Power of deep, all-exon resequencing for discovery of human trait genes

Gregory V. Kryukov[a], Alexander Shpunt[a,b], John A. Stamatoyannopoulos[c], and Shamil R. Sunyaev[a,1]

[a]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115; [b]Department of Physics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139; and [c]Department of Genome Sciences, University of Washington, 1705 Northeast Pacific Street, Seattle, WA 98195

Communicated by Helen H. Hobbs, University of Texas Southwestern Medical Center, Dallas, TX, December 19, 2008 (received for review June 30, 2008)

The ability to sequence cost-effectively all of the coding regions of a given individual genome is rapidly approaching, with the potential for whole-genome resequencing not far behind. Initiatives are currently underway to phenotype hundreds of thousands of individuals for major human traits. Here, we determine the power for de novo discovery of genes related to human traits by resequencing all human exons in a clinical population. We analyze the potential of the gene discovery strategy that combines multiple rare variants from the same gene and treats genes, rather than individual alleles, as the units for the association test. By using computer simulations based on deep resequencing data for the European population, we show that genes meaningfully affecting a human trait can be identified in an unbiased fashion, although large sample sizes would be required to achieve substantial power.

association studies | polymorphism | rare variants | sequencing

Whole-genome association studies based on genotyping have recently demonstrated potential for identifying SNPs and haplotypes associated with a range of common clinical phenotypes (1–3). However, only a small fraction of observed phenotypic variation is currently attributable to identified allelic variants. Association studies are fundamentally limited by previously known genetic variation, featuring predominantly high-frequency SNPs. By contrast, deep resequencing has the potential to reveal a vast trove of low-frequency alleles. Low-cost–high-throughput sequencing technologies hold the potential to propel discovery of gene–phenotype associations incorporating low-frequency allelic variation on a large scale.

Although knowledge of all variants segregating in the population would seem to increase the power of genetic analysis, this prospect faces daunting statistical challenges, because an expanding pool of variants requires more stringent multiple testing correction, whereas the power to detect association with low-frequency variants is reduced. This problem may be surmounted by pooling allelic variants in a single candidate gene (4–6) or pathway (7–9). However, if most variation in a gene or pathway is neutral, this pooling strategy will not provide a sufficient signal-to-noise ratio (10). To enrich variation in functionally significant alleles, the analysis should be limited to nonsynonymous coding variation as one obvious functional class. Site-directed mutagenesis and comparative genomics have shown that the large fraction of de novo missense mutations are of functional significance (11–14). Consequently, many mildly deleterious coding variants are expected to be segregating in the human population at low allele frequencies, as was originally proposed by Tomoko Ohta in the "nearly neutral theory" of molecular evolution (15). Indeed, the statistically significant excess of combined rare missense variation in individuals at phenotypic extremes was detected in candidate gene studies for several phenotypes (4–6).

Results

Simulation of Resequencing Studies. Although success with highly targeted candidate gene resequencing has been reported (4–6), the potential for unbiased discovery of new gene–phenotype associations by resequencing large numbers of genes in large numbers of individuals has not been considered. It is therefore unclear whether clinical populations of realistic size would provide sufficient statistical power for new gene discovery by using this approach. We address these questions through simulation of resequencing studies (Fig. 1*A*). We focus on quantitative rather than qualitative phenotypes (Fig. 1*B*). However, our strategy involves comparison of two groups of individuals at phenotypic extremes and, therefore, can be extended to dichotomous traits with specified penetrance. Quantitative traits allow for additional flexibility in selecting most informative individuals. Our simulations make no assumptions about the existence of specific causal variants with specified allele frequencies. Instead, they rely on the influx of new mutations and resulting collective effect of low-frequency hypomorphic alleles.

Assessing the feasibility of identifying human trait genes from "all-exon" sequence data is tantamount to determining the potential for detecting the impact of cumulative variation within an individual gene on a phenotype at a high level of statistical significance ($2.5 \cdot 10^{-6}$, equivalent to a level of $P < 0.05$ following Bonferroni correction for 20,000 genes). We combine multiple rare variants from the same gene and treat genes rather than individual alleles as the units for the association test.

We based our computational model of coding human genetic variation on existing population sequencing data and extrapolated this model to even larger population samples. We analyzed the deepest systematic resequencing dataset currently available, comprising 58 genes (exons plus flanking intronic and intergenic regions) that were resequenced in 757 individuals of European ancestry (8). The present analysis is focused exclusively on individuals of European ancestry because of the lack of deep resequencing data from other populations.

The feasibility of gene mapping by resequencing coding regions depends on mutation rate, population demographic history, selection coefficients, and phenotypic effects associated with new missense mutations. We assumed a mutation rate of $1.8 \cdot 10^{-8}$ per generation (16), and used the conventional four-parametric model of the history of the European population with long-term constant size followed by a bottleneck and then by an exponential expansion (17).

