation Sequencing Data (ANGSD), http://www.popgen.dk/software/index.php/Main_Page#ANGSD
AssotesteR, https://github.com/gastonstat/AssotesteR
ClinicalTrials.gov, http://clinicaltrials.gov/
dbSNP, http://www.ncbi.nlm.nih.gov/snp
KBAC, https://code.google.com/p/kbac-statistic-implementation/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
Primer3, http://bioinfo.ut.ee/primer3-0.4.0/primer3/
SeattleSeq Annotation server, http://snp.gs.washington.edu/SeattleSeqAnnotation137/
Sequence Kernel Association Test (SKAT), http://www.bios.unc.edu/~mcwu/software/
UCSC Genome Browser, http://genome.ucsc.edu

References

1. King, R.A., Rotter, J.I., and Motulsky, A.G. (2002). The Genetic Basis of Common Diseases (New York, NY: Oxford University Press).
2. Risch, N.J. (2000). Searching for genetic determinants in the new millennium. Nature *405*, 847–856.
3. Stranger, B.E., Stahl, E.A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. Genetics *187*, 367–383.
4. Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic mapping in human disease. Science *322*, 881–888.
5. Park, J.H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat. Genet. *42*, 570–575.
6. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
7. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.
8. Gibson, G. (2011). Rare and common variants: twenty arguments. Nat. Rev. Genet. *13*, 135–145.
9. Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. *11*, 415–425.
10. Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev. *19*, 212–219.
11. Eyre-Walker, A. (2010). Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc. Natl. Acad. Sci. USA *107* (*Suppl 1*), 1752–1756.
12. Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. *69*, 124–137.
13. Ahituv, N., Kavaslar, N., Schackwitz, W., Ustaszewska, A., Martin, J., Hebert, S., Doelle, H., Ersoy, B., Kryukov, G., Schmidt, S., et al. (2007). Medical sequencing at the extremes of human body mass. Am. J. Hum. Genet. *80*, 779–791.
14. Cohen, J., Pertsemlidis, A., Kotowski, I.K., Graham, R., Garcia, C.K., and Hobbs, H.H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in *PCSK9*. Nat. Genet. *37*, 161–165.
15. Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science *305*, 869–872.
16. Romeo, S., Pennacchio, L.A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H.H., and Cohen, J.C. (2007). Population-based resequencing of *ANGPTL4* uncovers variations that reduce triglycerides and increase HDL. Nat. Genet. *39*, 513–516.
17. Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J.A. (2009). Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. Science *324*, 387–389.
18. Albrechtsen, A., Grarup, N., Li, Y., Sparsø, T., Tian, G., Cao, H., Jiang, T., Kim, S.Y., Korneliussen, T., Li, Q., et al.; D.E.S.I.R. Study Group; DIAGRAM Consortium (2013). Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. Diabetologia *56*, 298–310.
19. Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. *26*, 1135–1145.
20. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*, 272–276.
21. Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010). Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. Nat. Genet. *42*, 790–793.
22. Ng, S.B., Nickerson, D.A., Bamshad, M.J., and Shendure, J. (2010). Massively parallel sequencing and rare disease. Hum. Mol. Genet. *19* (R2), R119–R124.
23. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson,

