Searching for Missing Heritability - Designing Rare Variants Association Studies: Supplementary Information

Contents

List of Figures

List of Tables

List of Symbols

L	Target size (number of sites, in nucleotides)	7
$\mu_{C,nc}$	Mutation rate (per-nucleotide per-generation)	7
μ_C	Mutation rate (per-class per-generation)	7
s	Selection coefficient	7
K	Total number of generations since ancestral population in a demographic model	8
N	Effective population Size	8
S	Re-scaled selection coefficient ($S = 4Ns$)	8
f	Derived allele frequency	8
t_s	Mean time until fixation/extinction for an allele with selection coefficient s	8
ψ_s	Density distribution of allele frequency for selection coefficient s	8
Ψ_s	Cumulative distribution of allele frequency for selection coefficient s	9
A_s	Probability distribution of age of alleles	9
γ	Euler's constant, $\gamma \sim 0.5772$	10
Ei	Exponential integral function, $Ei(x) \equiv \int_{-\infty}^{x} \frac{e^t}{t} dt$	10
λ	Effect size: relative in disease risk conferred by an allele	17
α	Fraction of null missense alleles at births	17
ρ_s	Fraction of null missense alleles for alleles below a certain frequency	24
π	Disease prevalence	29
n	Number of samples in an RVAS study	29
d	Number of observed derived alleles in an RVAS study	29
$\mathcal{LL}$	Log-likelihood	30
$\mathcal{LLR}$	Log-likelihood ratio	30
NCP	Non-Centrality-Parameter	30
V	Heritability: Total variance explained by a set of alleles	85

Software

Matlab code performing most of the mathematical calculations and simulations used in the paper is available as part of "heritability calculator" package `http://www.broadinstitute.org/mpg/hc`. A separate software package for simulations of demographic model was written by Stephen Schaffner (`sfs@broadinstitute.org`).

1 Allele Frequency Distribution for Different Populations: Theory

In this section we show how to compute allele frequency distribution of a class C of alleles, having a known selection coefficient s. Throughout the paper, we assume that the dynamics of allele frequencies throughout human demographic history follow a generalized Wright-Fisher model [1, 2].

1.1 Generalized Wright-Fisher Model

We assume a generalized Wright-Fisher model incorporating mutation, selection, and changes in population size throughout history; see for example [3]. (Other aspects such as migration and admixture are ignored for simplicity). We consider a target region, which we can imagine to be the coding region of a gene; however, the analysis could apply equally to other targets (such as a set of coding regions, or an extended genomic region). Within the region, each nucleotide (site) is assumed to have two possible alleles: the ancestral allele and the derived allele.

1.1.1 Model Parameters

We consider a class C of alleles that are functionally and selectively equivalent. (for example, disruptive alleles; silent alleles; or missense alleles that abolish gene function). Our model has the following parameters,

1. Target size: L denotes the length of the target (measured in nucleotides).
2. Mutation rate: $\mu_{C,nc}$ denotes the rate per-generation per-nucleotide at which new derived alleles in class C arise. μ_C denotes the rate for the entire target (for example, all mutation in a certain class in a gene), given by $\mu_C = L\mu_{C,nc}$.
3. Selection coefficient: s denotes the selection coefficient for derived alleles in C. The selection coefficient s is defined as the relative reduction in the expected number of offspring for carriers of the allele compared to non-carriers. An allele with frequency f_t in one generation has expected frequency $f_{t+1} = \frac{(1-s)f}{1-fs}$ in the next generation. Given that the individual allele frequencies under consideration are tiny, $f_{t+1} \approx (1-s)f_t$. (Simulations (see Section 1.3.2) performed with the exact or approximate formula yield essentially indistinguishable results).

We assume an additive model for fitness in which selection on an allele does not depend on the genetic background (that is, the value of other alleles).

4. Population size: $N^{(k)}$ denotes the population size at generation k, where k ranges from 1 to the total number of generations, K, under consideration. For the special case of population at equilibrium, we denote the effective population size by N or N_{eq}.

5. Re-scaled selection coefficient: $S = 4Ns$ denotes the selection coefficient re-scaled by the population size, where N is the population size at a given time.

1.1.2 Stochastic Process Formulation

We next formulate the generalized Wright-Fisher model as a discrete-time discrete-space stochastic process. We can formulate the process in terms of allele frequencies: we denote the derived allele frequency of a polymorphic allele in the population by f. More specifically, we define $f_i^{(k)}$ the allele frequency of the i-th allele at generation k ($i = 1, .., L$). Similarly, we denote $X_i^{(k)}$ the number of carriers of the i-th derived allele at generation k. Both $f_i^{(k)}$ and $X_i^{(k)}$ are random variables, whose distributions are given by the following generative process,

1. Initialize $X_i^{(0)} = f_i^{(0)} = 0$ for $i = 1, .., L$ - all sites are assumed to carry the ancestral allele.

2. For $k = 1$ to K,

   (a) Draw $X_i^{(k+1)}$ according to a binomial distribution,

   $$X_i^{(k+1)} \sim Binomial\Big(2N^{(k+1)}, f_i^{(k)} \frac{1 - s_i}{1 - f s_i}\Big). \tag{1.1}$$

   (b) Set $f_i^{(k+1)} = \frac{X_i^{(k+1)}}{2N^{(k+1)}} + \mu_{C,nc}$ to account for mutations.

We are interested in several properties of the allele frequency distribution as a function of the parameters of the model defined above.

1. Time spent at each frequency: Each newly born allele spends a certain amount of time (number of generations) in a polymorphic state at each given frequency $0 < f < 1$ until it is either fixed or becomes extinct. We denote by t_s(f) the *density* of the mean time the allele spends in a polymorphic state at frequency f. That is, the number of generations spent between frequencies f and Δf is $\int_f^{f+\Delta f} t_s(f')df'$.

2. Total time spent at polymorphic state: Each newly born allele spends a certain amount of time (number of generations) in a polymorphic state until it is either fixed or becomes extinct. We denote by t_s the mean time the allele spends in a polymorphic state.

3. Probability density function: We denote the allele frequency probability density function by ψ_s, where we sample alleles with probability proportional to their derived allele frequencies f_i. That is, we consider the distribution of allele frequencies obtained when sampling a

random individual from the population, and then sampling a random derived allele from this individual, and denote $\psi_s(f)$ the density obtained for this distribution.

4. Cumulative distribution function: We denote the cumulative probability function for allele frequencies by Ψ_s. Here we sample alleles with probability proportional to their derived allele frequencies f_i as above.

5. Age distribution of alleles: We denote by $A_s(k)$ the probability that a randomly sampled alleles on a randomly sampled chromosome from the population is of age k, i.e. was born k generations ago.

The model parameters μ_C, s and $N^{(k)}$ determine all quantities of interest above.

For population at equilibrium, Kimura solved the Wright-Fisher process analytically in the limit of large population size N, by using a diffusion approximation; see for example [3–6]. In Section 1.2, we use these solutions to give analytic formulas for the above quantities for a population at equilibrium. For general demographies, there is no known closed form solution, although some partial results and special cases are known (for example [7–9]), and one needs to resort to numerical calculations. In Section 1.3, we compute the quantities above by using simulations for various demographies.

1.2 Allele Frequency Distribution at Equilibrium: Analytical Formulas

In this section we compute parameters of interest for a *single* allele using a diffusion approximation for large N. We assume throughout the generalized Wright-Fisher model from eqs. (2). Kimura and Crow [6] expressed the *expected time* a newly born allele spends at allele frequency f as,

$$t_s(f) = \frac{2[1 - e^{-S(1-f)}]}{f(1-f)(1-e^{-S})}. \tag{1.2}$$

This is the density distribution of expected number of generations spent at frequency f (such that the number of generations spent between frequency f and $f + \Delta f$ is the integral of $t_s(f)$ in this range). From $t_s(f)$ we can compute all parameters of interest for the model using standard integrations and asymptotic expansions.

For example, we define the mean total *weighted* frequency T_s that a newly born allele accumulates throughout its life in a polymorphic state, until it is either lost or fixed in the population. The units of T_s are (frequency $\times$ generations). For example, for an allele which survive on average three generation as a singleton and two generation as a doubleton on average we will have $T_s = 3 \times \frac{1}{2N} + 2 \times \frac{2}{2N} = \frac{7}{2N}$. We have the following for the mean total weighted frequency T_s,

Proposition 1 *The total weighted frequency a newly born allele accumulates throughout its life time, until it dies, is given by, (in units of frequency $\times$ generations) is*

$$T_s \equiv \int_0^1 f t_s(f) df = \frac{2[\log(-S) + \gamma - Ei(-S)]}{1 - e^{-S}} \tag{1.3}$$

where γ is Euler's constant ($\gamma \sim 0.5772$), and Ei is the exponential integral function, defined as,

$$Ei(f) \equiv \int_{-\infty}^{f} \frac{e^x}{x} dx. \tag{1.4}$$

Proof 1 *Using the definition of T_s, we integrate to get*

$$\begin{aligned} T_s &= \int_0^1 f t_s(f) df \\ &= \frac{2}{1 - e^{-S}} \int_0^1 \frac{1 - e^{-S(1-f)}}{(1-f)} df \\ &= \frac{2}{1 - e^{-S}} \Big| - \log(1-f) + e^S Ei(S(f-1)) \Big|_{f=0}^{1} \\ &= \frac{2}{1 - e^{-S}} \Big[\lim_{x \to 0} \big(- \log(x) + e^S Ei(S(-x)) \big) - e^S Ei(-S) \Big] \\ &= \frac{2[\log(-S) + \gamma - Ei(-S)]}{1 - e^{-S}}. \end{aligned} \tag{1.5}$$

where here and throughout, we use the following asymptotic expansions for $Ei(x)$ in different regimes,

$$\begin{aligned} Ei(x) &\sim \log(x) + \gamma + x, \quad x \to 0 \\ Ei(x) &\sim \frac{e^x}{x}(1 + \frac{1}{x} + \frac{2}{x^2}), \quad x \to \pm\infty \end{aligned} \tag{1.6}$$

■

Proposition 2 *For a class C of alleles with total mutation rate μ_C, the Combined Allele Frequency (CAF) at equilibrium is*

$$f_C = 2N\mu_C T_s = \frac{4N\mu_C[\log(-S) - Ei(-S) + \gamma]}{1 - e^{-S}}. \tag{1.7}$$

In addition, we have the following asymptotic approximations for weak selection (nearly-neutral alleles, $S \to 0$) and very strong selection ($S \to \infty$),

$$f_C \approx \begin{cases} 4N\mu_C & S \ll 1 \\ \frac{\mu_C}{s} & S \gg 1 \end{cases} \tag{1.8}$$

Proof 2 *The combined allele frequency is given by multiplying the population mutation rate per generation, $2N\mu_C$, by the total time spent in polymorphic state for a given allele, T_S, giving $f_C = [2N\mu_C]T_S$. Plugging into eq. (1.3) gives the formula for f_C.*

For the asymptotic relations, we have,

$$\begin{aligned}\lim_{S\to 0} f_C &= 4N\mu_C \lim_{S\to 0} \frac{\log(-S)+\gamma-Ei(-S)}{1-e^{-S}} \\ &= 4N\mu_C \lim_{S\to 0} \frac{\log(-S)+\gamma-[\log(-S)+\gamma-S]}{S} \\ &= 4N\mu_C. \end{aligned} \tag{1.9}$$

$$\begin{aligned}\lim_{S\to \infty} f_C &= 4N\mu_C \lim_{S\to \infty} \frac{\log(-S)+\gamma-Ei(-S)}{1-e^{-S}} \\ &= 4N\mu_C \lim_{S\to \infty} \frac{-e^{-S}/S}{-e^{-S}} = \frac{4N\mu_C}{4Ns} \\ &= \frac{\mu_C}{s}. \end{aligned} \tag{1.10}$$

■

For strong selection, the above relation $f_C \approx \frac{\mu_C}{s}$ can also be obtained by considering mutation-selection balance, as observed early on by Fisher [1] and Haldane [10].

1.2.1 Individual Allele Frequency Distribution

We compute here the distribution of individual allele frequencies at equilibrium - recall that we use *weighted* sampling, where individual alleles are sampled with *weight* proportional to their frequency (alternatively, this is the distribution obtained when sampling a random chromosome in the population, sampling a random derived allele from this chromosome, and observing the frequency of this allele). The individual allele frequency distribution is given by the following proposition (the proof is given by simple integration),

Proposition 3 *For a population at equilibrium, at the large N limit, we have,*

1. *The probability density function is*

$$\psi_s(f) = \frac{f t_s(f)}{\int_0^1 x t_s(x) dx} = \frac{1-e^{-S(1-f)}}{(1-f)[\gamma - Ei(-S) + \log(-S)]}. \tag{1.11}$$

2. *The cumulative distribution function is,*

$$\Psi_s(f) = \int_0^f \psi_s(x)dx = \frac{Ei\Big(S(f-1)\Big) - \log(1-f) - Ei(-S) + \log(-S)}{\gamma + \log(-S) - Ei(-S)}. \tag{1.12}$$

Proof 3 *1. By definition, and plugging in eqs. (1.2, 1.3),*

$$\begin{aligned}\psi_s(f) &= \frac{f t_s(f)}{\int_0^1 x t_s(x) dx} \\ &= \frac{1-e^{-S(1-f)}}{(1-f)(1-e^{-S})} \Big/ \frac{\log(-S)+\gamma-Ei(-S)}{1-e^{-S}} \\ &= \frac{1-e^{-S(1-f)}}{(1-f)[\gamma-Ei(-S)+\log(-S)]}. \end{aligned} \tag{1.13}$$

2. Integrating, we get

$$\begin{aligned}\Psi_s(f) &= \int_0^f \psi_s(x) dx \\ &= \frac{1}{\gamma-Ei(-S)+\log(-S)} \int_0^f \frac{1-e^{-S(1-x)}}{1-x} \\ &= \frac{1}{\gamma-Ei(-S)+\log(-S)} \Big[-\log(1-x) + e^S Ei(S(x-1)) \Big]_{x=0}^{f} \\ &= \frac{Ei(S(f-1)) - \log(1-f) - Ei(-S) + \log(-S)}{\gamma+\log(-S)-Ei(-S)}. \end{aligned} \tag{1.14}$$

■

1.3 Allele Frequency Distribution for Different Populations: Simulations

The human population is not at equilibrium, but has changed in size. Recent human evolution has seen an exponential explosion in population size [11]. In addition, migration events, such as the out-of-Africa event [12] or the peopling of isolated regions such as Iceland [13] and Finland [14], have created population bottlenecks. We study several simple demographic models that capture key features of population expansion and bottlenecks. For each model, we used simulations to compute the resulting allele frequency distribution and the parameters of interest.

