Supplementary Material

1. Study design and data generation

We selected 1,034 individuals from 20 pedigrees that are part of San Antonio Family Heart Study (SAFHS) (Mitchell et al. 1996; McClellan and King 2010) and San Antonio Family Diabetes/Gallbladder Study (SAFDGS) projects (Hunt et al. 2005; Puppala et al. 2006). Prevalence of type 2 diabetes (T2D) in the families was 30%, significantly higher than the estimated prevalence of 15.7% in general Mexican American population of the region. Estimated heritability within 1,034 pedigree samples was 0.47 for T2D, 0.41 for fasting glucose, and 0.63 for fasting insulin. Each individual has T2D and glycemic trait measured longitudinally up to five times.

1.1. Sequencing and initial variant calling

We used ExomePick software (Sidore et al. 2015) to select 600 individuals to be whole genome sequencing to gain maximal genetic information for sample size 600. Almost all individuals in our dataset had been previously genotyped using GWAS genotyping arrays providing the scaffold of our imputation protocol. Whole-genome sequencing and initial variant calling was done by Complete Genomics (CGI). CGI sequencing generates reads based on self-assembling DNA nanoballs (DNBs). Each DNB generate 70 base reads, which corresponds to paired-end reads of 35 bases at each end in other sequencing technologies (e.g. Illumina). DNB reads are mapped against reference genome and reads in the regions that are likely to differ from reference sequence are assembled by local *de novo* assembly. NCBI build 37 human reference genome was used with CGI pipeline version 1.0. Details on CGI sequencing technology is described in (Drmanac et al. 2010) and details about CGI computational pipeline including mapping, local assembly, and variant calling is described in (Carnevali et al. 2012).

1.2. Sequenced sample quality control

We first examined sample identity, pedigree info, and sequencing quality outliers on sequenced samples using initial variant calls, the reported pedigree structure as defined by the San Antonio family studies, and summary reports generated by the CGI pipeline about sequencing metrics and variant calling statistics.

Pedigree verification: We validated the kinship structure of reported pedigrees and to avoid potential sample mix-up using KING (Manichaikul et al. 2010). We calculated pairwise kinship coefficients between all subject pairs and compared these to kinship coefficients expected from the reported pedigrees. From kinship analysis, we identified two errors in reported pedigrees due to incorrect sample IDs and corrected them before further analysis.

Principal components analysis: We visualized population structure using principal components analysis to identify potential outlying ancestral differences using EIGENSTRAT (Price et al. 2006). Population outliers were not excluded and this information was taken into account when examining summary data output from the sequencing pipeline to identify poor quality subjects.

Sample exclusion: All summary data output from the CGI pipeline were pooled together to generate distributions for all metrics provided. Subjects were filtered based on the number of outlying metrics by variant class.

- **Outlier variant statistics:** Subjects were excluded from SNV analysis if they have ≥3 outlying (|z-score| ≥3) metrics (which may have included coverage variability) or a disproportionately high number of novel variants. However, we utilized ancestry information inferred from the PCA analysis so that we did not filtered out samples with African ancestry even if they presented outlying number of SNPs (|z-score|≥3).
- **Mendelian Inheritance Errors**: We also eliminated individuals from the analysis if they presented relatively high numbers (|z-score|>3) of Mendelian inheritance errors where data were available.

1.3. Variant identification and quality control

Initial variant calls were produced by the standard CGI variant calling pipeline by processing samples individually. It has been reported that multi-sample filtering, which leverages information across samples, is more effective to filter out possible experimental artifacts. We used the GotCloud pipeline to filter out variants using multi-sample statistics (Jun et al. 2015). GotCloud uses a support vector machine (SVM) classifier, a machine-learning algorithm, to separate likely true positive and false positive variant sites. For SVM filtering, we collected a series of features related to the quality of each potential variant site, including sequencing depth, fraction of samples with coverage, fraction of reference bases in heterozygous individuals (allele balance, ideally close to 50%), correlation of alternative alleles with strand and read position (strand and cycle bias), and fraction of Mendelian inconsistencies in genotypes. The features were collected from sequence reads on each variant site, which were stored in the CGI-generated BAM files containing reads from the variant regions. We also applied stringent call rate filters by removing SNVs with call rates less than 97% and indels and substitutions with call rates less than 99%. Different cut-offs were used to ensure comparable genotype accuracies for SNVs and other categories, because indels had higher genotype error rates than SNVs with the same call rates. Genotype error rates were measured by Merlin using Mendelian inconsistencies within extended pedigrees (Abecasis et al. 2002).