Demographic Model. To estimate parameters of the demographic model, we computed a likelihood function for the observed site-frequency spectrum of synonymous and noncoding SNPs by using diffusion approximation of the Wright–Fisher model. We used the infinite-number-of-sites model and

Author contributions: G.V.K., J.A.S., and S.R.S. designed research; G.V.K. and A.S. performed research; G.V.K., A.S., and S.R.S. analyzed data; and G.V.K., A.S., J.A.S., and S.R.S. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: ssunyaev@rics.bwh.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0812824106/DCSupplemental.

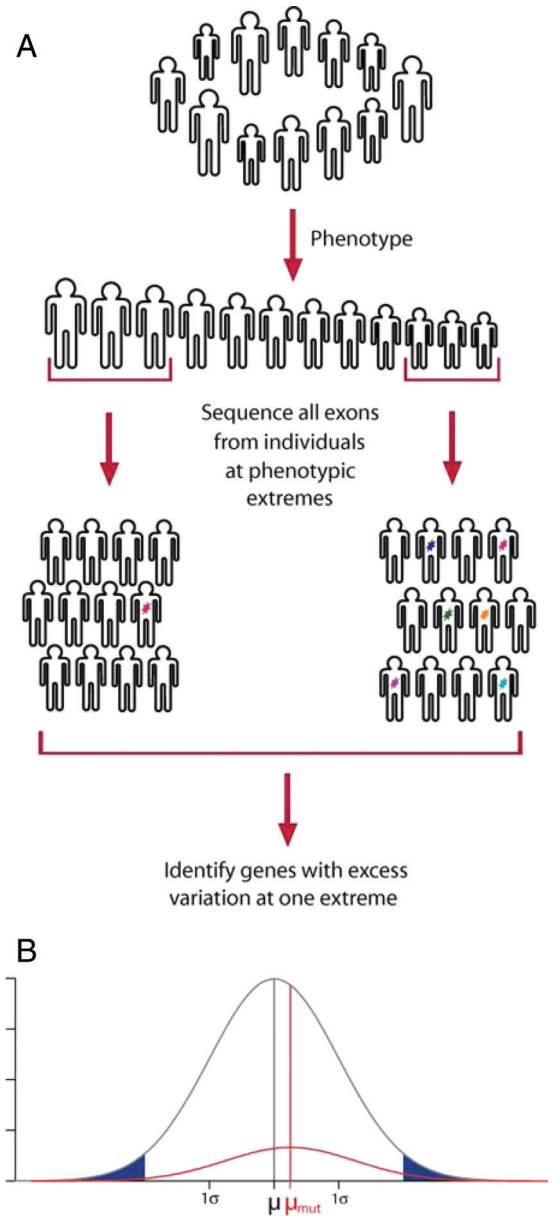

Fig. 1. Simulated resequencing study. (*A*) Design of simulated resequencing study. Color marks represent various new mutations discovered in sequenced individuals. (*B*) Modeling the effect of mutations effect on phenotype. Distribution of quantitative trait for noncarriers is shown in gray. Distribution of the same quantitative trait for individuals that carry at least one moderately deleterious mutation is shown in red. Observed distribution of QT in the whole population is the sum of these distributions. Individuals from the phenotypic extremes marked by blue are subjects for resequencing.

obtained a time-dependent solution for the population with the variable effective size (18, 19) (see *Materials and Methods*). We verified this analytical approach using forward simulations [supporting information (SI) *Appendix*, Fig. S1].

The likelihood has a single well-defined maximum in the space of the four parameters of the model (Fig. 2, *SI Appendix*, Fig. S3). The largest uncertainty was observed in the bottleneck population size due to sparse data on high-frequency SNPs in the resequencing dataset that we used.

Our demographic model reproduces observed site-frequency spectrum well (Fig. 3*A*). Under this model, the spectrum of neutral genetic variation in European populations is explained best by population growth that started approximately 7,500–9,000 years ago (average, 20–25 years per generation), coinciding with the spread of agriculture in Europe (20). As expected, our analysis of deep resequencing data resulted in a larger estimate of current effective population size (900,000) compared with demographic models constructed by using smaller population samples (17, 21–25). Explicit modeling of exponential population growth enabled us to extrapolate more accurately and thus simulate outcomes of resequencing projects involving significantly larger numbers of individuals.

The deep resequencing dataset used in our study is not informative about the ancient population bottleneck, a well-known feature of the demographic history of Europeans. This is not surprising because almost all variants in this dataset have low population frequency. According to the demographic model, essentially all rare variants originated by mutations in the recent population growth phase, whereas most frequent SNPs originated by mutations prior to the growth phase (*SI Appendix*, Fig. S4*A*).

To ensure that this is not an artifact of our method, we applied it to SeattleSNPs sequencing dataset, which covers as many as 320 genes in as few as 23 individuals of European origin. The inferred population history is in general agreement with previous coalescence-based analyses and reproduces a deep ancient bottleneck (21–25). However, SeattleSNPs dataset is not informative about the recent population growth, i.e., different growth rates have comparable likelihoods. Combined analysis of the two datasets (maximizing the product of the two likelihoods) results in a demographic history, which includes both deep ancient bottleneck and recent growth. This demographic history model generates numbers of rare SNPs highly similar to the model based solely on the deep sequencing dataset (*SI Appendix*, Fig. S4*B*). This model also predicts that the overwhelming majority of rare variants originated in the recent population growth phase.