1.3.1 Description of Demographic Models

We considered six models:

1. **Equilibrium:** The ancestral human population, assumed to have constant population size.
2. **Expansion1:** simple exponential expansion
3. **Expansion2:** two-phase exponential expansion, incorporating a recent period of more rapid growth (as suggested by recent human genetic data).
4. **Europe:** a model of the peopling of Europe, with a bottleneck roughly 1280 generations ago.

5. **Finland:** a model of the peopling of Finland, with a severe bottleneck roughly 100 generations ago
6. **Iceland:** a model of the peopling of Finland, with a severe bottleneck roughly 50 generations ago

All models have an exponential expansion stage, which corresponds to recent expansion in human populations. Some of the models contain bottleneck stages.

The parameters describing the models are as follows,

1. ϵ_1 - the relative increase in population size, per generation, first stage.
2. k_1 - the number of generations in the first stage.
3. $N_e(B)$ - the population size (number of individuals) surviving the first bottleneck.
4. ϵ_2 - the relative increase in population size, per generation, second stage.
5. k_2 - number of generation in second stage, after the population bottleneck.

The parameters used for different models are shown in Table S1.

Model	ϵ_1	k_1	$N_e(B)$	ϵ_2	k_2	$N_e(final)$
Equilibrium	0.0%	0	-	0.0%	0	$10,000$
Expnasion1	0.462%	1000	-	0.0%	0	$1,000,000$
Expansion2	0.197%	1230	-	9.4%	50	$10,000,000$
Europe$^{(*)}$	0.396%	1230	775	9.4%	50	$9,000,000$
Finland	0.462%	900	50	12.2%	100	$5,000,000$
Iceland	0.462%	950	500	13.2%	50	$250,000$

Table 1: Parameters for different demographic models. We also display the resulting final population size. (*) For Europe, we assume an *ancient* bottleneck, before the two expansion periods. For Iceland and Finland we assume a *recent* bottleneck, between the two expansion periods. For Europe we stopped the simulation ~ 50 generations ago at a population of ~ 9 millions for computational reasons. The last few generations which have a huge population size and are expected to produce only alleles with tiny frequencies. Since our results all depend on the *weighted* allele frequency distributions, ignoring these generations should not alter our results significantly.

1.3.2 Description of Simulations

We initialized the population size of all models to $N_{eq} = 10,000$, and run at this constant population size for a "burn-in period" of $k_B = 20N_{eq} = 200,000$ simulations, to reach equilibrium. The mutation rate (per-chromosome per-generation) was set to $\mu_D = 1.7 \times 10^{-6}$. We varied

the selection coefficient on a logarithmic scale with increments of 0.1 in the exponent, getting $s = 10^{-1}, 10^{-1.1}, 10^{-1.2}, ..., 10^{-4.9}, 10^{-5}$, and in addition we used $s = 0$ representing neutral alleles.

For each selection coefficient and population, we performed $I = 50,000$ independent simulations (representing histories for $50,000$ equivalent genes).

The simulations roughly followed the stochastic process described in Section 1.1.2. Specifically, we describe the Monte-Carlo simulation algorithm we used in Box 1. The values we used for all other parameters not described above are given in Table S1.

Box 1: Algorithm for Simulating Allele Frequency Distribution

Input Parameters:

N_{eq} - Initial population size at equilibrium

s - selection coefficient

μ_C - mutation rate

$\epsilon_1, \epsilon_2, k_1, k_2, N_e(B)$ - demographic parameters (see Table S1)

k_B - 'burn-in' parameter (number of generations run at ancestral population at equilibrium)

I - number of iterations to simulate

Output:

$\hat{f_C}$ - Mean Combined allele frequency

$\hat{\Psi}_s(f)$ - Cumulative individual allele frequency distribution

$\hat{\psi}_s(f)$ - Density of individual allele frequency distribution

$\hat{A}(k)$ - Age probability distribution

Steps:

1. Repeat I times, for $i = 1$ to I,

   (a) Initialize n_i, the counter of the number of derived alleles: set $n_i = 0$.

   (b) Set $K = k_B + k_1 + k_2$, the total number of generations to simulate.

   (c) For $k = 1$ to K,

      i. Set population size for next generation,

$$N^{(k+1)} = \begin{cases} N & k < k_B \\ N_e(B) & k = k_B \quad (^{(*)} \text{ only for Europe}) \\ \lfloor N^{(k)}(1+\epsilon_1) \rfloor & k_B < k < k_B + k_1 \\ N_e(B) & k = k_B + k_1 \quad \text{(only for Iceland and Finland)} \\ \lfloor N^{(k)}(1+\epsilon_2) \rfloor & k_B + k_1 < k \end{cases} \tag{1.15}$$

      ii. For $j = 1$ to n_i (all existing derived alleles)

         A. Draw $X_{ij}^{(k+1)}$ according to a binomial distribution,

$$X_{ij}^{(k+1)} \sim Binomial\Big(2N^{(k+1)}, f_{ij}(1-s_i)\Big)$$

         B. Update the frequency: set $f_{ij} = \frac{X_{ij}^{(k+1)}}{2N^{(k+1)}}$

      iii. A. Draw $n_{new} \sim Binomial(2N^{(k+1)}, \mu_C)$, the number of newly born alleles.

B. Set singleton frequency for newly born alleles $f_{ij} = \frac{1}{2N^{(k+1)}}$ for $j = n_i + 1, .., n_i + n_{new}$.

C. Record the generation at which each new allele is born, $a_{ij} = k$.

D. Update number of polymorphic alleles: set $n_i = n_i + n_{new}$.

(d) Record final individual allele frequencies, f_{ij} for all $i = 1, .., I$ and $j = 1, .., n_i$.

2. Compute empirical summary statistics from final outputs:

(a) Combined allele frequency

$$\hat{f_C} = \frac{1}{I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} f_{ij} \tag{1.16}$$

(b) Cumulative individual allele frequency

$$\hat{\Psi}_s(f) = \frac{1}{I\hat{f_C}} \sum_{i=1}^{I} \sum_{j=1}^{n_i} 1_{\{f_{ij} \leq f\}} f_{ij} \tag{1.17}$$

(c) Individual allele frequency density: Split the unit interval $[0, 1]$ into bins $B_0 = 0 < B_1 < .. < B_M = 1$. Then, for $f \in [B_i, B_{i+1})$, we have,

$$\hat{\psi}_s(f) = \frac{1}{I\hat{f_C}[B_{i+1} - B_i]} \sum_{i=1}^{I} \sum_{j=1}^{n_i} 1_{\{B_i \leq f_{ij} < B_{i+1}\}} f_{ij} \tag{1.18}$$

(d) Age distribution: ($\hat{A}_s(k)$ is the estimated probability of an allele to be born k generations before the end of the simulation)

$$\hat{A}_s(k) = \frac{1}{I\hat{f_C}} \sum_{i=1}^{I} \sum_{j=1}^{n_i} 1_{\{a_{ij} = K-k\}} f_{ij} \tag{1.19}$$

2 Allele Frequency Distribution for Different Populations: Empirical Results

We now turn from theory to simulated empirical results.

In section 1 we described the theory needed for computing the properties of a single allele in the population. In practice, we want to use a heterogenous class of different alleles (namely, missense alleles) having different selection coefficients, different effects on gene function and phenotype, etc. To handle this heterogeneity, we propose a simple two-class model, described in the next section. We then show results for allele frequency distribution under this model and for the different populations describe in Section 1.3.

2.1 Two-Class Model

As described in the main text, we adopt a simple two-class model in which all alleles in a gene are either neutral or null. Consider a gene G and a disease D. Given an allele in G, let s_{allele} denote the selection coefficient for the allele and let $\lambda_{alllele}$ denote the excess relative risk for D conferred by the allele.

1. Neutral alleles have no effect on selection or on the disease under study - that is $s_{allele} = \lambda_{allele} = 0$.
2. Null alleles abolish gene function. We have $s_{allele} = s$ and $\lambda_{alllele} = \lambda$, where s and λ are constant for each gene.

Alleles in the coding region of a gene can be divided into three observable classes.

1. Silent alleles, which do not affect the amino-acid sequence. We assume that all silent alleles are neutral.
2. Disruptive alleles, which produce nonsense, frameshift, and splice-site changes. We assume that all disruptive alleles are null.
3. Missense alleles follow a two-class model. Each newly born missense allele is assumed to be either null (with probability α) or neutral (with probability $(1 - \alpha)$).

In the main text, we primarily assume that $\alpha = 25\%$, which is the average value across genes [15]. However, the value of α appears to vary across genes. By comparing the deficit of missense alleles relative to silent alleles in the NHLBI Exome Sequencing Project (`http://evs.gs.washington.edu/EVS`), one can estimate values of $\alpha = 5\%$ for *BRCA2* and $\alpha = 50\%$ for *CHD8*.

2.2 Combined Allele Frequencies

We computed the (expected) combined allele frequencies for the different populations, and for different selection coefficients, as described in Section 1.3 and shown in Figure 1(a) in the main text, reproduced here in Figure 1.

In Table S2 we give the computed values for selected values of s, showing that the CAF is very similar for different populations, as previously observed [16].

s	Equilibrium	Expansion1	Expansion2	Europe	Finland	Iceland
0	0.0756	0.076	0.0762	0.0712	0.0708	0.0764
10^{-5}	0.0681	0.0695	0.0683	0.0638	0.0625	0.0691
$10^{-4.5}$	0.0533	0.0542	0.0542	0.0506	0.0511	0.0539
10^{-4}	0.0236	0.0241	0.0234	0.0231	0.0221	0.0238
$10^{-3.5}$	0.00561	0.00535	0.00516	0.00551	0.00497	0.00521
10^{-3}	0.00148	0.00145	0.0014	0.00144	0.00137	0.00137
$10^{-2.5}$	0.000485	0.000467	0.000446	0.000412	0.000469	0.000416
10^{-2}	0.000111	0.000162	0.000193	0.000162	0.00017	0.000137
$10^{-1.5}$	6.4e-005	8.24e-005	6.25e-005	7.37e-005	8.12e-005	8.01e-005
10^{-1}	2.23e-005	1.77e-005	1.82e-005	1.71e-005	2.06e-005	1.78e-005

Table 2: Combined Allele Frequency for different populations. We assume a mutation rate of $\mu_c = 1.7 \times 10^{-6}$. The values are nearly identical for all models, and well approximated by the equilibrium formula in eq. (1.7)

.

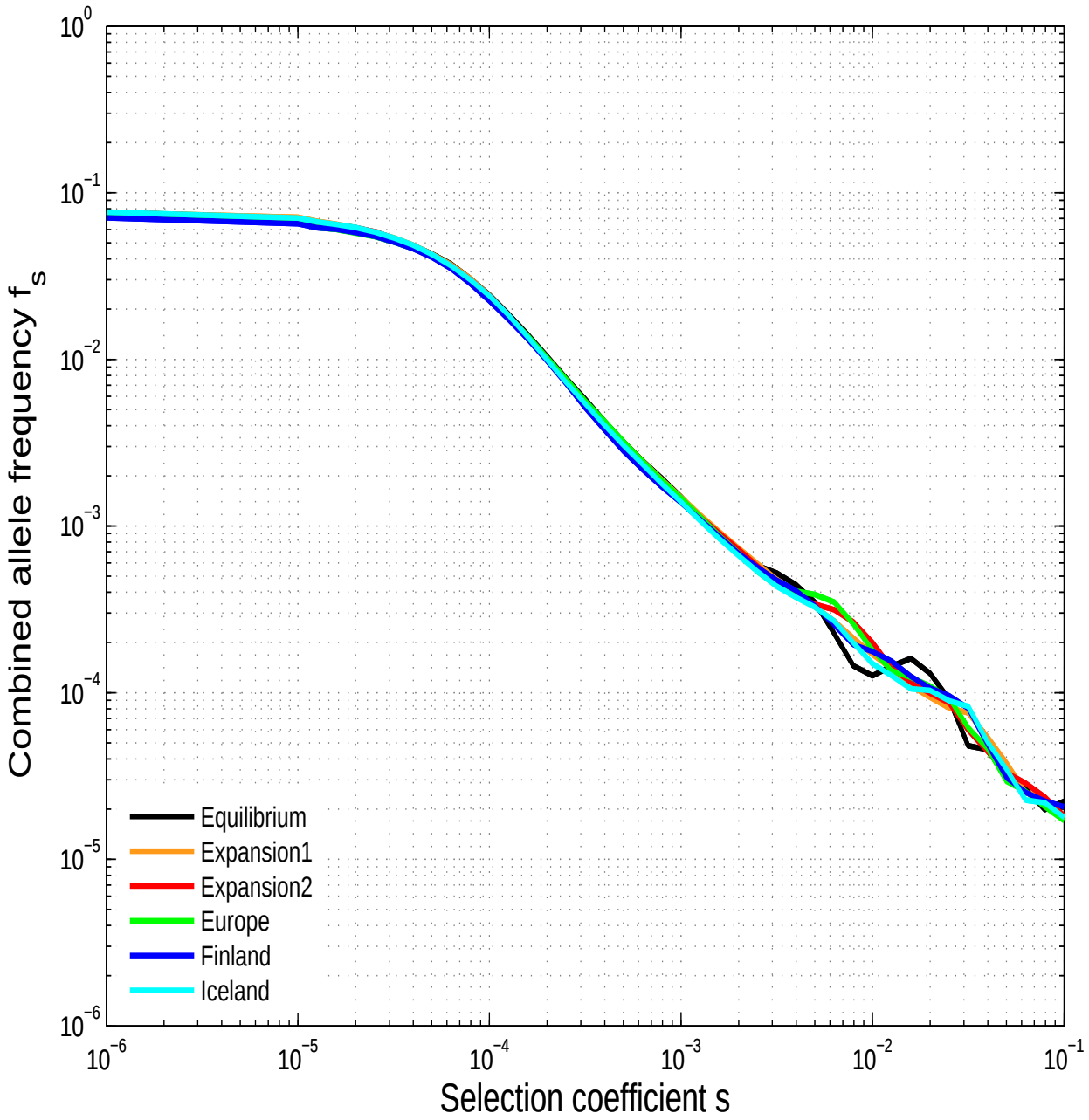

Figure 1: The Combined Allele Frequencies for all populations. Each curve shows the CAF for a different population.