Copy number variations (CNV) reported by the default CGI pipeline were generated by processing each sample's sequencing depth information only, without further using read mapping or multi-sample statistics. Consequently, default CG CNV variant calls are based on each 2Kb genomic intervals and have relatively higher error rates than short variants. We developed a novel algorithm that first identifies candidate intervals for large deletions by read mapping information and genotypes using multi-sample distribution of sequencing depths. In this method, we utilize SV candidate information based on read junctions generated by CGI software pipeline in addition to the depth-based CNV candidates in the standard CGI CNV calls. Then we collect GC-corrected read depth information on all candidate regions from all samples and perform expectation-maximization (EM)-based clustering with mixture of Gaussian models. Clustering results are then filtered based on the overlap of Gaussian components and Bayesian information criterion (BIC). We identified 3,144 bi-allelic large deletions in from sequenced samples with estimated genotype error rate of 0.95%, as measured by Merlin. The same algorithm was also applied to generate CNV calls on 433 CGI-sequenced samples in the 1000 Genomes Project Phase 3. Details

about this algorithm is described in the supplementary material (Section 5.5) of the 1000 Genomes Project's structural variation paper (Sudmant et al. 2015).

1.4. Pedigree-based imputation

We applied a family based imputation procedure to infer rare genotypes for individuals who were not sequenced using available GWAS data as a scaffold. These individuals were genotypes using four district GWAS platforms: Illumina HumanHap550v3, Human1M-Duov3, Human1Mv1, and Human660W-Quad_v1 GWAS array. Initially we defined the set of genetic variants present in all platforms and using this information we imputed the genotypes of all missing genetic variants. The known pairwise sample relationship was used to detect and correct any Mendelian inheritance errors. The whole genome sequencing imputation method consisted of two steps: 1) construction of sub-pedigree around the imputed sample and 2) imputation within sub-pedigree.

For construction of sub-pedigree around the imputed sample, we first selected the five most closely related sequenced individuals for a given GWAS sample. Next, to improve IBD estimation, we added the sequenced parents of each selected individual. Finally, we generated sub-pedigrees by removing all uninformative individuals (not selected and not needed to connect selected individuals in the pedigree).

For imputation within each sub-pedigree, we used Merlin (Abecasis et al. 2002). The resulting output contained the expected number of copies for the tested allele (dosage), the posterior probabilities for the three alternative genotypes, and the most likely genotype (the genotype with the highest posterior probability, set to missing if the highest posterior probability < 0.5).

Family-based imputation provides highly accurate imputation especially for rare variants because there is little ambiguity in the transmission of rare alleles. We evaluated accuracies of imputed genotypes with ExomeChip genotypes. Genotype accuracy was 96.7% for rare (MLE MAF<1%) variants imputed as heterozygotes, and it was 99.3% for rare variants imputed as homozygotes, much higher than 60% accuracy for population MAF<1% variants by population-based imputation as reported in the 1000 Genomes project.

2. Statistical analysis

2.1. Phenotype transformation and covariate selection

We analyzed T2D related metabolic traits: fasting glucose, fasting insulin, 2 hour glucose, 2 hour insulin total cholesterol level, LDL cholesterol level, HDL cholesterol level, and triglyceride level. Trait values were measured at up to five exams. Regressions were performed at each exam adjusting for covariates as appropriate, producing exam-specific residuals. The exam-specific residuals were then averaged over multiple measurements and an inverse-normal transformation was applied to averaged residuals. Covariates were chosen to align with strategies taken by consortia participating in meta-analysis of GWAS of the given traits, as well as the T2D-GENES and GoT2D consortia's trait transformation strategy (Fuchsberger et al. 2016) and included age, age^2, sex, and BMI. T2D samples were excluded from glycemic trait analyses and cholesterol levels were pre-adjusted by a fixed amount per lipid medication status.

2.2. Empirical kinship estimation

To account for known and unknown relatedness among study participants, we used empirically estimated kinship matrices. We estimated kinship matrices using EMMAX (Kang et al. 2010) included in the EPACTS (http://genome.sph.umich.edu/wiki/EPACTS) software package. Kinship coefficients were estimated from GWAS genotypes for most of the samples. For 10 samples with sequence data only, sequenced genotypes overlapping with GWAS markers were used for kinship estimation.

2.3. T2D association analysis

We tested for association with each genetic marker using a mixed model variance component approach implemented in SOLAR (Blangero et al. 2013). For residual non-independence, we used the empirical kinship matrix described above. The variance component approach uses a likelihood ratio test employing a standard measured genotype fixed effect to identify associated variants. Recent simulation results suggest that simple analysis of affection status as a binary pseudo-quantitative trait recovers a valid test and the original parameters of the true liability scale after suitable transformation. We performed our analyses using this much faster approximation and confirmed that the results from pseudo-quantitative approach conform well to the probit model (Duggirala et al. 1997; Williams et al. 1999).

2.4. Quantitative trait analysis

For association analyses of quantitative and expression traits, we used FamRVtest (Feng et al. 2015) with a variance component model using the empirical kinship matrix estimated by EMMAX.