It is likely that modeling the demographic history of Europeans by using datasets rich in both rare and frequent variants would require more complex demographic models. However, because almost all rare variants originated in the population growth phase, the recent history mostly determines results of our study. The parameters of our model corresponding to the recent history were estimated with a high degree of confidence by using the deep resequencing dataset. Therefore, we present here results obtained by using this demographic model.

Distribution of Selection Coefficients. As a next step we added a distribution of selection coefficients for new missense mutations to the model and estimated this distribution from the site-frequency spectrum of nonsynonymous SNPs by using forward simulations. We modeled a distribution of selection coefficients by a gamma distribution with parameters estimated by maximum likelihood. Simulated site-frequency spectrum agrees well with the observed spectrum for nonsynonymous SNPs (Fig. 3*B*). All suboptimal distributions consistent with the data have the maximum probability mass in the range of selection coefficients between 0.0006 and 0.004 (Fig. 4), which is in good agreement with multiple recent studies (11–14). There is a consensus that the majority of new missense mutations are moderately deleterious both in humans (11–14) and flies (26).

We note that this analysis assumed complete neutrality of synonymous sites. Incorporating weak negative selection acting on synonymous sites would result in slightly higher estimates of selection coefficients for nonsynonymous mutations and in lower estimates of the recent population growth rate.

Simulation of Phenotypic Variation. To simulate population phenotypic variation, we first simulated nonsynonymous genetic variation in the average human protein coding gene (~500 aa) by using the estimated demographic history and distribution of selection coefficients. To simulate corresponding quantitative trait (QT)

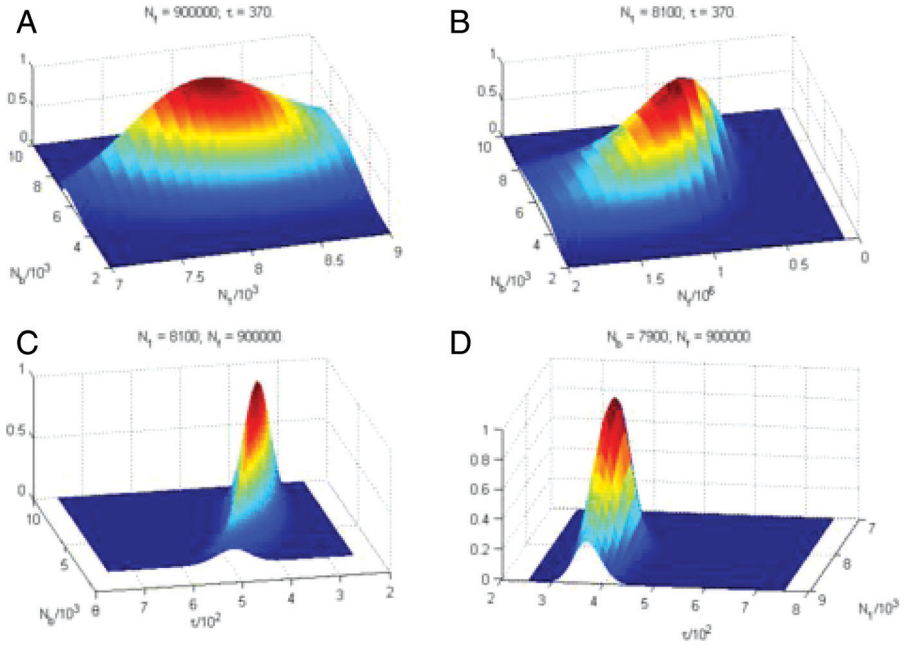

Fig. 2. Two-dimensional sections of the likelihood surface for the demographic model that was fitted to the systematic resequencing data. Population history model, long-term constant population size is followed by a bottleneck and subsequent exponential population growth. The model has four parameters and limited to the European population: N_1, ancestral population size; Nb, bottleneck population size; N_f, final population size; τ, time of the population expansion since the bottleneck.

variation, we used a simple model of genotype–phenotype relationship. We assumed that QT is distributed normally and individuals carrying at least one functional nonsynonymous allele have QT values distributed with a shifted mean, but with the same variance (Fig. 1*B*). In our simulation we subdivided new nonsynonymous mutations into functional (i.e., affecting QT) and nonfunctional classes by using two different approaches. First, we assumed all missense mutations with selection coefficient $>10^{-4}$ to be functional. This assumption implies neither that a given mutation is under selection due to its effect on QT of interest, nor that QT of interest is under selection at all. The simple reasoning behind this assumption is that if amino acid change is visible to purifying selection through its effect on at least some traits, it affects protein molecular function.