2.3 Individual Allele Frequencies

While the combined allele frequency is not sensitive to the demography, different populations show substantial differences in the distribution of individual allele frequencies.

We first demonstrate this by showing the median of the individual allele frequency distribution. Figure S2 shows the median of the individual allele frequency distribution $\Psi_s(f)$ for different demographic models as a function of the selection coefficient s, and Table S3 shows the median IAF for representative values of s. The figure shows very different behaviors for the different models. For large populations with no severe bottlenecks (Expansion1, Expansion2 and Europe), the median IAF is low. For populations with bottleneck (Finland and Iceland), the median IAF is much high, because the (weighted) IAF distribution is dominated by alleles born before the bottleneck and surviving the bottleneck, which therefore have high frequency. For a population at equilibrium, the median IAF is also high, due to variations in genes.

For these three populations, and when selection is strong, the median IAF is even higher than the CAF! for example, for $s = 10^{-2}$ we have $f_{null} \approx 0.02\%$ for all populations, but the median IAF for Finland is $\approx 1.5\%$. At a first glance it seems impossible that the median IAF, representing the frequency of *one* allele, exceeds the CAF, representing the combined frequency of *all* alleles in a gene; but careful examination of the two quantities can explain this phenomena. Recall that f_{null} is the *expected* CAF for a gene, and that the IAF distribution represents sampling a random allele from *all* genes on a random chromosome in the population, and hence weights alleles by their frequency. If the CAF was constant across genes - then indeed each individual allele frequency, and in particular the median IAF, would have been lower. However, since there may be substantial variation in the CAF for different genes, a sampled allele will be more likely to be sampled from a gene with high CAF (possibly much higher than the expected CAF), and therefore might have frequency far exceeding the expected CAF. The bottleneck and equilibrium populations show high variation in the CAF, and therefore a high median IAF. We elaborate further on the consequences of variation in CAF in Section S8.

Figure S3 shows the entire cumulative distribution $\Psi_s(f)$ for different models and different values of the selection coefficient s. Again, we see clear differences between the different populations.

s	Equilibrium	Expansion1	Expansion2	Europe	Finland	Iceland
0	0.514	0.502	0.514	0.54	0.527	0.502
10^{-5}	0.49	0.478	0.478	0.514	0.502	0.478
$10^{-4.5}$	0.434	0.414	0.424	0.467	0.467	0.424
10^{-4}	0.243	0.226	0.226	0.302	0.295	0.237
$10^{-3.5}$	0.0707	0.0396	0.0396	0.109	0.118	0.048
10^{-3}	0.0211	0.00108	0.00251	0.0113	0.0642	0.00973
$10^{-2.5}$	0.006	9.36e-005	0.000346	0.000547	0.0377	0.00545
10^{-2}	0.00175	2.3e-005	4.87e-005	4.53e-005	0.00223	0.0027
$10^{-1.5}$	0.000648	6.54e-006	2.26e-006	3.32e-006	5.14e-006	0.000148
10^{-1}	0.000252	2e-006	5.04e-007	5.04e-007	5.97e-007	1.2e-005

Table 3: Median Allele Frequency for different populations. Obtained with $50,000$ simulations using $\mu_C = 1.7 \times 10^{-6}$, where alleles are sampled with weights proportional to their frequency, as described in Section 1.3. The values are sensitive to the demographic model, except perhaps for neutral alleles. Populations encountering bottleneck, such as Iceland and Finland, exhibit much higher allele frequencies, (yet fewer individual alleles), compared to populations with only expansion.

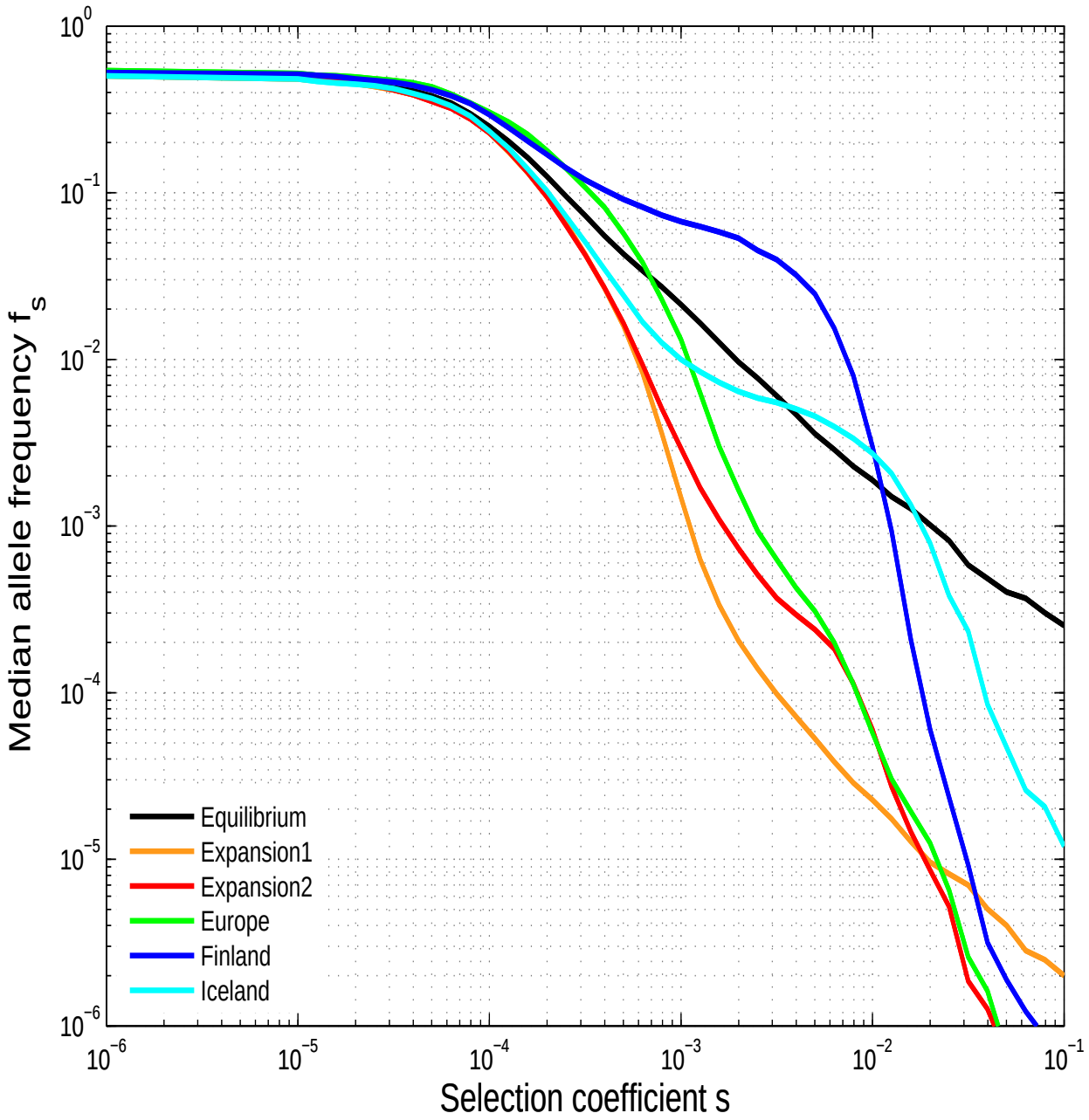

Figure 2: Median Individual Allele Frequencies in Populations for all populations. Each curve shows the median of the IAF distribution for a different population. Obtained with $50,000$ simulations using $\mu_C = 1.7 \times 10^{-6}$, where alleles are sampled with weights proportional to their frequency, as described in Section 1.3.

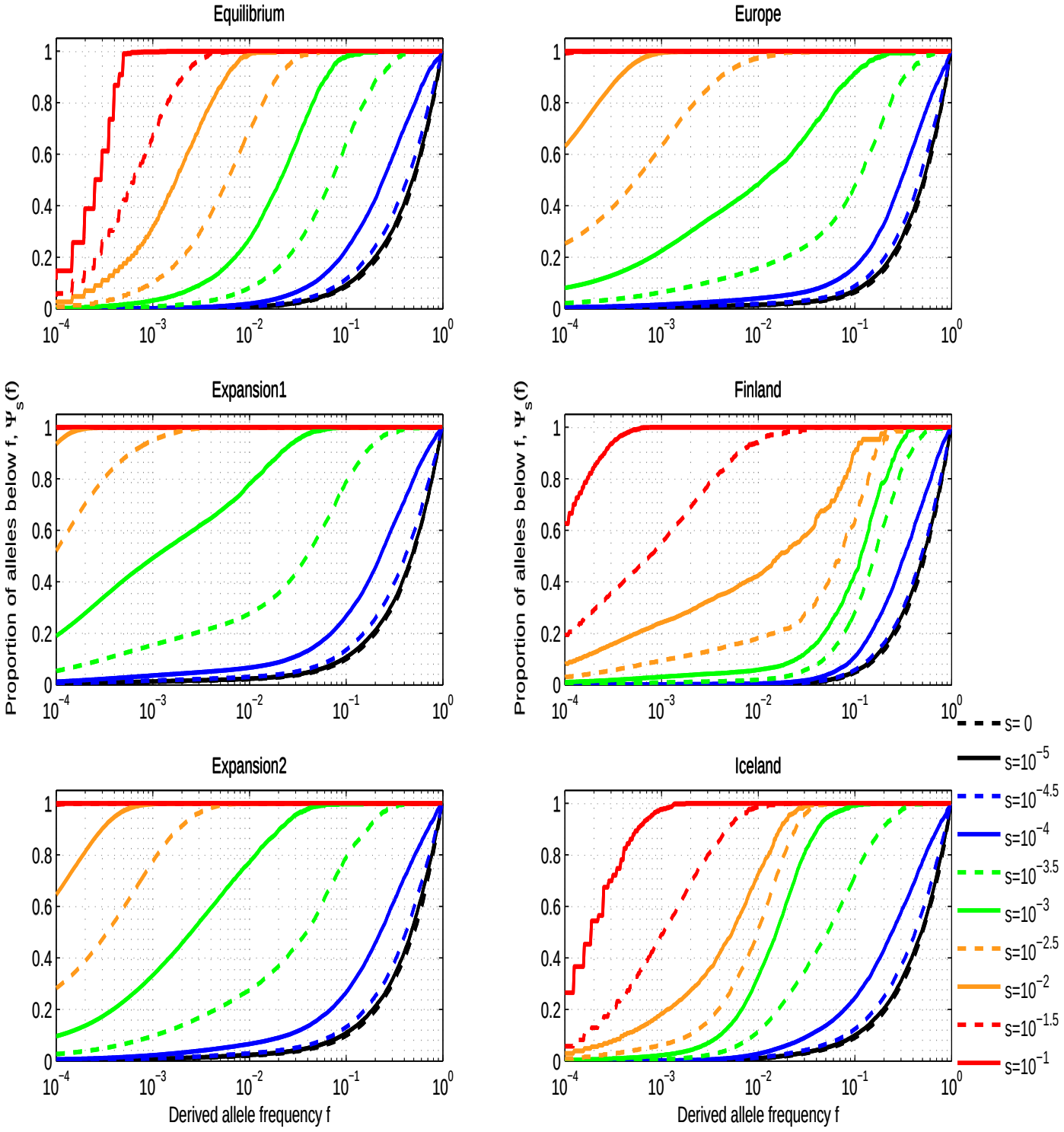

Figure 3: Cumulative Allele Frequencies in Populations for all populations. Each panel shows the cumulative distribution for a different population.

2.4 Age Distribution of Alleles

It has been recently suggested that the rapid expansion of the human population has led to a dramatic increase in the role of recent variants in human disease [17]. We explored this suggestion by using the simulations (described in the previous section) to study how the age distribution of alleles is affected by population demography. Figure S4 shows the median age of an allele on a randomly chosen chromosome in the population, as a function of the selection coefficient s, for the six populations studied. Figure S5 shows the cumulative distribution of allelic age for the different populations and for different selection coefficients. The figures show that the distribution of ages of alleles is essentially *insensitive* to demography - similarly to the situation for the combined allele frequency, and in contrast to the situation for the individual allele frequency distribution.

2.5 Proportion of Null Alleles among Missense Alleles

Missense alleles are the most challenging class to interpret, because they are a mixture of both null and neutral alleles - with the proportion varying according to the allele frequency range under consideration. Given (i) the IAF for both null and neutral alleles and (ii) the proportion of newly born missense alleles that are null, we can calculate the proportion $\rho_s(f)$ of missense alleles with frequency below f that are null,

$$\rho_s(f) = \rho_s(f;\alpha) = \frac{\alpha\Psi_s(f)f_s}{\alpha\Psi_s(f)f_s + (1-\alpha)\Psi_0(f)f_0}. \tag{2.1}$$

The proportion is shown in Figure S6, for various populations and selective coefficients.