2.5. eQTL analysis

We obtained whole-genome expression profiles using the Illumina Sentrix Human Whole Genome (WG-6) Series I BeadChip. After performing both BLAT (Kent 2002) and LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) analyses on 47,293 probe sequences, we removed probes that could not be mapped, probes that were previously reported as not expressed, and probes that harbored a sequenced variant nearby that had a potential to cause hybridization artifacts. We performed single variant eQTL association analysis using FamRVtest on the remaining 21,677 transcripts. We defined a *cis*-region as 1Mb from a gene or probe's start and end coordinates, and defined trans as all other regions. We set the significance threshold of $\alpha = 7.0\times10^{-6}$ for *cis*-eQTL based on an estimated FWER of 5% by permuting *cis*-regions with random genes in different chromosomes. Using the permuted expression traits, we set the FWER threshold by sorting the top signals from all expression traits and selecting the p-value at the 5% (1083-th out of 21,677) point.

2.6. Gene-based rare variant aggregated test

We aggregated rare variants with MLE MAF < 1% in exonic regions for gene-based tests using MAFs estimated by SOLAR using maximum-likelihood estimation (Blangero et al. 2013). We generated four different masks based on the functional annotations as explained in the main text. We categorized protein truncating variants (PTVs) and missense variants using five different functional prediction algorithms (LRT, Mutation Tester, PolyPhen2-HumDiv, PolyPhen2-HumVar, SIFT) into variants predicted to be

deleterious by at least one of five algorithms (NS_{broad}) and variants predicted to be deleterious by all five algorithms (NS_{strict}) (Chun and Fay 2009; Kumar et al. 2009; Adzhubei et al. 2010; Schwarz et al. 2010).

We performed aggregation tests for T2D and quantitative traits using FamRVtest in a variance component framework. We employed the collapsing test (CMC count) (Li and Leal 2008), collapsing with differential weighting(Madsen and Browning 2009), kernel-based (SKAT) (Wu et al. 2011), and variable threshold (VT) tests (Price et al. 2010).

3. Power analysis

3.1. Single marker power for private variants - quantitative traits

We computed the power to detect a private risk variant (defined as a singleton among founders) in our pedigrees via simulation. We first created realizations for the segregation of private alleles within these pedigrees using the gene-dropping simulation routine in Merlin. For each of the 413 founders, we simulated the segregation of 10,000 unlinked singleton variants through their respective pedigree lineage and recorded the observed allele count (AC) in sequenced or genotyped/imputed samples. We estimated the probability distribution for private variant observed AC in our pedigrees, $Pr(AC = k)$, as the fraction of gene-drop realizations with exactly k observed alleles in the pedigrees.

Power to detect a private risk allele conditional on being sampled in a founder (Figure 2A): We computed the power to detect a private risk allele for a quantitative trait conditional on observed AC. For each simulated private variant, we augmented the observed BMI transformed trait value of carriers by an effect size of δ and tested each variant individually for single marker association using EPACTS. We used the empirical transformed BMI values as the baseline simulated phenotypes because (1) the trait showed no evidence of genetic effect for private variants and (2) using the empirical phenotypes retains any potential phenotypic correlations among pedigree members due to observed or unobserved covariates in the real data. We controlled for relatedness using the empirical kinship coefficient matrix in a mixed model framework. We computed the power at level α for effect size δ conditional on a given allele count AC, $P_{AC}(\delta, \alpha)$, as the fraction of gene drop realizations with exactly AC observations having association p-value <α. We then computed association power for a private variant observed in the dataset with effect size δ at level α as $P(\delta, \alpha) = \sum_{AC} P_{AC}(\delta, \alpha) \times Pr(AC)$.

Power to sample and then detect a private risk allele (Figure 2B): Next, we accounted for the fact that a private variant must be captured by exactly one of the 413 pedigree founders (or 826 founder alleles) to be eligible for association testing. Assuming a private risk allele with population allele frequency p and effect size δ, the power to capture and detect that risk allele is then $P(p, \delta, \alpha) = (826 \times p \times (1-p)^{825}) \times P(\delta, \alpha)$. Finally, if we assume that N such private risk alleles each with population allele frequency p and effect size δ exist in the population, our power to detect at least one of them in our pedigrees is $1 - \left(1 - P(p, \delta, \alpha)\right)^N$, and the number of such risk variants we expect to detect is $N \times P(p, \delta, \alpha)$.

3.2. Single marker power for private variants - dichotomous traits

We used a similar procedure to estimate power for dichotomous traits using the actual T2D phenotypes. In this analysis, for each private variant gene drop realization, we flipped the phenotype of carriers from an unaffected to a case with a fixed probability that we treated as a penetrance parameter. If a carrier was in truth a case we left it as a case.