In addition, to analyze a possibility that the effect of mutations on QT of interest is independent of the integrated strength of natural selection (either direct or pleiotropic), we randomly assigned mutations to functional and nonfunctional classes irrespective of the selection coefficients associated with them.

We further assumed that all functional missense mutations in a given gene bias QT in one direction. This assumption is based on the fact that the vast majority of de novo amino acid mutations with a measurable effect reduce protein activity. Gain-of-function

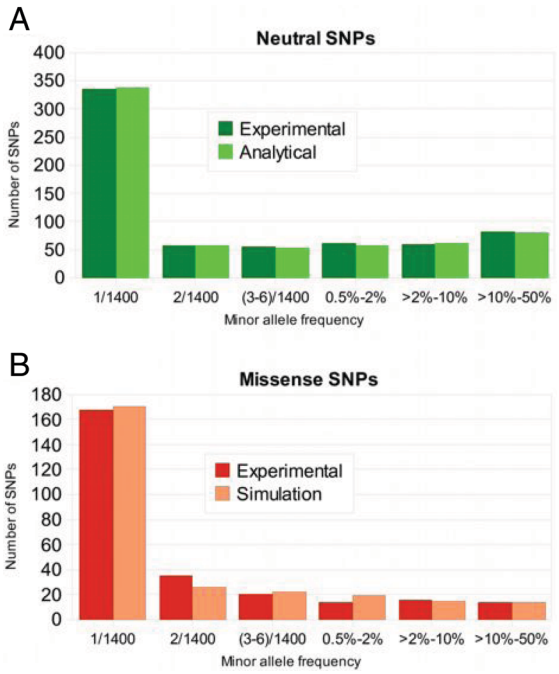

Fig. 3. Agreement of experimental allele frequency spectra with the modeled spectra (*A*) neutral SNPs and (*B*) missense SNPs.

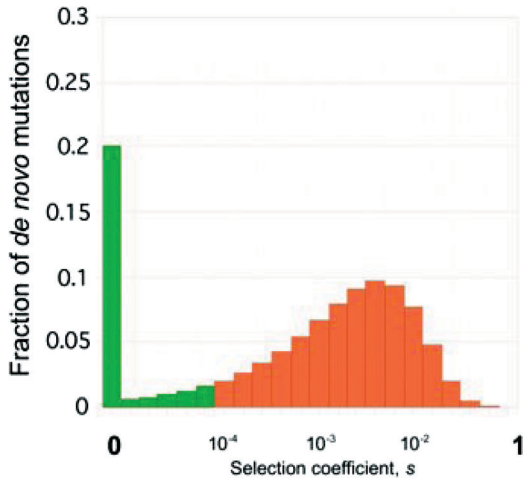

Fig. 4. Distribution of selection coefficients for de novo missense mutations. Distribution was modeled by gamma distribution and fitted to deep resequencing data by the maximum likelihood method. Mutations with selection coefficient $<10^{-4}$ were assumed to have no effect on quantitative phenotype in our model and are shown in green. Mutations assumed to be functional are shown in orange.

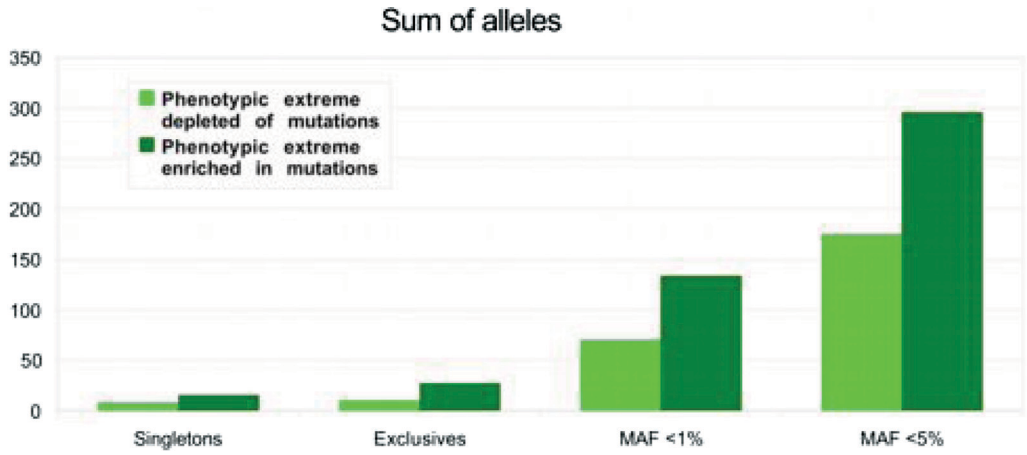

Fig. 5. Excess of missense variants at one of the phenotypic extremes. Data are shown for simulation of 5,000 individuals sequenced at each of two 5% phenotypic extremes. Sums of the alleles (cumulative frequencies of pooled variants) were averaged over 10,000 simulations. Shift of quantitative trait median in mutation carriers is assumed to be equal to 0.5σ.

mutations are much less frequent and are restricted to specific residues or domains. Although very important in a number of rare, Mendelian syndromes, gain-of-function mutations do not represent a sizable fraction of human deleterious variation, and thus can be disregarded in our analysis. Loss-of-function mutations in different genes can bias QT in different directions (depending on the molecular function and position in biological networks), but loss-of-function mutations in the same gene should bias QT in the same direction.