D.A., et al. (2010). Exome sequencing identifies the cause of a Mendelian disorder. Nat. Genet. *42*, 30–35.
24. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755.
25. Bamshad, M.J., Shendure, J.A., Valle, D., Hamosh, A., Lupski, J.R., Gibbs, R.A., Boerwinkle, E., Lifton, R.P., Gerstein, M., Gunel, M., et al.; Centers for Mendelian Genomics (2012). The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. Am. J. Med. Genet. A. *158A*, 1523–1525.
26. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.
27. Liu, D.J., and Leal, S.M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet. *6*, e1001156.
28. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.
29. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
30. Asimit, J., and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. Annu. Rev. Genet. *44*, 293–308.
31. Pan, W., Basu, S., and Shen, X. (2011). Adaptive tests for detecting gene-gene and gene-environment interactions. Hum. Hered. *72*, 98–109.
32. Permutt, M.A., Wasson, J., and Cox, N. (2005). Genetic epidemiology of diabetes. J. Clin. Invest. *115*, 1431–1439.
33. Almgren, P., Lehtovirta, M., Isomaa, B., Sarelin, L., Taskinen, M.R., Lyssenko, V., Tuomi, T., and Groop, L.; Botnia Study Group (2011). Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. Diabetologia *54*, 2811–2819.
34. Kaprio, J., Tuomilehto, J., Koskenvuo, M., Romanov, K., Reunanen, A., Eriksson, J., Stengård, J., and Kesäniemi, Y.A. (1992). Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. Diabetologia *35*, 1060–1067.
35. Poulsen, P., Kyvik, K.O., Vaag, A., and Beck-Nielsen, H. (1999). Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance—a population-based twin study. Diabetologia *42*, 139–145.
36. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al.; Wellcome Trust Case Control Consortium; Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. *44*, 981–990.
37. Alberti, K.G., and Zimmet, P.Z. (1998). Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabet. Med. *15*, 539–553.
38. Jørgensen, T., Borch-Johnsen, K., Thomsen, T.F., Ibsen, H., Glümer, C., and Pisinger, C. (2003). A randomized non-pharmacological intervention study for prevention of ischaemic heart disease: baseline results Inter99. Eur. J. Cardiovasc. Prev. Rehabil. *10*, 377–386.
39. Glümer, C., Jørgensen, T., and Borch-Johnsen, K.; Inter99 study (2003). Prevalences of diabetes and impaired glucose regulation in a Danish population: the Inter99 study. Diabetes Care *26*, 2335–2340.
40. Lauritzen, T., Griffin, S., Borch-Johnsen, K., Wareham, N.J., Wolffenbuttel, B.H., and Rutten, G.; Anglo-Danish-Dutch Study of Intensive Treatment in People with Screen Detected Diabetes in Primary Care (2000). The ADDITION study: proposed trial of the cost-effectiveness of an intensive multifactorial intervention on morbidity and mortality among people with Type 2 diabetes detected by screening. Int. J. Obes. Relat. Metab. Disord. *24* (*Suppl 3*), S6–S11.
41. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.
42. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
43. Kim, S.Y., Lohmueller, K.E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., Grarup, N., Jiang, T., Andersen, G., Witte, D., et al. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinformatics *12*, 231.
44. Skotte, L., Korneliussen, T.S., and Albrechtsen, A. (2012). Association testing for next-generation sequencing data using score statistics. Genet. Epidemiol. *36*, 430–437.
45. Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS ONE *7*, e37558.
46. Browning, B.L., and Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am. J. Hum. Genet. *85*, 847–861.
47. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
48. Lage, K., Karlberg, E.O., Størling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol. *25*, 309–316.
49. Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA *101*, 6062–6067.

50. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
51. Do, R., Kathiresan, S., and Abecasis, G.R. (2012). Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum. Mol. Genet. *21* (R1), R1–R9.
52. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838.
53. Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al. (2012). Exome sequencing and the genetic basis of complex traits. Nat. Genet. *44*, 623–630.
54. Long, A.D., and Langley, C.H. (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. *9*, 720–731.
55. Pilia, G., Chen, W.M., Scuteri, A., Orrú, M., Albai, G., Dei, M., Lai, S., Usala, G., Lai, M., Loi, P., et al. (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet. *2*, e132.
56. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
57. Bainbridge, M.N., Wang, M., Wu, Y., Newsham, I., Muzny, D.M., Jefferies, J.L., Albert, T.J., Burgess, D.L., and Gibbs, R.A. (2011). Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol. *12*, R68.
58. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al.; Broad GO; Seattle GO; NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.
59. Challis, D., Yu, J., Evani, U.S., Jackson, A.R., Paithankar, S., Coarfa, C., Milosavljevic, A., Gibbs, R.A., and Yu, F. (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics *13*, 8.
60. Ellard, S., and Colclough, K. (2006). Mutations in the genes encoding the transcription factors hepatocyte nuclear factor 1 alpha (*HNF1A*) and 4 alpha (*HNF4A*) in maturity-onset diabetes of the young. Hum. Mutat. *27*, 854–869.
61. Osbak, K.K., Colclough, K., Saint-Martin, C., Beer, N.L., Bellanné-Chantelot, C., Ellard, S., and Gloyn, A.L. (2009). Update on mutations in glucokinase (*GCK*), which cause maturity-onset diabetes of the young, permanent neonatal diabetes, and hyperinsulinemic hypoglycemia. Hum. Mutat. *30*, 1512–1526.
62. Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R., and Amos, C.I. (2008). Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am. J. Hum. Genet. *82*, 100–112.
63. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.
64. Rockman, M.V. (2012). The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution *66*, 1–17.
65. Orr, H.A. (1998). The population genetics of adaptation: The distribution of factors fixed during adaptive evolution. Evolution *52*, 935–949.
66. Hill, W.G., Goddard, M.E., and Visscher, P.M. (2008). Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. *4*, e1000008.
67. Thornton, K.R., Foran, A.J., and Long, A.D. (2013). Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. PLoS Genet. *9*, e1003258.
68. Li, Y., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Jiang, T., Jiang, H., Albrechtsen, A., Andersen, G., Cao, H., Korneliussen, T., et al. (2010). Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat. Genet. *42*, 969–972.
69. Stahl, E.A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B.F., Kraft, P., Chen, R., Kallberg, H.J., Kurreeman, F.A., et al.; Diabetes Genetics Replication and Meta-analysis Consortium; Myocardial Infarction Genetics Consortium (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat. Genet. *44*, 483–489.
70. Vattikuti, S., Guo, J., and Chow, C.C. (2012). Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. PLoS Genet. *8*, e1002637.
71. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. *43*, 519–525.
72. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA *109*, 1193–1198.
73. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D.B. (2010). Rare variants create synthetic genome-wide associations. PLoS Biol. *8*, e1000294.
74. Chang, D., and Keinan, A. (2012). Predicting signatures of "synthetic associations" and "natural associations" from empirical patterns of human genetic variation. PLoS Comput. Biol. *8*, e1002600.
75. McClellan, J., and King, M.C. (2010). Genetic heterogeneity in human disease. Cell *141*, 210–217.
76. Lupski, J.R., Belmont, J.W., Boerwinkle, E., and Gibbs, R.A. (2011). Clan genomics and the complex architecture of human disease. Cell *147*, 32–43.
77. Need, A.C., McEvoy, J.P., Gennarelli, M., Heinzen, E.L., Ge, D., Maia, J.M., Shianna, K.V., He, M., Cirulli, E.T., Gumbs, C.E., et al. (2012). Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. Am. J. Hum. Genet. *91*, 303–312.
78. Heinzen, E.L., Depondt, C., Cavalleri, G.L., Ruzzo, E.K., Walley, N.M., Need, A.C., Ge, D., He, M., Cirulli, E.T., Zhao, Q., et al. (2012). Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy. Am. J. Hum. Genet. *91*, 293–302.
79. Liu, L., Sabo, A., Neale, B.M., Nagaswamy, U., Stevens, C., Lim, E., Bodea, C.A., Muzny, D., Reid, J.G., Banks, E., et al. (2013). Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. PLoS Genet. *9*, e1003443.

80. Hunt, K.A., Mistry, V., Bockett, N.A., Ahmad, T., Ban, M., Barker, J.N., Barrett, J.C., Blackburn, H., Brand, O., Burren, O., et al. (2013). Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature *498*, 232–235.
81. Helgason, H., Sulem, P., Duvvari, M.R., Luo, H., Thorleifsson, G., Stefansson, H., Jonsdottir, I., Masson, G., Gudbjartsson, D.F., Walters, G.B., et al. (2013). A rare nonsynonymous sequence variant in *C3* is associated with high risk of age-related macular degeneration. Nat. Genet. *45*, 1371–1374.
82. Seddon, J.M., Yu, Y., Miller, E.C., Reynolds, R., Tan, P.L., Gowrisankar, S., Goldstein, J.I., Triebwasser, M., Anderson, H.E., Zerbib, J., et al. (2013). Rare variants in *CFI, C3* and *C9* are associated with high risk of advanced age-related macular degeneration. Nat. Genet. *45*, 1366–1370.
83. Zhan, X., Larson, D.E., Wang, C., Koboldt, D.C., Sergeev, Y.V., Fulton, R.S., Fulton, L.L., Fronick, C.C., Branham, K.E., Bragg-Gresham, J., et al. (2013). Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. Nat. Genet. *45*, 1375–1379.