3.3. Power of single variant eQTL association tests

To simulate eQTL power for both rare and common variants, we repeated the gene dropping technique used for private variants but allowed for the possibility of variants to enter the dataset through multiple founders. We first computed power conditional on a fixed allele count in founders AC_F. For each AC_F, we simulated 10,000 realizations in which we dropped the risk allele through AC_F randomly selected pedigree lineages and augmented the observed expression level of each carrier by effect size δ. We selected an expression trait (hmm9375-S) that did not have any statistically significant eQTL results and therefore represented an approximate null trait distribution. The power $P(\delta, AC_F)$ to detect an eQTL with effect size δ sampled in AC_F pedigree founders was computed as the fraction of 10,000 gene drops realizations statistically significant at $\alpha = 7.0\times10^{-6}$.

We estimated the power to sample and detect an eQTL of effect size δ and MLE MAF p is then $\sum_N P(\delta, AC_F = N)\times \Pr(AC_F = N \mid MAF)\times \Pr(MAF)$. We used the empirical site frequency spectrum derived from 3,886 Mexican-American chromosomes in a large, multi-ethnic exome sequencing study of T2D (Fuchsberger et al. 2016) as the prior for allele frequency $Pr(MAF)$, and used Monte Carlo draws to compute the above power.

3.4. Possible limitations/pitfalls on power estimation

The power simulations were designed in a fair and realistic manner. However, a major assumption of our simulations is perfect genotyping and imputation of alleles segregating through the pedigrees. As evidenced by Figure 2A of the main text, the power to detect an association is highly dependent upon the number of times that allele is observed in the pedigree, particularly when allele count is low. Not surprisingly then, imperfect genotyping and imputation would reduce power since fewer risk alleles will be called than actually exist in the pedigrees. Our evaluation of imputation accuracy showed that genotype calls are of high quality and likely to have minimal impact on the power analyses. We have also presented power results assuming relatively large effect sizes for rare risk variants. Whether alleles of such magnitude exist for our phenotypes is unclear. However, we clearly state that this enrichment through pedigree sequencing strategy would require very large effect sizes to be successful at the given sample sizes.

4. Supplementary results

The traits tested for genetics associations were T2D and quantitative traits: fasting glucose, fasting insulin, 2-hour glucose, 2-hour insulin, LDL cholesterol, HDL cholesterol and triglyceride.

4.1. T2D single variant analysis

Among 153 common variant loci with published associations with T2D, our study shows nominal significance ($p<0.05$) with SNVs near *KCNQ1* (p=0.004), *CDC123*/*CAMK1D* (p=0.04), *TSPAN*/*LGR5* (p=0.03), *C2CD4A*/*C2CD4B* (p=0.01), *GCKR* (p=0.02), and *ADAMTS9* (p=0.009). No single variant reached genome-wide significance for T2D ($p<5\times10^{-8}$).

4.2. Quantitative trait single variant analysis

Two common SNVs in a small region of chromosome 7 located 42 kb downstream of the gene *PDK4* showed marginal genome-wide significant associations with 2-hour glucose (rs10249057 with $p<4.4x10^{-8}$ and rs10268456 with $p<4.4\times10^{-8}$). The ENCODE project data showed that the chromosomal region where the variant rs10268456 is located has an open chromatin formation and is accessible to transcription factors (Encode Project Consortium 2004). These variants also showed interesting pleiotropic effects on other diabetes related traits including 2-hour insulin concentration. The pyruvate dehydrogenase kinase 4 (*PDK4*) gene is part of the glycolysis/glucogenesis gene pathway and its role in T2D has been speculated on with no clear conclusions (Cadoudal et al. 2008). The gene product enhances hepatic glucogenesis through inhibition of the PDC complex and thus preservation of pyruvate for gluconeogenesis, and is upregulated during starvation and insulin resistant state. The association of the transcription factor *TCF7L2* and T2D is well known and interestingly the *PDK4* expression is mediated by the presence of *TCF7L2*. We found that rs964184, a known GWAS SNV (Geier et al. 2010; Global Lipids Genetics Consortium et al. 2013) for hypertriglyceridemia in Hispanics, is associated with triglyceride levels ($p=4.8x10^{-8}$).

4.3. Gene-based rare variant association analysis

In gene-based aggregation tests of rare variants, we found no gene significantly associated with T2D with gene-wide significance ($\alpha<2.5\times10^{-6}$). In the gene-based analyses of quantitative traits, we found significant association of *LDLR* with LDL cholesterol levels, *CYP3A4* with fasting glucose, and *OR2T11* with 2-hour insulin levels (Supplementary Table S1). Although *LDLR* is a known gene previously reported to be associated with LDL cholesterol in various studies (Geier et al. 2010; Global Lipids Genetics Consortium et al. 2013), we could not find additional evidences of association for *CYP3A4* and *OR2T11*. We then investigated individual rare variants included in the analyses in these genes (Supplementary Table S2). Out of seven variants included in the *LDLR*, five of them were private variants unique to a pedigree, although main association signal was driven by a non-private variant (19:11210913 G/A). Three rare variants out of four in the *CYP3A4* gene-based result were private variants, but all these three variants were singletons that were not enriched through the pedigrees. One private and one non-private rare variants were included in the *OR2T11* result. The non-private variant in *OR2T11* was defined as 'rare' based on MLE MAF, but had >1% MAF in the 1000 Genomes Project data.