The exact distribution of effects of new missense mutations on QTs is unknown, but on average, de novo missense mutation in a gene highly relevant to a given phenotype should have phenotypic effect smaller than new complete loss-of-function mutations but larger than segregating nonsynonymous SNPs.

De novo mutations associated with dominant syndromes caused by haploinsufficiency are known to cause shifts in QT means exceeding one standard deviation (σ). For example, loss-of-function mutations in NSD1 causing Sotos syndrome increase mean occipitofrontal head circumference by $>2\sigma$ (27); mutations in JAG1 causing Alagille syndrome decrease mean height by 2σ (28); dominant-negative mutations in FBN1 producing Marfan syndrome reduce mean bone mineral density by 1.5σ (29); and LDL receptor mutations causing hypercholesterolemia increase mean LDL-bound cholesterol level by $>4\sigma$ (30). However, a segregating nonsynonymous SNP A390P in CETP is associated with 0.4σ decrease in plasma HDL-C levels (31). We therefore concentrated on values of 0.25σ and 0.5σ for new missense mutations but also considered a wide range of QT means shifts (*SI Appendix*, Fig. S5*A*). In the context of a trait such as human height, 0.25σ would correspond to a change in mean height of 0.5 inch.

Simulation of Resequencing of Individuals from Phenotypic Extremes. Next, we simulated results of resequencing of individuals with extreme QT values. Simulated individuals were sorted according to their QT values and gene sequence information was extracted for individuals from lower and upper percentiles. Although it is easy to incorporate sequencing errors in the model, in this work we assumed ideal quality of sequence data.

Simulation of Detection of QT Affecting Genes. Finally, we tested whether lower and upper tails of the trait distribution were significantly (at the genome-wide level) different in the total amount of rare nonsynonymous variants. Current candidate gene-based studies focus on variants observed exclusively in one of the tails. With increasing sample size, this approach will have decreased power because at would ignore variants observed predominantly, albeit not exclusively in one of the extremes (Fig. 5). We used two alternative approaches: one in which only rare variants were analyzed and the other in which information on common variants was also included to further improve power (32).

In the "rare variants only" approach SNPs with the observed frequency above $>5\%$ in either of the two tails were considered to be common and were excluded from further analysis. We then calculated the cumulative frequency of rare variants separately at each phenotypic extreme, by summing counts of rare SNPs. We calculated statistical significance of departure of rare variants distribution between two phenotypic tails from 50%/50% ratio using the χ^2 test. If the obtained P-value was <0.05 after Bonferroni correction to 20,000 genes, the association between the gene and the phenotype was counted as identified.

To incorporate information about common alleles we compared contingency table of allele counts between two phenotypic extremes considering all SNPs with minimal observed frequency above $>5\%$ independently and the remaining rare SNPs grouped into one category with all allele counts summed. Dissimilarity between distributions of allele counts in the two phenotypic extremes was detected by using χ^2 test.

Estimation of Number of Required Individuals and Power of Resequencing Studies. For each set of parameters the simulation described above was run 10,000 times. Power of the test was calculated as the fraction of simulations in which enrichment of rare variants in the proper phenotypic tail was detected. As shown in Table 1, achieving substantial power for individual gene discovery by using this study design would require large population samples. Sequencing of 10,000 individuals (with 5,000 from the upper and 5,000 from the lower fifth percentiles) would require 100,000 phenotyped individuals and would provide 77% and 40% power for genes with the effect of new missense mutations of 0.5σ and 0.25σ, respectively. The power estimates were 78% and 49% for the above parameters, under the model that assumed disassociation between a mutation's effect on phenotype and the strength of purifying selection acting on the mutation.

Incorporating frequent variants in addition to rare variants increases the power by 3%. It should be noted, however, that our analysis was focused on the utility of rare, deleterious SNPs in identification of genes affecting human traits. Our simulations do not incorporate balancing selection, selection reversal, antagonistic pleiotropy, or any other evolutionary forces that might lead to high-frequency functional variants. Correspondingly, the real gain of power from analysis of common variants in addition to pooled rare variants might be significantly higher that these estimates.

Sampling of suboptimal demographic histories and distributions of selection coefficients (*SI Appendix*, Fig. S6 and Table S1) suggests that our power estimates are within the range of 20%

Table 1. Estimated power of gene mapping by complete resequencing

Effect of functional mutations (in fractions of standard deviation)	No. of sequenced individuals	No. of phenotyped individuals				
		12,500	25,000	50,000	100,000	200,000
0.25σ	5,000	0.11	0.18	0.24		
	10,000		0.24	0.31	0.40	
	20,000			0.38	0.51	0.59
0.5σ	5,000	0.36	0.47	0.57		
	10,000		0.56	0.69	0.77	
	20,000			0.76	0.84	0.88

to 70%. Dependencies of the power estimates on other model parameters are shown in *SI Appendix*, Fig. S5.