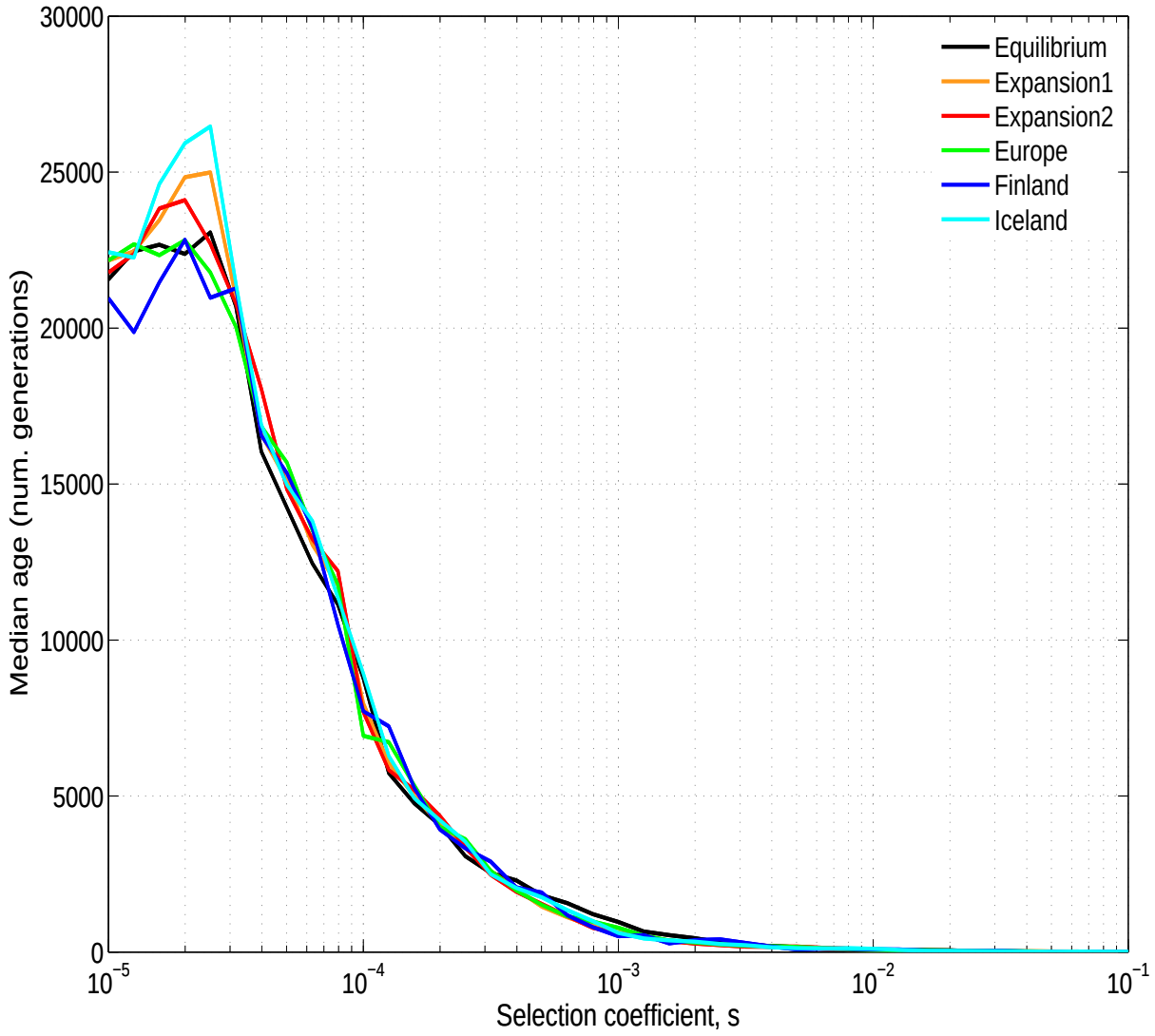

Figure 4: The median age (number of generations) of an allele drawn from a random chromosome in the population is shown as a function of the selection coefficient s. To generate this plot, alleles were drawn with probability proportional to their frequency. When selection is strong, ($s > 10^{-3}$), alleles survive for a short time in the population, and the distribution is dominated by recent alleles (last $\sim$1000 generations). When selection in weak ($s < 10^{-3}$), older alleles are more likely to survive and alleles born more than 10000 generations ago may comprise the majority of the genetic variation. For all values of s, the median age is very similar for the different populations, hence the effect of demography on the allelic age distribution is minor.

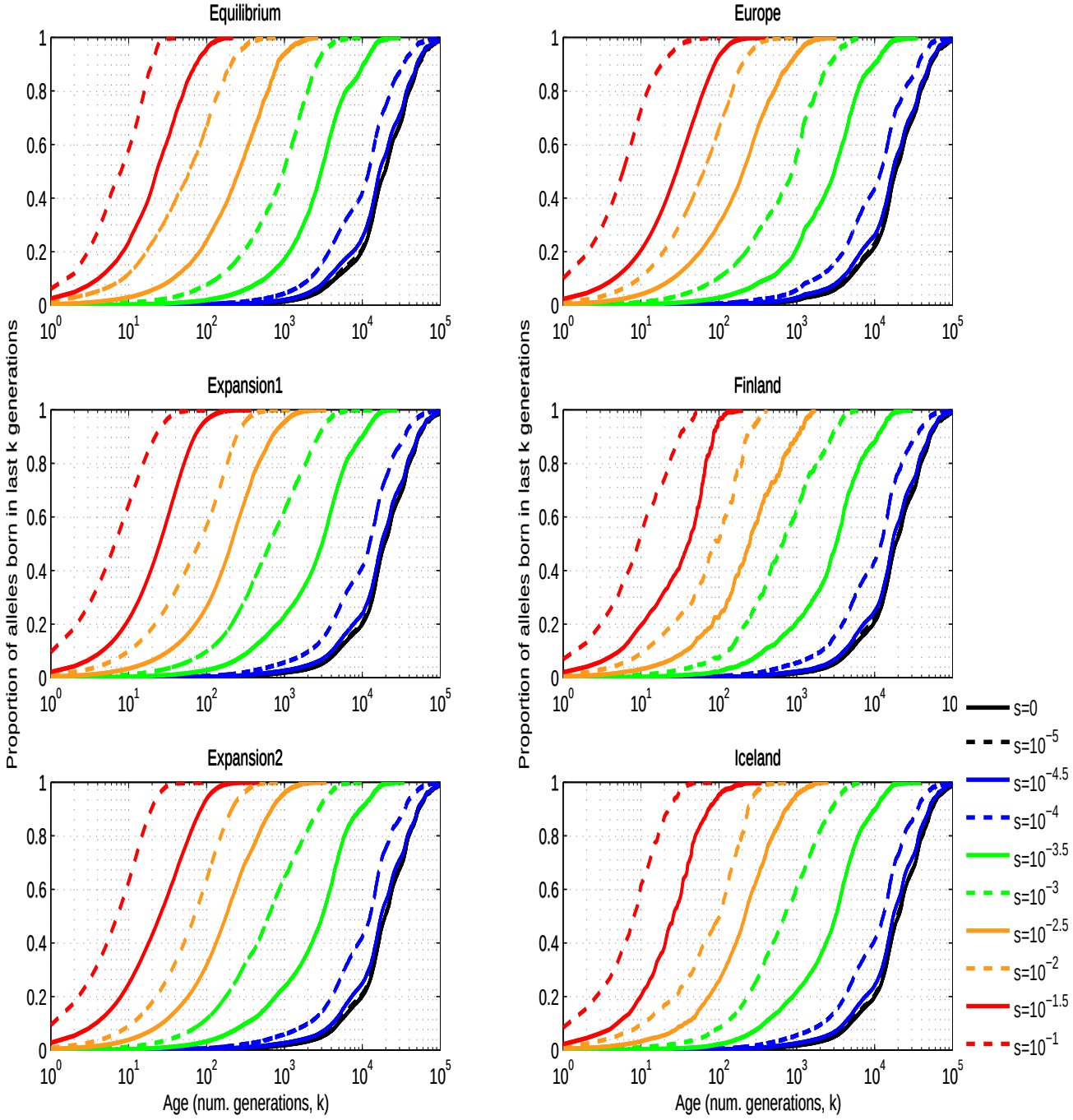

Figure 5: The cumulative age distribution (number of generations) for alleles drawn from a random chromosome in the population is shown for different populations (different panels) and different selection coefficients s (different colored curves). For strong selection, the distribution is concentrated at recent alleles, and older alleles are likely to survive for weaker selection. The different cumulative distirbutions are very similar for the different populations, showing that demography does not alter significantly the allelic age distribution.

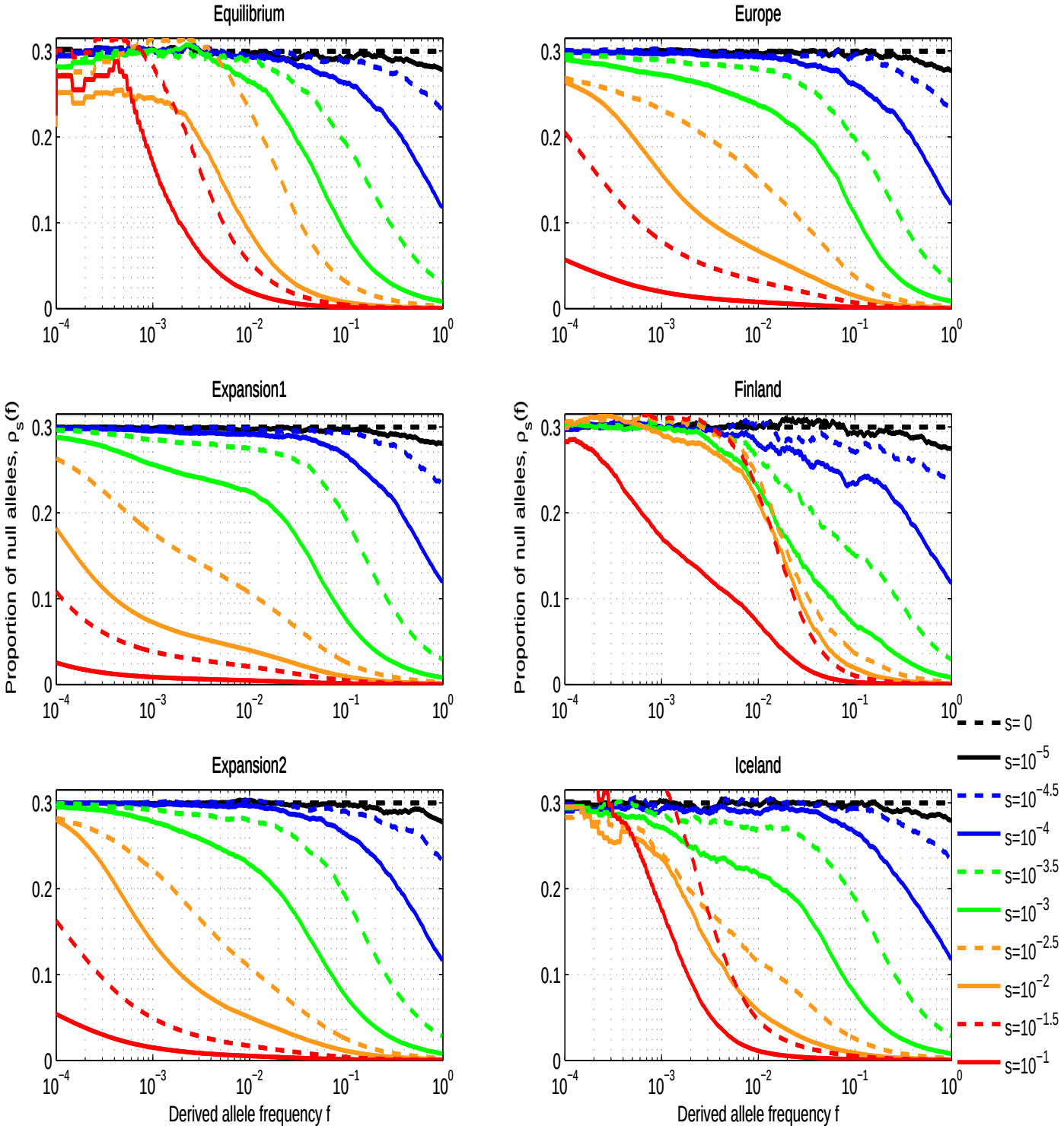

Figure 6: Fraction of missense alleles with frequency less than f that are null, for various populations. Each panel shows the proportion of null alleles $\rho_s(f)$ for a randomly sampled allele with frequency $\leq f$ for a different population. We assume $\alpha = 0.25$ - that is, $\frac{1}{4}$ of newly born alleles are null.

2.6 Mutation Rates for Observable Classes

For protein-coding regions, three classes of mutations can be directly observed: silent (S), missense (M) and disruptive (D, defined as nonsense, splice site and frameshift changes, which severely disrupt protein structure).

For any human gene, we can obtain good estimates of the mutation rates for these observable classes (S, M, D) based on (i) the length and sequence composition of the gene and (ii) the rate and mutational spectrum for point mutations in the local region [18–21]. (The latter quantities can be estimated from comparative genomics [18–21], medical genetics [22] or large-scale human parent-offspring trio sequencing studies [23].)

We estimated mutation rates (S, M, D) by extending a model [24] that incorporates gene length, sequence context and local variation in mutation rate (as inferred by fixed divergence between human and macaque). Briefly, local sequence context was used to determine the probability of each base in the coding region mutating to any each other possible base and determining the coding impact of each possible mutation.

For power calculations, we focus on a 'typical' human gene with a coding region of 1500 bases and a median mutation rate. For this typical gene, the mutation rates for silent, missense and disruptive mutations are $5.6 \times 10^{-6}, 12.8 \times 10^{-6}$, and 1.7×10^{-6} per chromosome per generation, respectively (Table 1 in the main text).

In the main text, we cite specific mutation rates D for disruptive mutations in two genes:

- *LDLR* (coding region 2583 bases, including the stop codon) $= 3.8 \times 10^{-6}$.
- *CHD8* (coding region 7746 bases, including the stop codon) $= 5 \times 10^{-6}$.

The rate for *LDLR* is slightly higher and for *CHD8* considerably lower than expected by simply adjusting for gene length, with the differences being due to sequence context (largely GC content).

3 Calculating Sample Sizes required for RVAS

In order to understand the prospects of RVAS for identifying disease-associated loci, it is important to study the detection power. Power calculations are also useful for guiding the design of RVAS (e.g. what sample size should be used, what populations, what classes of alleles and tests to employ etc.). We describe how to calculate the sample size for RVAS, using a burden test for a class C of alleles. We start with the simple case of a 'pure' class of alleles with identical effect size (such as disruptive alleles). Then, we build on these sample size calculations to analyze the gains in power achieved by different RVAS strategies.

3.1 Estimating Aggregate Effect Size

We consider a disease D with prevalence π in the population. We let f denote the cumulative allele frequency of alleles in class C, and let $1+\lambda$ denote the average relative risk conferred by alleles in C. (We thus allow the possibility that C is a mixture of alleles with different properties). By Bayes' formula, the probability that an individual carrying an allele in class C is affected is $f(1+\lambda)$.

$$Pr(X_C = 1|Z = 1) = \frac{Pr(Z = 1|X_C = 1)Pr(X_C = 1)}{Pr(Z = 1)} = \frac{f\pi(1+\lambda)}{\pi} = f(1+\lambda). \tag{3.1}$$

where Z is an indicator variable for disease status (with 1 indicating that the individual is affected and 0 indicating that the individual is not affected) and X_C is an indicator variable for carrier status for an allele in class C (with 1 indicating a carrier and 0 indicating a non-carrier).