4.4. Candidate regions follow up.

Linkage: Each normalized phenotype was used to perform a linkage association scan using previously defined genome-wide IBD identities of all T2D-GENES participants. A peak was considered significant if presented a minimum LOD-score of 3 and the limits of the linkage peak defined when a reduction of

single LOD-score was observed. The SNVs located under a linkage peak were selected, and a new local multiple hypotheses significance threshold was defined for each linkage peak. The WGS genotype-phenotype association statistics are now interpreted using this corrected significance threshold.

GWAS and Monogenic gene lists: We performed a comprehensive scientific literature search and identified all genotypic associations for T2D and diabetes related traits as available on 7/8/2013. We compiled a list of genes that have been implicated in monogenic diabetes, obesity, severe insulin resistance and lipodystrophy (Fuchsberger et al. 2016). The genomic coordinates of those associations were updated using LiftOver to make these coordinates comparable with T2D-GENES sequencing data. For each GWAS variant, we defined a 500kb window upstream and downstream of the previously detected association. Similarly, we defined a 500kb window upstream and downstream from the monogenic genes. The set of genetic variants identified inside this window was defined and a new multiple hypotheses significance threshold was defined. The genotypic associations are interpreted using this newly corrected significance threshold.

5. Comparing effect sizes of rare and common variants

Each SNV located near a probe was classified as rare or common based on its SOLAR-estimated MLE MAF. Genetic variants were classified as rare if MLE MAF<0.01 and as common otherwise. The association results from rare and common variants within eQTL cis-regions were compared in two different ways: 1) using only statistically significant associations and extrapolating using simulated power and 2) using all p-values to estimate overall distributions of effect sizes.

5.1. Estimation of common and rare variant contributions from significant associations and power simulations

Given the differences in power, it is not surprising that we identified many more common than rare cis eQTLs in our analysis. However, it is possible and indeed likely that many more rare eQTL variants remain to be found. We therefore sought to estimate the potential number and contribution to heritability of rare eQTL variants. To do so we used simulation to compute power to detect rare (MLE MAF <1%) and common (MLE MAF>1%) risk alleles of various effect sizes using our pedigree strategy. Based on this power and our empirical findings, we can estimate the total numbers of rare and common eQTLs that exist in the population. Identifying independent risk alleles of given a frequency and effect size can be thought of as Bernoulli trials with “success” probability equal to single marker association power. If we assume power of α and successfully identify k risk alleles, we estimate the total number of risk alleles of that frequency and effect size that exist in the population to be k/α.

Given that the average effect size of a significant common eQTL was approximately 0.5 standard deviation units, we estimate that the population contains as many as 23K common eQTL variants among the 21,677 traits we tested for eQTL signals, explaining 5.8% of overall trait variances. Accurately predicting the contribution of rare variants is complicated by the fact that effect estimates of genome-wide significant rare eQTLs are biased by the winner’s curse. That is, although our average effect size for significant rare eQTLs was 2.0 standard deviation units, the true average effect is likely smaller.

However, we can still make predictions regarding rare eQTLs in relation to that of common eQTLs. If we assume that rare variants have an average effect size of δ =1.0, so twice that of common variants, we predict that 222K significant rare eQTLs exist in the population and can account for 0.9% of total variances across the expression traits. Because it is based on tests of significance, these are underestimates of the overall effect of common and rare cis-acting variants on gene expression.

Our results indicate that many large effect rare eQTLs remain to be identified by larger, better powered sequencing studies. However, despite the large number of rare eQTLs that may exist, their contribution to overall expression heritability is unlikely to match that of common eQTLs.

5.2. Comparing effect sizes of rare and common variants using all eQTL results

The comprehensive set of eQTLs associations was classified as rare or common variant based on the MLE MAF estimates. The distributions of the signed effect sizes were compared in Figure 5 in the main text where rare genetic variants presented larger effect. The observed effect size variances were 0.022 for common variants and 0.125 for rare variants. This estimate of total effect size variance can be viewed as the sum of two components: the true biological effect size variance and the error variance. A higher error variance for estimated effect sizes is expected for rare variants due to a higher sampling error resulting from the small number of minor allele copies. To correct for this structural difference between rare and common effect size distributions, we calculated the error variance of the effect size estimations using the formula $e = \delta^2/\chi^2$, where *e* is the estimation variance, δ is the estimated effect size and χ^2 is the chi-square obtained in the likelihood ratio test (LRT) applied for the linear mixed model. The Supplementary Table S3 shows the observed distribution of error variances with rare variants exhibiting much higher error variances than common ones. The observed error variances were then subtracted from the observed effect size variances to obtain a composite maximum likelihood estimate of the true biological variance. These estimates were then statistically compared using a variance ratio test.