A much smaller sample size of 1,000 individuals would have >75% power to detect effect sizes of 2σ. If pathways rather than individual genes would be considered as units of the association test, a sample of 1,000 would be sufficient to achieve over >60% power (assuming 20 genes per pathway and effect size of 0.5σ).

Although 100,000 individuals is substantially larger than any currently existing well-phenotyped clinical population, efforts to aggregate populations on this scale are underway and will be greatly facilitated by electronic medical records. Notably, the total number of unique individuals in published systematic retrospective clinical studies currently available for genetic analysis substantially exceeds 100,000. Genome-wide association studies have already been conducted on >15,000 individuals (2). Also, numerous trait genes may in fact be uncovered with far shallower levels of resequencing.

Meta-Analysis of MC4R Gene Resequencing Studies. To illustrate the feasibility of this approach we performed a meta-analysis of four resequencing datasets of the MC4R gene, mutations that are known to be strongly overrepresented among severely obese individuals (8, 33–35) (*SI Appendix*, Table S2). Forty-three individuals with rare missense variants were found among 2,940 obese individuals (99th BMI percentile), while only 3 individuals with rare missense variants were detected in 1,925 lean individuals (<15th BMI percentile). With the described approach, MC4R would be readily detected with P-value $< 5 \cdot 10^{-7}$, achieving genome-wide statistical significance. As such, this gene could have been discovered by resequencing a sample of <5,000 individuals.

Discussion

Importantly, our simulations do not assume that protein-coding variation is solely responsible for phenotypic variation. In contrast to previous studies (36–38) the goal of these simulations was not to predict the allelic spectrum of human disease, but rather to probe the utility of rare missense SNPs in discovery of genes influencing human traits. The considered study design is similar in concept to a genetic screen utilizing naturally occurring mutations. Even if a large fraction of trait variance were explained by noncoding *cis*-regulatory variation (e.g., causing changes in expression levels of genes associated with phenotypes), hypomorphic coding mutations arising in the same genes would lead to some level of functional polymorphism detectable by resequencing of these genes in a sufficiently large population. Accordingly, resequencing of only coding regions should be sufficient to identify genes affecting traits of interest, but to explain population variation in phenotype, resequencing of noncoding region might be necessary. We note, however, that extending the very same type of analysis to highly conserved non-coding regions is unlikely to be successful because the effect of mutations in these regions is probably much weaker than the effect of coding mutations. Deletions of ultraconserved regions were not shown to lead to any detectable phenotypes (39) and population genetics studies suggest that conservation of many of the conserved regions is maintained by weak selective pressure (40–42).

We conclude that a simple study design combining the pending availability of both low-cost–high-volume sequencing technology and large well-phenotyped population samples has the potential to facilitate discovery of genes relevant to human phenotypes.

Materials and Methods

Demographic Model. Overall, the demographic model included four parameters: (*i*) long-term ancestral effective population size; (*ii*) bottleneck population size; (*iii*) duration of exponential growth in generations; and (*iv*) recent effective population size. In the analysis, experimental site frequency spectrum was represented by synonymous and noncoding variants combined. Allelic spectra of these categories of SNPs were similar, and pooling them together resulted in an increase of the amount of data. Although we used data on 757 individuals sequenced (1,514 chromosomes), due to failed base calls in some cases, we approximated number of sequenced chromosomes as 1,400. The combined number of sequenced neutral nucleotide sites was approximated as 63,000 (see *SI Appendix*, *Methods*).

Neutral Wright–Fisher Model for Variable Population Size. Let $\phi(q, t|p)$ be the probability density that the allele frequency in the tth generation is between x and $x + dx$, $(0 < x < 1)$ given its starting frequency p. It has been shown (43–47) that $\phi(q, t|p)$ satisfies the forward *Kolmogorov* equation:

$$\frac{\partial \phi}{\partial t} = \frac{1}{4N_t} \cdot \frac{\partial^2}{\partial q^2}\{q(1-q)\phi\}.$$

For a constant population size, $N_t = N_0$, the transient solution:

$$\phi(q, t|p, N_0) = \sum_{i=1}^{\infty} \frac{(2i+1)(1-(1-2p)^2)}{i(i+1)} \cdot C_{i-1}^{3/2}(1-2p) \cdot C_{i-1}^{3/2}(1-2q) \cdot e^{-\frac{i(i+1)}{4N_0}t},$$

where $C_i^{3/2}$ is the *Gegenbauer* polynomial with $\lambda = 3/2$. For general time-dependent population size N_t, we can obtain $\phi(q, t|p)$ by defining *effective time* t', such that $dt' = (N_0/N_t)dt$. Written in terms of the effective time t', forward *Kolmogorov* equation becomes

$$\frac{\partial \phi}{\partial t'} = \frac{1}{4N_0} \cdot \frac{\partial^2}{\partial q^2}\{q(1-q)\phi\}.$$