Suppose that we gather data from n affected individuals (cases), and observe d total alleles of class C. We can estimate the allele frequency in cases by, $\hat{f}_{cases} = \frac{d}{2n}$ (where we divide by the total number of *chromosomes*). From the above, we have $E[\hat{f}_{cases}] = f(1+\lambda)$.

3.1.1 Risk for Carriers of Multiple Allele

Our model assumes that a carrier of a null rare allele has risk $\pi(1+\lambda)$, regardless of the number of null alleles carried. We thus assume that individuals who are homozygous (or compound heterozygous) for null alleles have the same disease risk as for heterozygotes. In fact, these individuals are likely to have a greater disease risk. However, because the cumulative frequency of rare alleles is low ($f \ll 1$), the frequency of these individuals is so small (on the order of f^2) and their higher disease risk does not significantly affect our calculations. We thus ignore any increased disease risk associated with homozygosity (or compound heterozygosity) for null alleles.

3.1.2 Protective Alleles

While the text typically speaks about null alleles increasing disease risk, our mathematical calculations apply equally well to alleles that decrease disease risk (that is, protective alleles) [25, 26]. In this case the parameter λ is simply negative, $-1 \leq \lambda < 0$. For example, $\lambda = -1$ corresponds to fully protective alleles (i.e. carriers are completely immune to the disease). As described in the main text, detecting protective alleles requires larger sample sizes than detecting harmful alleles (see Figure S8).

3.2 Calculating Sample Size when f is Already Known

We first suppose that f is already known to high precision - for example, from a large population survey.

Suppose that we gather data on n cases, and observe d total alleles of class C. We model our observations using a binomial distribution $d \sim Binomial(2n, f_{cases})$. We use the estimator for the allele frequency in cases, $\hat{f}_{cases} = \frac{d}{2n}$, and perform a simple hypothesis test, with,

$$\begin{aligned} H_0 &: f_{cases} = f \ , \\ H_1 &: f_{cases} > f \end{aligned} \tag{3.2}$$

We use the likelihood-ratio test statistic

$$\begin{aligned} 2\mathcal{LLR}(d, n; f) &= 2[LL(d, n; \hat{f}_{cases}) - \mathcal{LL}(d, n; f))] \\ &= 2[d \log \frac{\hat{f}_{cases}}{f} + (2n - d) \log \frac{1 - \hat{f}_{cases}}{1 - f}] \end{aligned} \tag{3.3}$$

where $\mathcal{LL}$ denotes the Log-Likelihood and $\mathcal{LLR}$ denotes the Log-Likelihood-Ratio. Under the null hypothesis H_0, the test statistic $2\mathcal{LLR}$ has a $\chi^2(1)$ distribution. Under the alternative H_1, it has a non-central $\chi^2(1; NCP)$ distribution, with the non-centrality parameter NCP,

$$\begin{aligned} NCP = NCP(n, \lambda, f) &= E[2\mathcal{LLR}] \\ &= n\Big[(1 + \lambda) f log(1 + \lambda) + (1 - (1 + \lambda) f) log \frac{1 - (1 + \lambda) f}{1 - f}\Big]. \end{aligned} \tag{3.4}$$

The power to detect an allele, when setting a significance threshold a (for example, $a = 2.5 \times 10^{-6}$ to get 0.05 false-positives and account for testing 20,000 genes), is approximately

$$1 - b = 1 - b(n, \lambda) = 1 - F_{\chi^2}\Big(F^{-1}_{\chi^2}(1 - a, 1), 1, NCP\Big) \tag{3.5}$$

where b is defined as the false-negatives rate (type-2 error).

We checked that the above formula provides an excellent analytic approximation to power, by comparing it to Fisher's exact test (Figure S7). Similarly good approximations are obtained for the other tests described in the next sections.

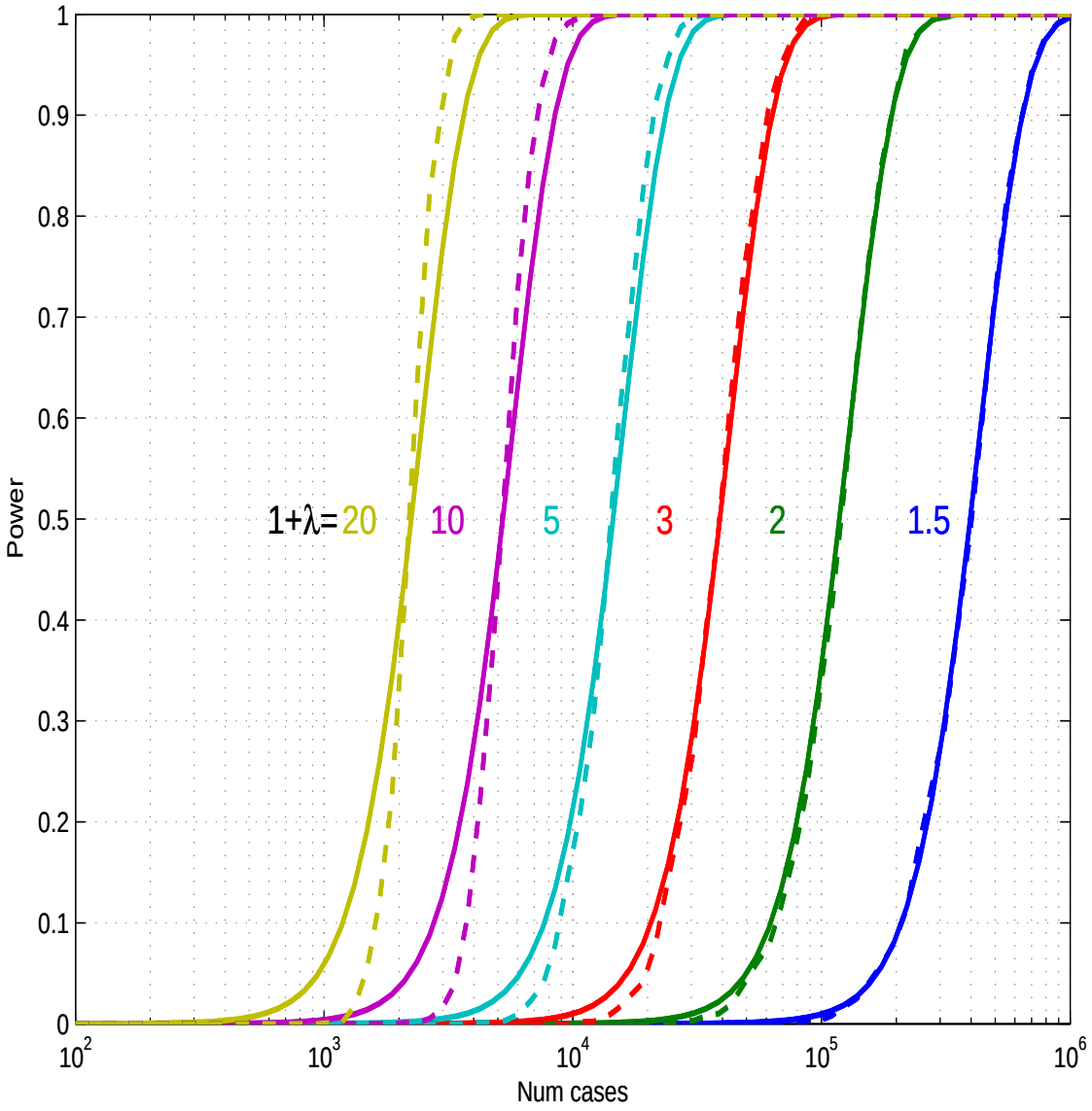

Figure 7: Power to detect an association as function of sample size (x-axis, number of cases) and effect size (different colored curves), comparing the analytic approximation from eq. (3.5) (solid curves) and simulations using Fisher's exact test (dashed curves). For each value of the sample size (number of cases), we simulated 1000 different realization of cases and computed the empirical power as the fraction of simulations in which the test statistics exceeded the threshold set by the significance level α. (We assumed cumulative allele frequency $f = 0.1\%$, significance level $\alpha = 0.05/20,000$ and a disease prevalence of 5%.) Overall, there is a very good agreement between the analytic approximation and empirical power, especially for small effect sizes and for intermediate power ($\sim 50\%$). For large effect sizes, the analytic approximation slightly under-estimates power for high power (e.g. $\sim$90%), and slightly over-estimates power for low power (e.g. $\sim$10%).

For given type-1 error a and type-2 error b, we introduce a specific value of a non-centrality parameter denoted by $\nu_{a,b}$. We define $\nu_{a,b}$ to be the value such that if a test statistic has a non-central $\chi^2(1;\nu_{a,b})$ distribution under the alternative hypothesis H_1, then testing under H_0 using this test statistic, assuming a $\chi^2(1)$ distribution, and setting the significance level to a, would indeed give power of $1-b$. The specific value $\nu_{a,b}$ for given a and b is obtained by solving equation (3.5) above for NCP. It depends only on the type-1 and type-2 errors. We can compute the required sample size by matching this value to the NCP actually obtained in a given study for a given effect size and sample size. Specifically, from eq. (3.4), the sample size is then simply given by,

$$n_{a,b} = \frac{\nu_{a,b}}{4[(1+\lambda)f log(1+\lambda) + (1-(1+\lambda)f) log\frac{1-(1+\lambda)f}{1-f}]}. \tag{3.6}$$

When f is small, $f \ll 1$, we can approximate $n_{a,b}$ by,

$$n_{a,b} \approx \frac{\nu_{a,b}}{4f[(1+\lambda)\log(1+\lambda) - \lambda]} = \frac{\nu_{a,b}}{4fg(\lambda)}. \tag{3.7}$$

where we define

$$g(\lambda) \equiv (1+\lambda)\log(1+\lambda) - \lambda. \tag{3.8}$$

Thus, the sample size (number of cases) is roughly inversely proportional to the combined allele frequency f and the value of $g(\lambda)$. The function $g(\lambda)$ is plotted in Figure S8 for both harmful $(1+\lambda > 1)$ and protective $(1+\lambda < 1)$ alleles.

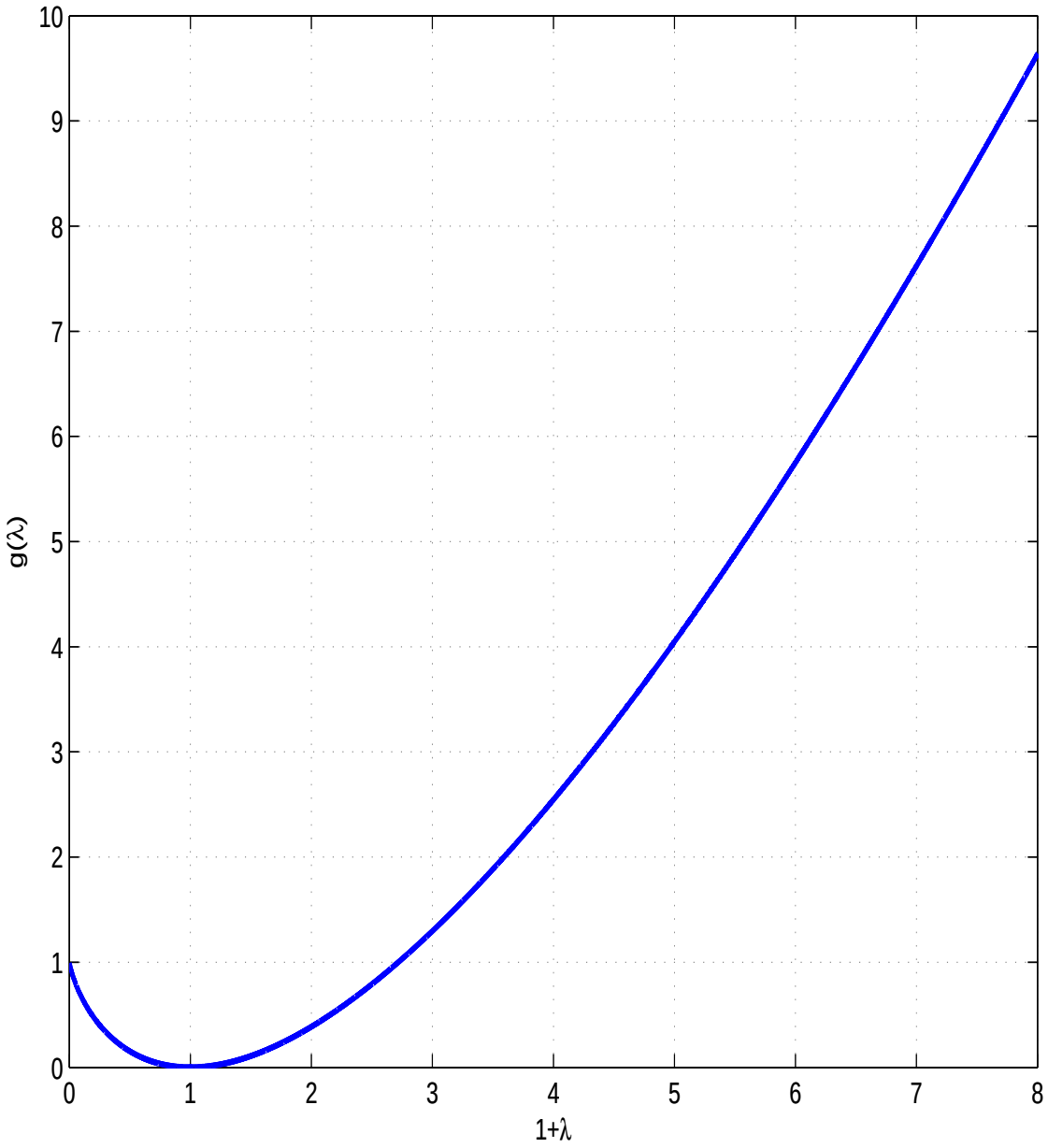

Figure 8: The function $g(\lambda)$, representing power, which is proportional to the inverse sample size as a function of λ, the effect size. $g(\lambda)$ is shown on the y-axis, plotted against the relative disease risk $1+\lambda$, shown on the x-axis for both harmful and protective alleles. In general, the function is $g(\lambda)$ is lower for protective alleles $(1+\lambda<1)$ than for harmful alleles $(1+\lambda>1)$, hence protective alleles have lower detection power and will require larger sample sizes for detection. For example, for a completely protective allele $(\lambda=-1)$ we have $g(-1)=g(e)=1$, for $e\approx 2.718$, thus the power to detect a completely protective allele is roughly equal to the power to detect a harmful allele with a relative risk of ≈ 2.7.