To assess the potential importance of rare variants on total genetic variance, we performed a simulation using the observed MAF distribution and the observed eQTL biological effect size distributions. We randomly selected one million variants from the total observed variant distribution and categorized them based on the number of minor allele copies as rare or common variants. Depending upon this classification, we then paired each chosen variant with a randomly drawn effect size from the appropriate distribution of standardized eQTL association results. This process generates an empirical convolution of the two random distributions. For each association, we calculated the genetic variance as:

$$G_i = 2\times MLE_i\times(1 - MLE_i)\times \beta_i$$

where MLE_i is the MLE MAF for the genetic variant i and β_i is the observed effect of the genetic variant i in the gene expression g as previously defined by the eQTL association analysis. The overall contribution of rare variants was defined as:

$$C_{rare} = \sum Gr_i \Big/ (\sum Gr_i + \sum Gc_i)$$

where $\sum Gr_i$is the sum of genetic variance of the complete set of *n* rare variant associations and is the sum of genetic variance of the complete set of *m* common variant associations. The total contribution of rare variants C_{rare} on our simulation was 24.97%, a substantial, although minority, proportion of total genetic variance.

6. Empirical Significance

We simulated a set of 1000 phenotypes without heritability ($h^2 = 0$) taking into account the known pairwise sample relationship of our large human pedigrees. For each phenotype, we performed a full WGS-based association analyses for all single-nucleotide variants (SNVs) identified on our data. All association analyses were performed using the linear mixed model implemented in SOLAR, allowing for residual non-independence between relatives. We sorted and defined the most extreme association of each simulated phenotype and constructed an extreme value distribution composed by one thousand extreme p-values (Šidák 1967). The extreme value distribution was fit into a theoretical beta distribution using a maximum likelihood estimation algorithm for estimation of the respective β parameter that is directly related to the effective number of tests being carried. This corrected number of effective tests is used to define a new corrected multiple hypothesis Bonferroni in the form of target $\alpha = 0.05 \ / \ nef$ where nef is the number of effective tests defined by the β parameter (Almeida et al. 2016). Based on our simulations, we identified only a set of 704,225 independent tests being carried at any whole genome association scan. This leads to a new corrected target significance threshold of 7.1×10^{-8}. The Supplementary Figure S2 presents the observed extreme distribution and the theoretical beta distribution that was optimally fitted. Our results clearly showed the high extent of LD contribution on large human pedigrees dealing with whole genome sequencing data.

Supplementary Table S1. Gene-wide significant ($p<2.5x10-6$) association results.

Trait	Gene	Variant Group	Test	P-value
LDL cholesterol	*LDLR*	PTV+missense	SKAT	$8.3x10^{-7}$
Fasting glucose	*CYP3A4*	PTV+NS_{broad}	CMC	$9.2x10^{-7}$
			VT	$1.8x10^{-6}$
2-hour insulin	*OR2T11*	PTV+missense	Madsen-Browning	$1.9x10^{-6}$

Supplementary Table S2. Variants in significantly associated genes.

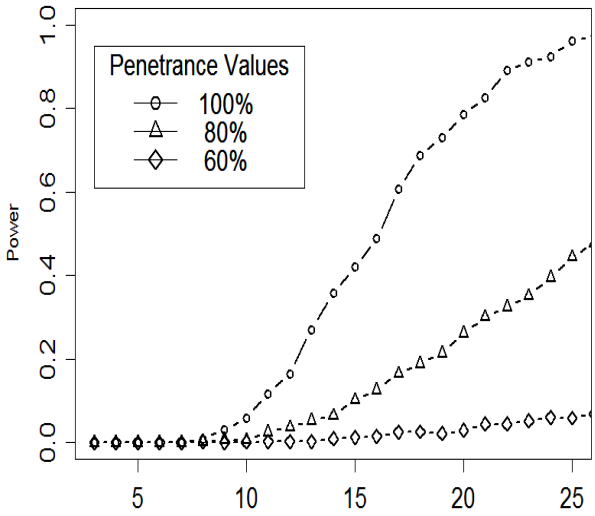
Gene	Variant	Allele	Effect size*	p-value*	Allele count	MLE MAF	Type
LDLR	19:11210913	G/A	1.10	2.54×10^{-6}	20	0.00777	Rare
	19:11211016	C/T	-0.133	0.744	6	0.00155	Private
	19:11218079	G/A	-0.0620	0.929	2	0.00155	Private
	19:11227645	G/A	0.668	0.176	4	0.00310	Rare
	19:11231177	G/A	-0.443	0.195	8	0.00156	Private
	19:11233886	C/T	1.60	0.0171	2	0.00155	Private
	19:11242072	A/AC	-0.0885	0.846	4	0.00156	Private
CYP3A4	7:99355806	G/GT	2.80	1.19×10^{-4}	1	0.00155	Private
	7:99359845	T/C	1.00	0.248	1	0.00155	Private
	7:99366121	A/G	1.49	0.0935	1	0.00155	Private
	7:99367392	C/G	1.91	4.56×10^{-4}	5	0.00311	Rare
OR2T11	1:248789504	C/T	-1.22	2.00×10^{-4}	60	0.00628	Rare**
	1:248789822	A/C	-2.80	2.23×10^{-3}	2	0.00157	Private