Consequently, the solution at time τ is

$$\phi(q, \tau' \,|\, p, N_0) = \phi\left(q, \left[\int_0^{\tau} \frac{N_0}{N_t} dt\right] | \, p, N_0\right).$$

For $N_t = N_0 \cdot e^{\gamma t}$,

$$\tau' = \int_0^{\tau} e^{-\gamma t} dt = \frac{1 - e^{-\gamma \tau}}{\gamma}.$$

Site-Frequency Spectrum at Present. The expected site frequency spectrum $f(q)$, produced by the above demographic model in today's population, is given by a sum of the contribution of ancestral mutations originated at the stationary phase prior to the bottleneck and the contribution of recent mutations occurring in the population with the cumulative rate $2N_t\mu$ per generation t.

$$f(q) = 4N_1\mu \sum_{i=1}^{2N_b-1} \frac{1}{i} \cdot \phi\left(q, \frac{1-e^{-\gamma\tau}}{\gamma} \,\middle|\, \frac{i}{2N_b}, N_b\right) + 2N_b\mu \cdot \sum_{t=1}^{\tau} e^{\gamma t} \phi\left(q, \frac{1-e^{-(\tau-t)\gamma}}{\gamma} \,\middle|\, \frac{1}{2N_b e^{\gamma t}}, N_b e^{\gamma t}\right).$$

GENETICS

Likelihood. In a sample of N_s chromosomes the number of segregating sites observed with polymorphic multiplicity i is given by

$$F_i = \int_0^1 \binom{N_s}{i} q^i (1-q)^{n-i} f(q) dq.$$

In absence of additional information about which allele is the ancestor, what is actually recorded is the minority multiplicity, i.e., $F_j^{(b)} = 63{,}000(F_j + F_{N_s-j})$, for $j = 1, \ldots, N_s/2 - 1$. The number of monomorphic sites is given by $F_0 = 63{,}000 - \sum_{j=1}^{N/2} F_j^{(b)}$. The probability that SNP will be found i times of N_s is then $P_j = F_j^{(b)}/63{,}000$.

Assuming independence between SNPs, the likelihood is given by

$$L(N_1, N_b, \gamma, \tau) = \prod_{j=0}^{Ns} P_j^{M_j}.$$

Note that because the likelihood includes monomorphic sites, maximizing it maximizes match in the observed units and not merely the distribution form. The 4D likelihood surface was studied extensively and shown to possess a single maximum. Its 2D projections are shown in Fig. 2.

Estimation of Selection Coefficient Distributions. Distribution of selection coefficients was modeled as a gamma distribution with addition of completely neutral sites (selection coefficient, s = 0) and sites under "lethally" strong selection (s = 1). Intermediate categories of sites with selection coefficients in the range of $10^{-5} - 10^{-1}$ were modeled by a gamma distribution:

$$f(s, k, \theta) = s^{k-1} \frac{e^{-\frac{s}{\theta}}}{\theta^k \Gamma(k)},$$

where $\Gamma(k)$ is a gamma function. Continuous gamma distribution was transformed to a 20-bin histogram by numerical integration within corresponding bin boundaries uniformly spaced in the logarithmic scale between 10^{-5} and 10^{-1}. We determined parameters of the distribution by using an exhaustive grid search. Fraction of neutral and invariable sites were allocated in 5% minimal "blocks," while remaining 20 bins described by gamma distribution were normalized so the sum of all bins was equal 1. All possible values from 0% to 100% were tested for values of extreme bins. Values of θ and k parameters were taken from the logarithmic grid: θ from 10^{-5} to 10^{-1} with $\sqrt{10}$ step; k from 10^{-2} to 10 with $\sqrt{10}$ step.

Allele frequency spectra were computed by using forward simulations with a predefined demographic history and a chosen distribution of selection coefficients. Log-likelihood LL, for each model was computed as:

$$LL = \sum_{i=1}^{M/2} n_i^{\text{exp}} \ln\left(n_i^{\text{simul}}/n^{\text{total}}\right),$$

where n_i^{exp} is a number of nonsynonymous SNPs in which minor allele was observed i times in the experimental resequencing dataset, n_i^{simul} is a number of SNPs with simulated minor allele count equal to i, n^{total} is the total number of simulated "missense" sites (20,000), and M is the number of sequenced chromosomes.

ACKNOWLEDGMENTS. We thank N. Ahituv and L. Pennacchio for sharing resequencing data, and A. Kondrashov, I. Kohane, and G. Church for helpful discussions. This work was supported by National Institutes of Health Grants GM078598 and MH084676 (to S.R.S.), R01GM071852 (to J.A.S.) and National Institutes of Health roadmap initiative grant U54LM008748.