Under strong selection, we have $f = \frac{\mu_C}{s}$ from eq. (1.8). This gives,

$$n_{a,b} \approx \frac{\nu_{a,b} s}{4\mu_C g(\lambda)}. \tag{3.9}$$

Thus, the sample size is (approximately) directly proportional to the selection coefficient.

When λ is relatively small, $\lambda \ll 1$, the above two formulas further simplify to,

$$n_{a,b} \approx \frac{\nu_{a,b}}{2f\lambda^2} \approx \frac{\nu_{a,b} s}{2\mu_C \lambda^2}. \tag{3.10}$$

Thus, the sample size is proportional to $\frac{1}{\lambda^2}$, which is the same as given by a simple z-score approximation (see next). However, when λ is large, this z-score approximation over-estimates power. In fact, the true power is inversely proportional to roughly $\lambda \log(\lambda)$ (and not λ^2), from eq. (3.7).

We can compare this to the following crude approximation, based on a z-score,

$$n_{a,b} \approx \frac{[\Phi^{-1}(a) - \sqrt{1+\lambda}\Phi^{-1}(b)]^2}{2\lambda^2 f} \tag{3.11}$$

which simplifies when $\lambda \ll 1$ to,

$$n_{a,b} \approx \frac{[\Phi^{-1}(a) - \Phi^{-1}(b)]^2}{2\lambda^2 f}. \tag{3.12}$$

For example:

1. If we take $b = 0.1$ (i.e. 90% power) and $a = 2.5 \times 10^{-6}$, we get, $\nu_{a,b} = 35.9$. For small λ and f, we get $n \approx \frac{17.9}{\lambda^2 f}$. The crude approximation gives us $n \approx \frac{17.1}{\lambda^2 f}$, which is slightly optimistic.
2. If we are satisfied with a nominal significance level $a = 0.05$ (that is, if we were certain that we would only ever test a single gene), we could use $\nu_{a,b} = 10.5$. For small λ and f, the required sample size is then roughly 3.5-fold lower at $\frac{5.3}{\lambda^2 f}$.

3.3 Calculating Sample Size when f is Unknown

In the above calculations, we assumed that the population frequency f is known precisely. If f is unknown, we can estimate it by performing a case-control study or a case-unaffected study. Here "controls" refer to random individuals in the general population and "unaffecteds" to individuals known not to have the disease (based on the diagnostic criteria in use).

3.3.1 Using unaffecteds

We begin by considering a case-unaffected study. The control frequency $f_{unaffected}$ satisfies $f = (1-\pi)f_{unaffected} + \pi f_{cases}$, which gives

$$f_{unaffected} = \frac{1-\pi-\pi\lambda}{1-\pi} f,$$

$$f_{cases} = \Big(1 + \frac{\lambda}{1-\pi-\pi\lambda}\Big) f_{unaffected}. \tag{3.13}$$

The observed numbers of carriers in cases and controls have the following binomial distributions,

$$d_{cases} \sim Binomial(2n_{cases}, f_{cases}),$$

$$d_{controls} \sim Binomial(2n_{controls}, f_{unaffected}). \tag{3.14}$$

This time, we test the hypothesis,

$$H_0 : f_{cases} = f_{unaffected}\ ,$$

$$H_1 : f_{cases} > f_{unaffected}. \tag{3.15}$$

i.e. we test whether the parameters of two binomial random variables are equal. Again, we write the likelihood ratio statistic

$$\begin{aligned} &2\mathcal{LLR}\big(n_{cases}, d_{cases}, n_{controls}, d_{controls}\big) \\ &= 2\Big[LL\big(n_{cases}, d_{cases}, n_{controls}, d_{controls}; (\hat{f}_{cases}, \hat{f}_{unaffected})\big) \\ &- LL\big(n_{cases}, d_{cases}, n_{controls}, d_{controls}; \hat{f}\big)\Big] \end{aligned} \tag{3.16}$$

with,

$$\hat{f}_{cases} = \frac{d_{cases}}{2n_{cases}},$$

$$\hat{f}_{unaffected} = \frac{d_{controls}}{2n_{controls}},$$

$$\hat{f} = \frac{d_{cases} + d_{controls}}{2(n_{cases} + n_{controls})}. \tag{3.17}$$

The non-centrality parameter is

$$\begin{aligned} NCP = E[2\mathcal{LLR}] &= 4n_{cases}h_b(f_{cases}) + 4n_{controls}h_b(f_{unaffected}) \\ &- 4(n_{cases} + n_{controls})h_b\Big(\frac{f_{cases}n_{cases} + f_{unaffected}n_{controls}}{n_{cases} + n_{controls}}\Big) \end{aligned} \tag{3.18}$$

where h_b is the binary entropy function (with logarithm taken at the natural basis):

$$h_b(x) = -[x\log(x) + (1-x)\log(1-x)]. \tag{3.19}$$

We assume that the case-controls ratio is $\rho : (1-\rho)$, and define $n_{cases,a,b}, n_{controls,a,b}$ the number of cases and controls, respectively, required to reach false-positive and false-negative rates a and b, respectively. The overall sample size (cases+controls) is $n_{a,b} = n_{cases,a,b} + n_{controls,a,b}$ with $n_{cases,a,b} = \rho n_{a,b}$ and $n_{controls,a,b} = (1-\rho)n_{a,b}$. From eq. (3.18) we get,

$$n_{a,b} = \frac{\nu_{a,b}}{4\Big\{\rho h_b((1+\lambda)f) + (1-\rho)h_b(\frac{1-\pi+\pi\lambda}{1-\pi}f) - h_b([\rho(1+\lambda) + (1-\rho)\frac{1-\pi+\pi\lambda}{1-\pi}]f)\Big\}}. \tag{3.20}$$

When f is small, we get the following approximate formula,

$$n_{a,b} \approx \frac{\nu_{a,b}}{4f[\rho g(\frac{\lambda}{1-\pi-\pi\lambda}) - g(\frac{\rho\lambda}{1-\pi-\pi\lambda})]} \tag{3.21}$$

When the effect size λ is also small, Taylor expansion gives,

$$n_{a,b} \approx \frac{\nu_{a,b}(1-\pi)^2}{2f\rho(1-\rho)\lambda^2} \tag{3.22}$$

and the number of cases is,

$$n_{cases,\alpha,\beta} \approx \frac{\nu_{a,b}(1-\pi)^2}{2f(1-\rho)\lambda^2}. \tag{3.23}$$

3.3.2 Using controls from the population

If we perform a case-control study (where individuals from the the control group are taken from the general population), the control frequency is simply f. Formula (3.20) changes to,

$$n_{a,b} = \frac{\nu_{a,b}}{4\Big\{\rho h_b((1+\lambda)f) + (1-\rho)h_b(f) - h_b((1+\rho\lambda)f)\Big\}}. \tag{3.24}$$

When f is small, we get the following approximate formula,

$$n_{a,b} \approx \frac{\nu_{a,b}}{4f[\rho g(\lambda) - g(\rho\lambda)]}. \tag{3.25}$$

When the effect size λ is also small, Taylor expansion gives,

$$n_{a,b} \approx \frac{\nu_{a,b}}{2f\rho(1-\rho)\lambda^2} \tag{3.26}$$

and the number of cases is,

$$n_{cases,\alpha,\beta} = \rho n_{a,b} \approx \frac{\nu_{a,b}}{2f(1-\rho)\lambda^2}. \tag{3.27}$$

Compared to the case-only test (eq. (3.7)), the number of cases required increases by roughly a factor of $1-\rho$. For example, if we take a balanced case-control study ($n_{cases} = n_{controls}$), for $a = 2.5 \times 10^{-6}, b = 0.1$ we get $n_{cases,a,b} = \frac{35.9}{f\lambda^2}$, compared to $n_{a,b} \approx \frac{17.9}{f\lambda^2}$ for the case-only test with known f (when effect size λ is small). The resulting number of cases is increased $\sim$ 2-fold (the overall sample size is increased 4-fold since we also need to genotype controls).

When the disease is rare ($\pi \ll 1$) we have $f \approx f_{unaffected}$, $f_{cases} = (1+\lambda)f \approx (1+\lambda)f_{unaffected}$ and the formulas for the case-control and case-unaffected studies coincide.

3.4 Tests Proposed in the Literature for RVAS

Given the early stage of the field, the analytical methodology for RVAS remains in flux. Many authors have proposed a rich collection of possible statistics (e.g. [27–35]; reviewed in [36] and [37]; tests compared in [38]. Our goal here is neither to evaluate their relative merits nor to propose alternatives. The tests include 'burden" tests, designed to detect allelic effects in the same direction, and "overdispersion" tests, designed to detect bi-directional allelic effects [27, 39, 40]. One can incorporate weights based on allele frequency [41–43], apparent effect size [39, 44] and presumed functional significance [39, 42, 45].

4 RVAS Strategy 1: Studying Disruptive Alleles

The simplest strategy would be to study just the disruptive alleles, which we assumed to be a homogenous class of null alleles with equal effect size and selection coefficient. To compute power, we can use the formulas for Section 3. In Figure S9 (reproduced Figure 2(a) from the main text) we show the power to detect harmful disruptive alleles as a function of either the selection coefficient s or the combined allele frequency f_D (assumed to be interchangeable through the expectation $f_D = \frac{\mu_D}{s}$) and their effect size λ. Similarly, in Figure S10 (reproduced Figure 2(b) from the main text) we show the power to detect protective disruptive alleles.

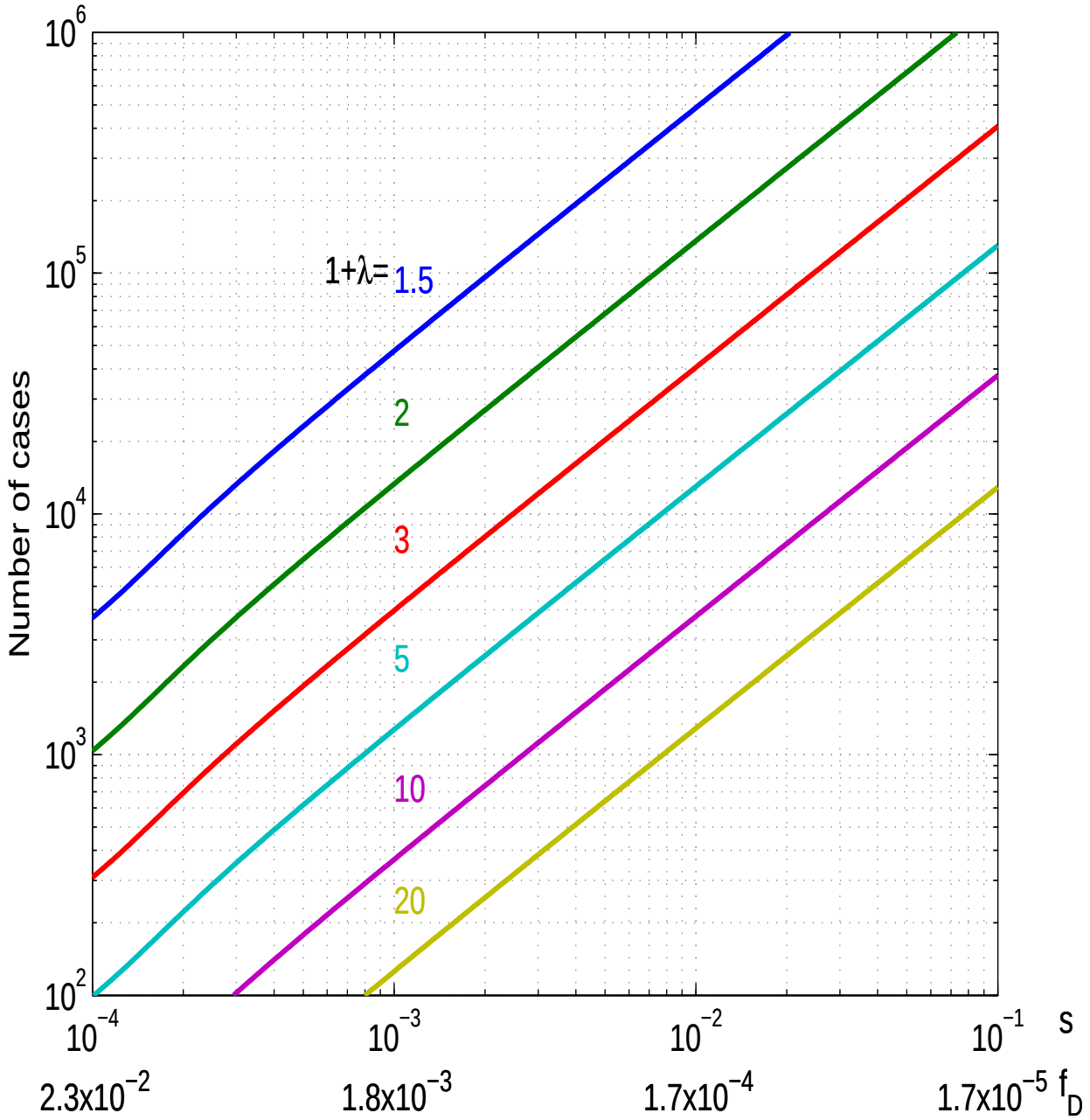

Figure 9: Sample size (number of cases) required to detect association using disruptive alleles in a case-only study with known population allele frequency, as function of the either f_D, the combined allele frequency or s, the selection coefficient (x-axis), and λ, the effect size (different colored curves) (Figure 2(a) in main text).