* Effect sizes and p-values based on single-variant association.

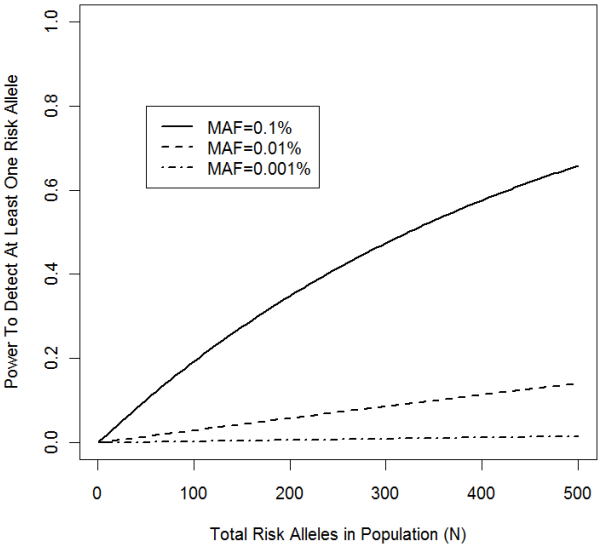
** This variant MLE MAF<1%, but has >1% MAF in 1000 Genomes Project data.

Supplementary Figure S1. Power simulations for dichotomous traits. A) Power to detect private risk variants for T2D conditional on the observed allele count. B) Power to detect at least one of N private T2D risk alleles with 80% penetrance.

A



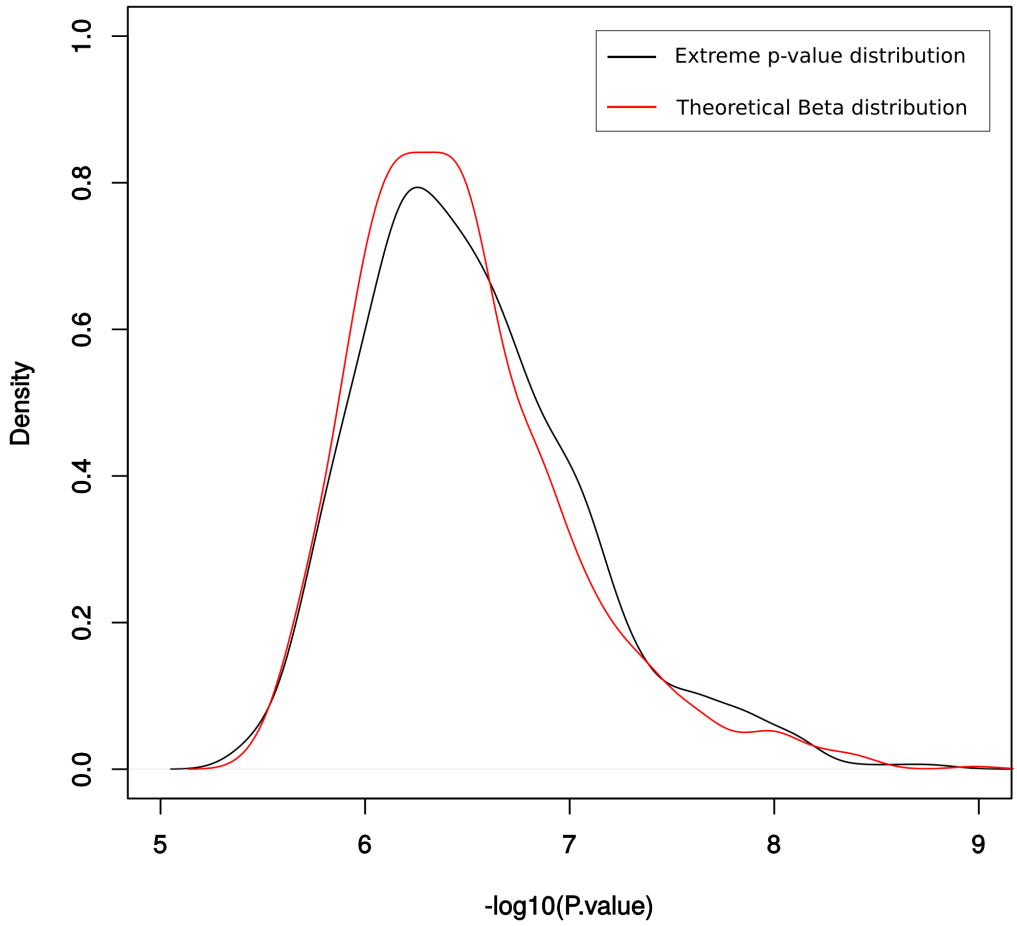
B



Supplementary Table S3. Summary statistics of the sampling error of effect size estimation.