1. Couzin J, Kaiser J (2007) Genome-wide association. Closing the net on common disease genes. *Science* 316:820–822.
2. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–678.
3. Frayling TM (2007) Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat Rev Genet* 8:657–662.
4. Cohen JC, *et al.* (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305:869–872.
5. Kotowski IK, *et al.* (2006) A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am J Hum Genet* 78:410–422.
6. Romeo S, *et al.* (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat Genet* 39:513–516.
7. Riley BM, *et al.* (2007) Impaired FGF signaling contributes to cleft lip and palate. *Proc Natl Acad Sci USA* 104:4512–4517.
8. Ahituv N, *et al.* (2007) Medical sequencing at the extremes of human body mass. *Am J Hum Genet* 80:779–791.
9. Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutat Res* 615, 28–56.
10. Hirschhorn JN, Altshuler D (2002) Once and again-issues surrounding replication in genetic association studies. *J Clin Endocrinol Metab* 87:4438–4441.
11. Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. *Nat Rev Genet* 8:610–618.
12. Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *Am J Hum Genet* 80:727–739.
13. Yampolsky LY, Kondrashov FA, Kondrashov AS (2005) Distribution of the strength of selection against amino acid replacements in human proteins. *Hum Mol Genet* 14:3191–3201.
14. Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173:891–900.
15. Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246:96–98.
16. Kondrashov AS (2003) Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* 21:12–27.
17. Adams AM, Hudson RR (2004) Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* 168:1699–1712.
18. Evans SN, Shvets Y, Slatkin M (2007) Non-equilibrium theory of the allele frequency spectrum. *Theor Popul Biol* 71:109–119.
19. Griffiths RC (2003) The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theor Popul Biol* 64:241–251.
20. Ammerman AJ, Cavalli-Sforza LL (1984) *The Neolithic Transition and the Genetics of Populations in Europe* (Princeton Univ Press, Princeton, NJ).
21. Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166:351–372.
22. Williamson SH, *et al.* (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci USA* 102:7882–7887.
23. Schaffner SF, *et al.* (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15:1576–1583.
24. Voight BF, *et al.* (2005) Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci USA* 102:18508–18513.
25. Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* 39:1251–1255.
26. Loewe L, Charlesworth B (2006) Inferring the distribution of mutational effects on fitness in Drosophila. *Biol Lett* 2:426–430.
27. Cole TR, Hughes HE (1994) Sotos syndrome: A study of the diagnostic criteria and natural history. *J Med Genet* 31:20–32.
28. Rovner AJ, *et al.* (2002) Rethinking growth failure in Alagille syndrome: The role of dietary intake and steatorrhea. *J Pediatr Gastroenterol Nutr* 35:495–502.
29. Moura B, *et al.* (2006) Bone mineral density in Marfan syndrome. A large case-control study. *Joint Bone Spine* 73:733–735.
30. Brown MS, Goldstein JL (1986) A receptor-mediated pathway for cholesterol homeostasis. *Science* 232:34–47.
31. Spirin V, *et al.* (2007) Common single-nucleotide polymorphisms act in concert to affect plasma levels of high-density lipoprotein cholesterol. *Am J Hum Genet* 81:1298–1303.
32. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am J Hum Genet* 83:311–321.
33. Hinney A, *et al.* (2006) Prevalence, spectrum, and functional characterization of melanocortin-4 receptor gene mutations in a representative population-based sample and obese adults from Germany. *J Clin Endocrinol Metab* 91:1761–1769.
34. Larsen LH, *et al.* (2005) Prevalence of mutations and functional analyses of melanocortin 4 receptor variants identified among 750 men with juvenile-onset obesity. *J Clin Endocrinol Metab* 90:219–224.
35. Hinney A, *et al.* (2003) Melanocortin-4 receptor gene: Case-control study and transmission disequilibrium test confirm that functionally relevant mutations are compatible with a major gene effect for extreme obesity. *J Clin Endocrinol Metab* 88:4258–4267.
36. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 69:124–137.
37. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. *Trends Genet* 17:502–510.
38. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: Common disease-common variant.or not? *Hum Mol Genet* 11:2417–2423.
39. Ahituv N, *et al.* (2007) Deletion of ultraconserved elements yields viable mice. *PLoS Biol* 5:e234.
40. Kryukov GV, Schmidt S, Sunyaev S (2005) Small fitness effect of mutations in highly conserved non-coding regions. *Hum Mol Genet* 14:2221–2229.
41. Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ (2005) Evolutionary constraints in conserved nongenic sequences of mammals. *Genome Res* 15:1373–1378.
42. Chen CT, Wang JC, Cohen BA (2007) The strength of selection on ultraconserved elements in the human genome. *Am J Hum Genet* 80:692–704.
43. Ewens W (2004) *Mathematical Population Genetics I. Theoretical Introduction* (Springer, Heidelberg).
44. Kimura M (1955) Solution of a process of random genetic drift with a continuous model. *Proc Natl Acad Sci USA* 41:144–150.
45. Kimura M (1964) Diffusion models in population genetics. *J Appl Prob* 1:177–232.
46. Joyce P, Tavaré S (1995) The distribution of rare alleles. *J Math Biol* 33:602–618.
47. Myers S, Fefferman C, Patterson N (2008) Can one learn history from the allelic spectrum? *Theor Popul Biol* 73:342–348.