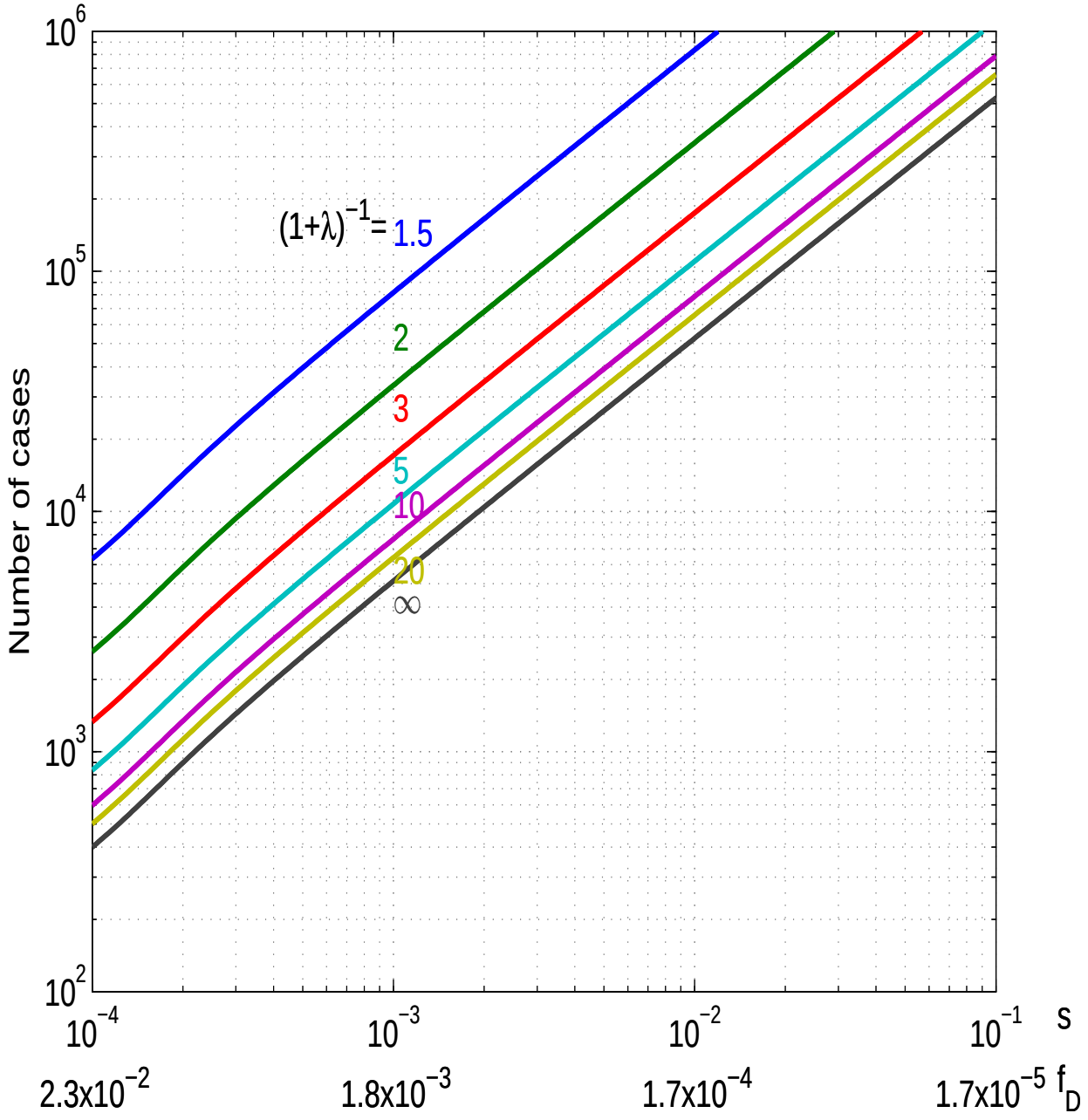

Figure 10: Sample size required to detect association using disruptive protective alleles, as function of the either f_D, the combined allele frequency or s, the selection coefficient (x-axis), and effect size λ (different colored curves) (Figure 2(b) in main text). For protective alleles, the effect size λ is negative - for example $\lambda = -0.95$ for a $(1+\lambda)^{-1} = 20$-fold depletion. The sample size needed to detect the depletion of rare protective alleles in cases is considerably higher than the sample size needed to detect the excess of rare harmful alleles (Figure S9). As protection increases, the sample size curve converges to a curve representing completely protective alleles (corresponding to $(1+\lambda)^{-1} = \infty$), and is roughly equal to the sample size required to detect alleles increasing disease probability by a factor of $e \approx 2.718$.

5 RVAS Strategy 2: Incorporating Missense Alleles

One attractive way to increase power is to incorporate missense alleles into the burden analysis. Only a fraction α of missense alleles are null and thereby contribute to disease association. If we could perfectly identify this subset, we could simply include both disruptives and missense-null alleles in our burden test. This would be advantageous because the number of missense-null alleles typically exceeds the number of disruptive alleles. Specifically, the expected ratio of null missense alleles to disruptive alleles is $\frac{\mu_{M-null}}{\mu_D} = \frac{\alpha\mu_M}{\mu_D}$. For our 'typical' gene with $\alpha = 0.25$ and $\frac{\mu_M}{\mu_D} = 7.3$ (Table 1 in the main text), this ratio is $\sim$2.8.

The problem is that we cannot readily distinguish the missense nulls from missense neutrals. The latter are more abundant (because they are not under purifying selection) and thus may swamp the former. We can try to avoid this problem by focusing only on missense alleles with frequency below a specified threshold T. There are two considerations:

1. The ratio of such alleles in cases vs controls will be smaller than the true effect size $1 + \lambda$. Instead, we will see an "apparent" effect size of $1 + \rho_s(T)\lambda$, because the signal is diluted by neutral alleles.

2. The total number of alleles under consideration is reduced to $\Psi_s(T)f_M$.

There is an important trade-off. Decreasing the frequency threshold T (1) enriches for missense null alleles and thereby increases the apparent effect size (although it never exceeds $1 + \alpha\lambda$) but (2) decreases the total number of alleles under consideration. Our goal then is to choose the optimal threshold T^* maximizing power. The optimal value depends on s, α, λ and demography.

For a given threshold T, the non-centrality-parameter is

$$\begin{aligned} NCP_{missense}(T) &= E[\mathcal{LLR}_{missense}(T)] \\ &\approx 4n\Big[(1 + \lambda\rho_s(T))\frac{\mu_M}{\mu_D} f log(1 + \lambda\rho_s(T)) \\ &+ (1 - (1 + \lambda\rho_s(T))\frac{\mu_M}{\mu_D} f) log\frac{1 - (1 + \lambda\rho(T))\frac{\mu_M}{\mu_D} f}{1 - \frac{\mu_M}{\mu_D} f}\Big] \end{aligned} \tag{5.1}$$

where we take the target size of missense alleles to be roughly $\frac{\mu_M}{\mu_D}$ times the target size of disruptive alleles. The optimal threshold is,

$$T^* \equiv \underset{T}{argmax}\, NCP_{missnense}(T). \tag{5.2}$$

We will assume that selection is strong enough that the total allele frequency f is quite low. In this case, we can approximate the power ratio as follows,

$$\frac{E[\mathcal{LLR}_{missense}(T^*)]}{E[\mathcal{LLR}_D]} \sim \frac{\mu_M \alpha \Psi_s(T^*) g(\rho_s(T^*)\lambda)}{\mu_D \rho_s(T^*) g(\lambda)}. \tag{5.3}$$

For small effect sizes ($\lambda \to 0$), we get,

$$\frac{E[\mathcal{LLR}_{missense}]}{E[\mathcal{LLR}_D]} \sim \frac{\mu_M}{\mu_D} \alpha \Psi_s(T^*) \rho_s(T^*). \tag{5.4}$$

For large λ (but with the constraint $f\lambda \ll 1$, that is, the risk of disease in carriers is significantly smaller than 100%), we get,

$$\frac{E[\mathcal{LLR}_{missense}]}{E[\mathcal{LLR}_D]} \sim \frac{\mu_M}{\mu_D} \alpha \Psi_s(T^*). \tag{5.5}$$

For large values of λ, the dilution $\rho_s(T^*)$ has little effect on power: the power is similar to the ideal situation where we can perfectly recognize null missense alleles. The reason is that when effect sizes are very large, the combined allele frequency in cases is barely altered due to the dilution effect.

In general, as the effect size λ increases, the effect of dilution $\rho_s(T^*)$ decreases, and it becomes more beneficial to set a higher threshold T^*, (up to the maximum frequency of $T^* = 1$).

Figures S11, S12 and S13 show the relative value of the set of missense alleles (normalized to the value of disruptive mutations), as a function of the frequency threshold T used to filter the missense alleles. The three figures correspond to medium, large and small effect sizes ($1 + \lambda = 4, 16$ and 1.5); the six panels within each figure correspond to different population models; and the curves within each panel correspond to different values of s.

The value of missense alleles is sensitive to the choice of T, and the optimal value T^* depends on the value of s. Although the optimal value T^* varies with s, the power achieved at this threshold is fairly similar across values of s. For medium effect size ($1 + \lambda = 4$), the optimal threshold T^* results in missense alleles contributing roughly an equal amount of information as disruptive alleles - that is, the peak is typically around 1.0. (When combining information from both disruptive and missense alleles, the sample size is thus $1.5 - 2.0$-fold smaller). For large effect size, the contribution is larger ($\sim$1.5-fold) for large effect size and smaller ($\sim$0.60-fold). The optimal threshold T^* itself is relatively insensitive to the effect size (increasingly only slightly with λ).

The optimal threshold T^* typically occurs around the 80% percentile of the IAF for null alleles (that is, we retain $\Psi(T^*) \sim 80\%$ of the null alleles) and the proportion of nulls among all missense alleles below this frequency threshold tends to be $\rho_s(T^*) \sim 15\% - 20\%$ (see Tables S4-S9). Figure S14 shows how the optimal threshold T^* relates to $\Psi(T^*)$ and $\rho_s(T^*)$.

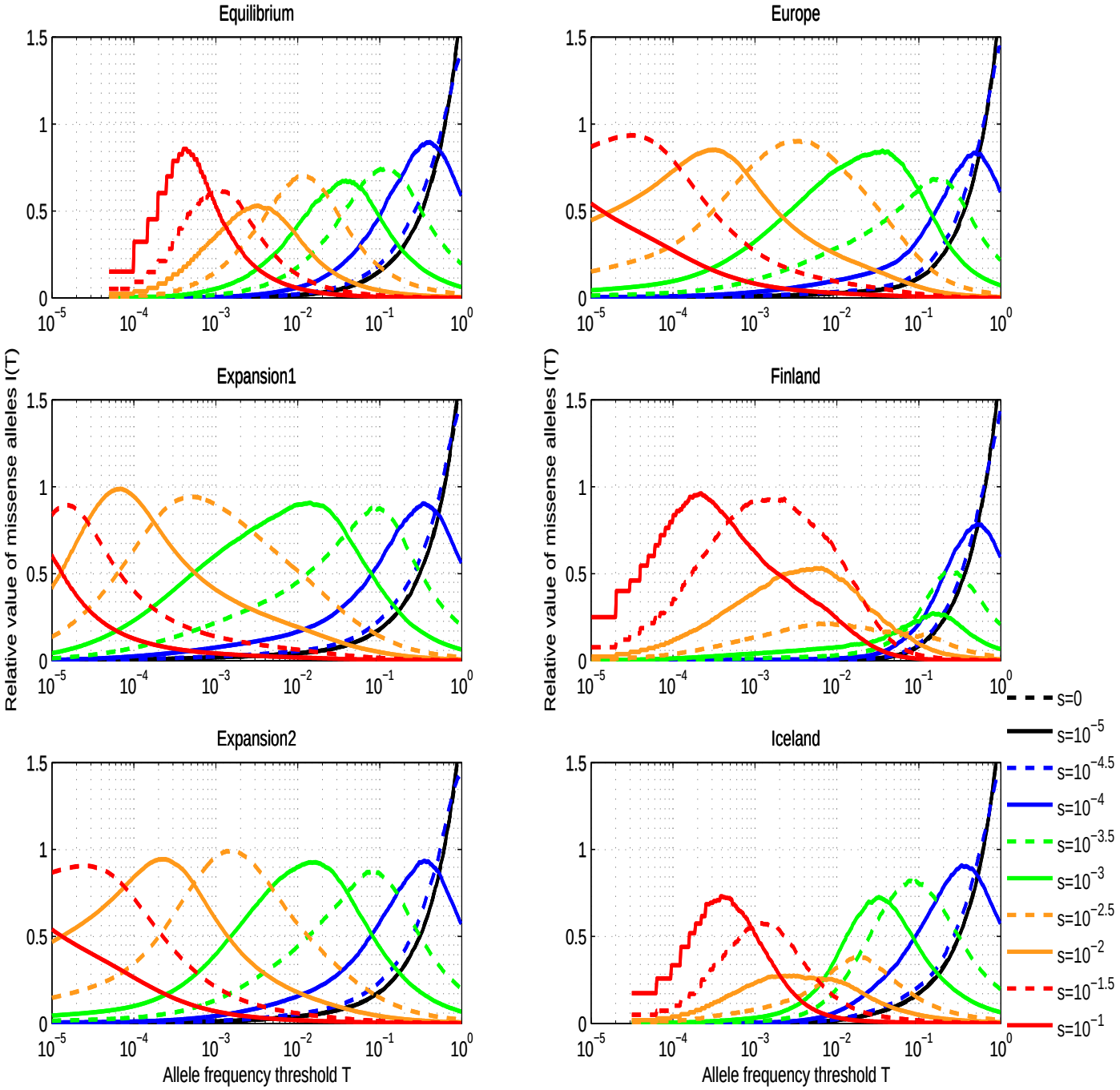

Figure 11: The relative contribution of missense alleles to power as function of threshold used, for different demographic models (shown in different panels), and different selection coefficients (colored curves). The precise value of threshold used depends heavily on the population and the selection coefficient, and is usually at the $\sim 80\%$ quantile of the allele frequency distribution. The relative power of missense alleles from using the optimal threshold is, however, more stable, at around $0.5 - 1$ fold in most cases, leading to a $1.5 - 2$-fold improvement over using only disruptive alleles. We assume that the proportion of null alleles among newborn missense alleles is $\alpha = 0.25$, and an effect size of $1 + \lambda = 4$.