	1st.Q	Median	Mean	3rd.Q
Common	0.004	0.008	0.0191	0.0242
Rare	0.079	0.11	0.1128	0.1443

Supplementary Figure S2. Kernel density plots of a -log transformed p-values of the extreme p-value distribution and the theoretical beta distribution.



References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR. 2002. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**: 97-101.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7**: 248-249.

Almeida M, Blondell L, Peralta JM, Kent JW, Jr., Jun G, Teslovich TM, Fuchsberger C, Wood AR, Manning AK, Frayling TM et al. 2016. Independent test assessment using the extreme value distribution theory. *BMC Proc* **10**: 245-249.

Blangero J, Diego VP, Dyer TD, Almeida M, Peralta J, Kent JW, Jr., Williams JT, Almasy L, Goring HH. 2013. A kernel of truth: statistical advances in polygenic variance component models for complex human pedigrees. *Adv Genet* **81**: 1-31.

Cadoudal T, Distel E, Durant S, Fouque F, Blouin JM, Collinet M, Bortoli S, Forest C, Benelli C. 2008. Pyruvate dehydrogenase kinase 4: regulation by thiazolidinediones and implication in glyceroneogenesis in adipose tissue. *Diabetes* **57**: 2272-2279.

Carnevali P, Baccash J, Halpern AL, Nazarenko I, Nilsen GB, Pant KP, Ebert JC, Brownley A, Morenzoni M, Karpinchyk V et al. 2012. Computational techniques for human genome resequencing using mated gapped reads. *J Comput Biol* **19**: 279-292.

Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes. *Genome Res* **19**: 1553-1561.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**: 78-81.

Encode Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**: 636-640.

Feng S, Pistis G, Zhang H, Zawistowski M, Mulas A, Zoledziewska M, Holmen OL, Busonero F, Sanna S, Hveem K et al. 2015. Methods for association analysis and meta-analysis of rare variants in families. *Genet Epidemiol* **39**: 227-238.

Fuchsberger C Flannick J Teslovich TM Mahajan A Agarwala V Gaulton KJ Ma C Fontanillas P Moutsianas L McCarthy DJ et al. 2016. The genetic architecture of type 2 diabetes. *Nature* doi:10.1038/nature18642.

Geier CF, Terwilliger R, Teslovich T, Velanova K, Luna B. 2010. Immaturities in reward processing and its influence on inhibitory control in adolescence. *Cereb Cortex* **20**: 1613-1629.

Global Lipids Genetics Consortium Willer CJ Schmidt EM Sengupta S Peloso GM Gustafsson S Kanoni S Ganna A Chen J Buchkovich ML et al. 2013. Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**: 1274-1283.

Hunt KJ, Lehman DM, Arya R, Fowler S, Leach RJ, Goring HH, Almasy L, Blangero J, Dyer TD, Duggirala R et al. 2005. Genome-wide linkage analyses of type 2 diabetes in Mexican Americans: the San Antonio Family Diabetes/Gallbladder Study. *Diabetes* **54**: 2655-2662.

Jun G, Wing MK, Abecasis GR, Kang HM. 2015. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research* **25**: 918-925.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**: 348-354.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**: 1073-1081.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83**: 311-321.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**: e1000384.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**: 2867-2873.

McClellan J, King MC. 2010. Genetic heterogeneity in human disease. *Cell* **141**: 210-217.

Mitchell BD, Kammerer CM, Blangero J, Mahaney MC, Rainwater DL, Dyke B, Hixson JE, Henkel RD, Sharp RM, Comuzzie AG et al. 1996. Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. The San Antonio Family Heart Study. *Circulation* **94**: 2159-2170.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei L-J, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics* **86**: 832-838.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904-909.

Puppala S, Dodd GD, Fowler S, Arya R, Schneider J, Farook VS, Granato R, Dyer TD, Almasy L, Jenkinson CP et al. 2006. A genomewide search finds major susceptibility loci for gallbladder disease on chromosome 1 in Mexican Americans. *Am J Hum Genet* **78**: 377-392.

Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**: 575-576.

Šidák Z. 1967. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association* **62**: 626-633.

Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, Zoledziewska M, Mulas A, Pistis G, Steri M, Danjou F et al. 2015. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet* **47**: 1272-1281.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75-81.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**: 82-93.