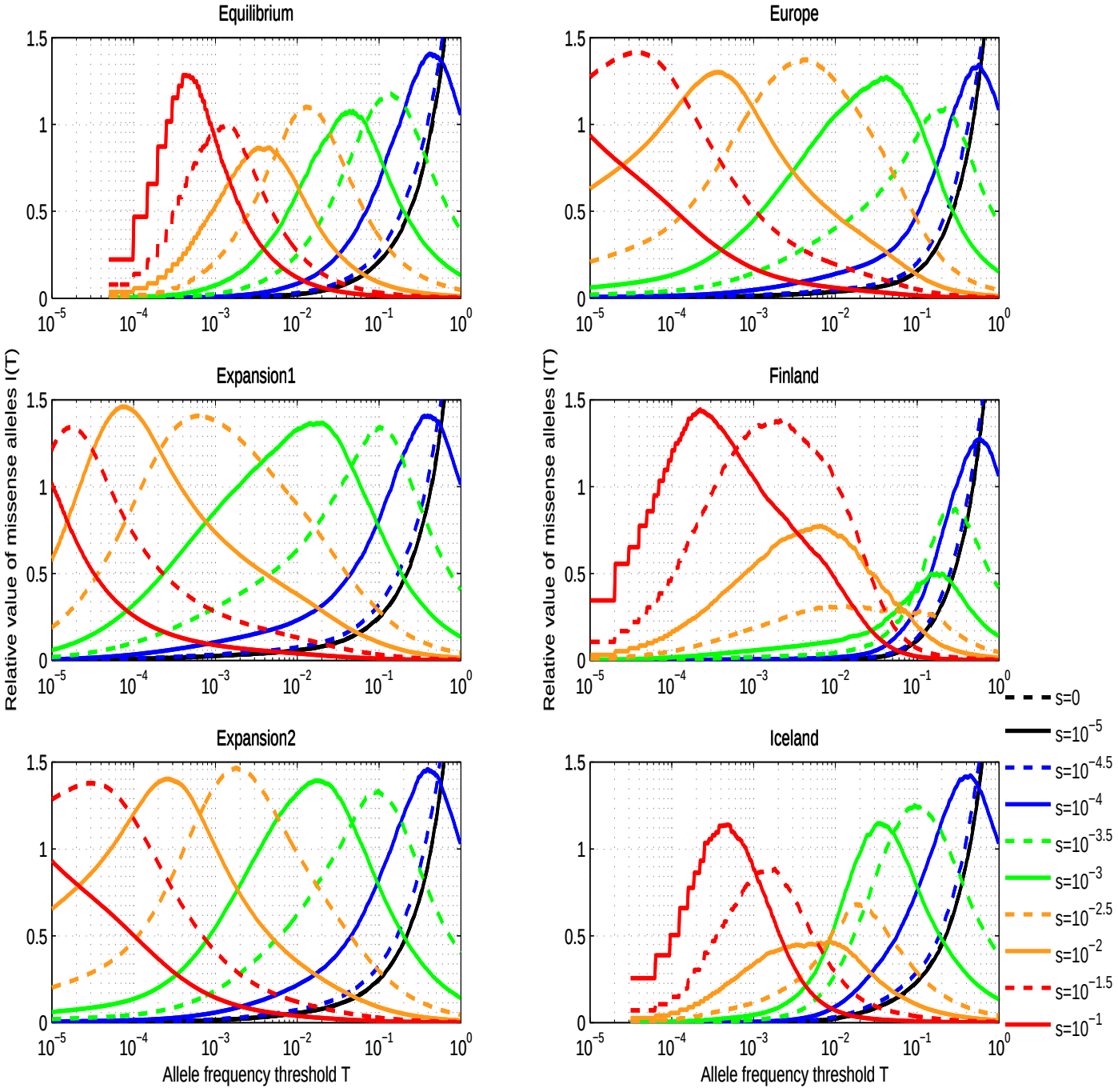

Figure 12: Similar to Figure S11, but for larger effect size ($1 + \lambda = 16$). For large effects, missense alleles are more beneficial, giving $\sim 1 - 1.5$-fold the power of disruptive alleles at the optimal threshold, thus using them, in combination with disruptive alleles, leads to a $\sim 2 - 2.5$-fold increase in power.

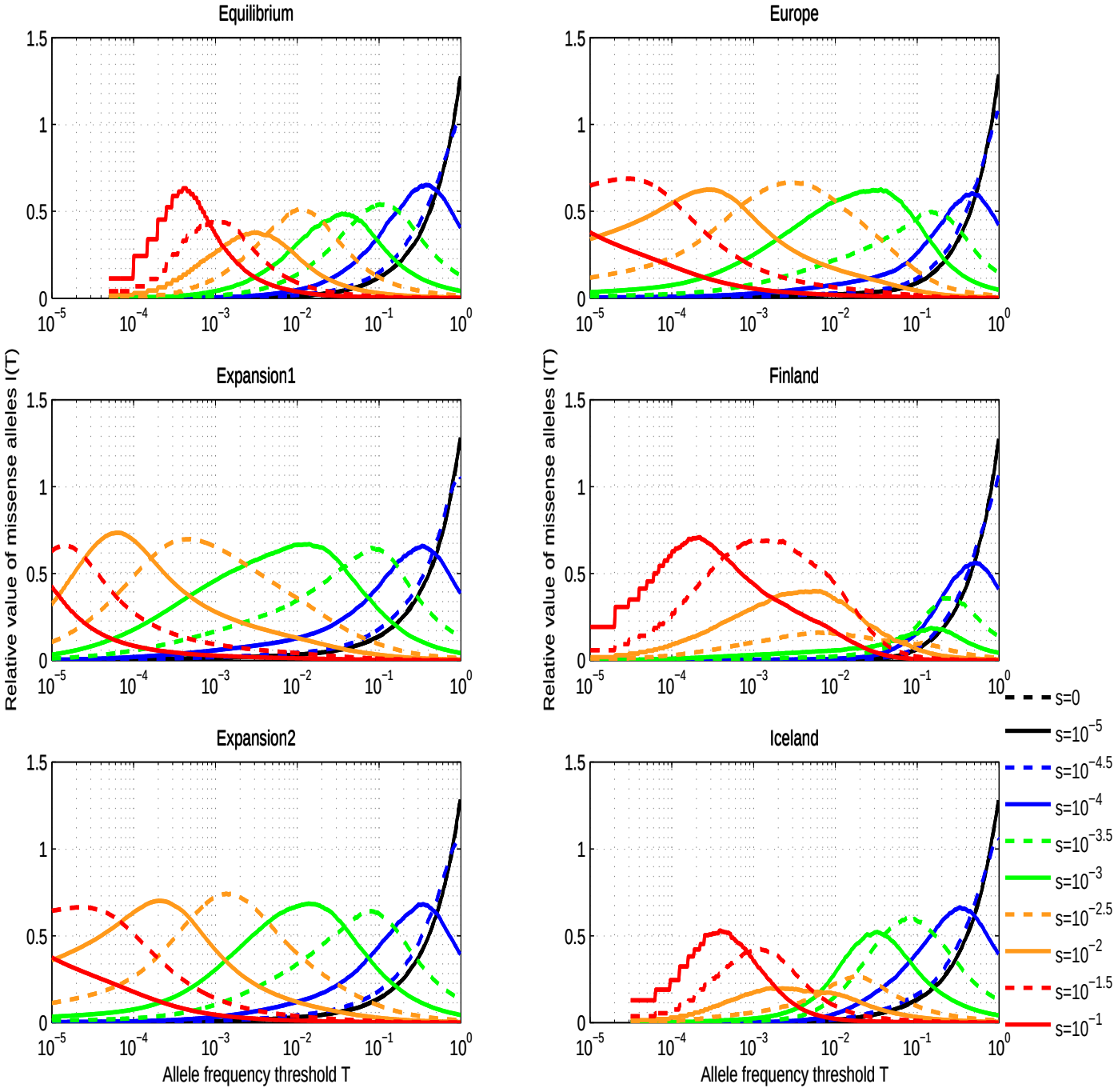

Figure 13: Similar to Figure S11, but for smaller effect size ($1+\lambda = 1.5$). For small effects, missense alleles are less beneficial, giving $\sim$0.5-fold the power of disruptive alleles at the optimal threshold, thus using them, in combination with disruptive alleles, leads only to a $\sim$1.5-fold increase in power.

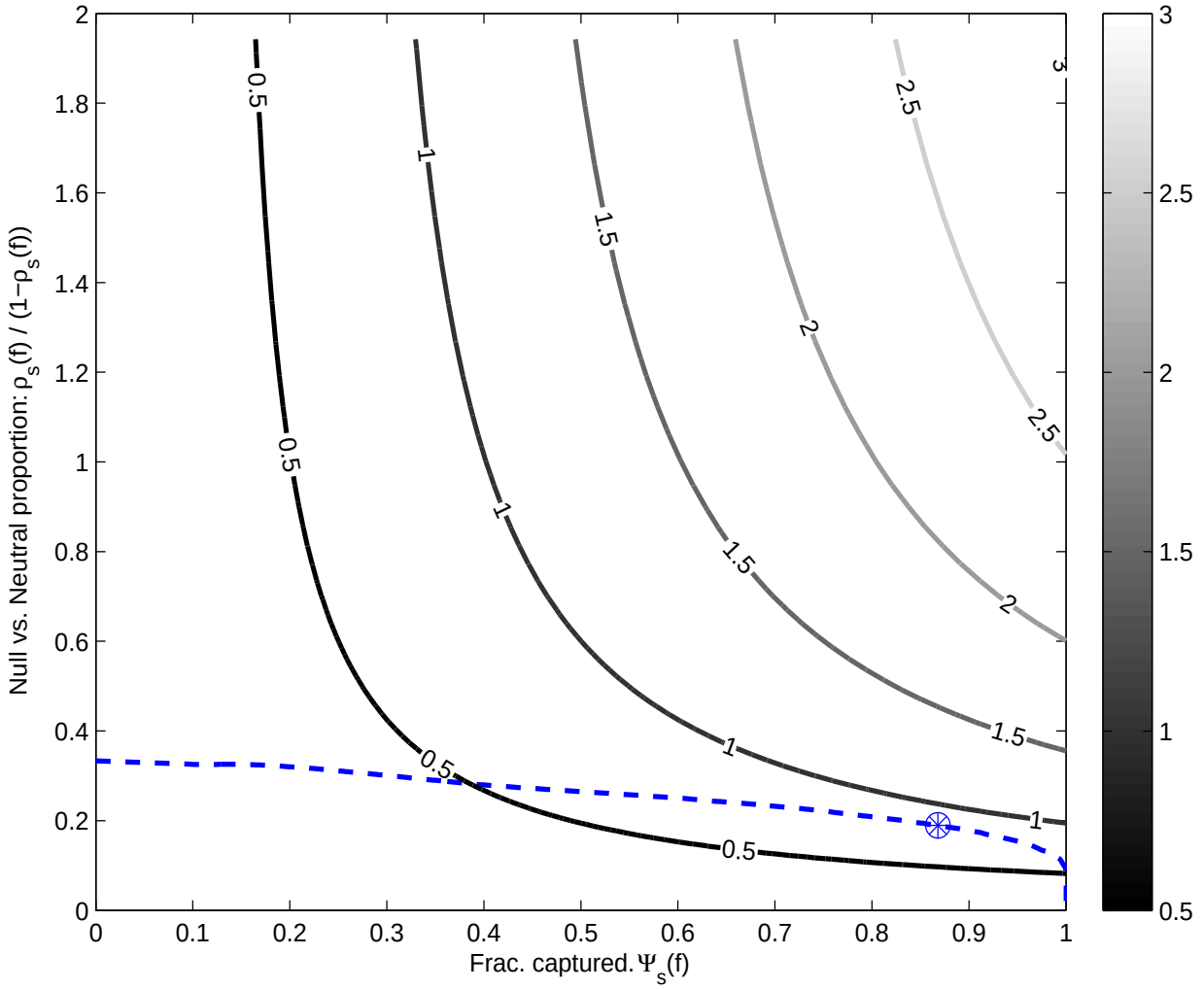

Figure 14: Relation between the fraction of null alleles which are included ($\Psi_s(T^*)$) to the fraction of null alleles among the set of included alleles ($\rho_s(T^*)$) as we vary the threshold T^* for the European population. The contour plots represent iso-power curves, where the value is the ratio of the contribution of missense alleles relative to disruptive alleles. The x-axis displays the fraction of null alleles captured, and the y-axis represent the proportion of null alleles among all missense alleles. Varying the threshold T corresponds to moving along the blue curve. The circled blue dot represents the values obtained for the optimal threshold. We have used $1 + \lambda = 4$, $\alpha = 0.25$, and $s = 0.01$. The behavior for different populations and different selection coefficients is not significantly different.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
$f_{1/2}$	0.51	0.49	0.43	0.24	0.071	0.021	0.006	0.0018	0.00065	0.00025
T^*	0.99	0.99	0.96	0.38	0.1	0.039	0.011	0.0032	0.0012	0.0004
$\rho_s(T^*)$	0.25	0.25	0.2	0.16	0.16	0.13	0.15	0.12	0.13	0.17
$\Psi_s(T^*)$	1	1	1	0.86	0.76	0.8	0.76	0.72	0.78	0.81
$\rho_s(1\%)$	0.25	0.25	0.25	0.25	0.23	0.2	0.16	0.054	0.019	0.0081
$\Psi_s(1\%)$	0.011	0.011	0.015	0.042	0.13	0.35	0.72	0.98	1	1
$\rho_s(0.1\%)$	0.25	0.25	0.25	0.25	0.24	0.24	0.24	0.18	0.13	0.081
$\Psi_s(0.1\%)$	0.001	0.001	0.0013	0.0039	0.013	0.039	0.11	0.33	0.72	0.99

Table 4: Optimal threshold for Equilibrium model. The tables show the optimal threshold T^*, compared to the median allele frequency $f_{1/2}$. In addition, we show the cumulative allele frequency Ψ_s and the proportion of null alleles ρ_s at the optimal threshold T^*, as well as the frequencies 0.1% and 1%. At the optimal threshold T^*, the cumulative frequency is in the range of $0.7-0.9$, and the proportion of null alleles is around $0.15-0.2$. Using a fixed threshold of 0.1% or 1% is sub-optimal, and may lead to severe dilution of null alleles (low ρ_s for strong selection.)

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
$f_{1/2}$	0.50	0.48	0.41	0.23	0.040	0.0011	9.4e-005	2.3e-005	6.5e-006	2e-006
T^*	0.99	0.99	0.96	0.33	0.08	0.014	0.00048	6.2e-005	1.4e-005	3.9e-006
$\rho_s(T^*)$	0.25	0.25	0.21	0.16	0.18	0.17	0.18	0.19	0.18	0.16
$\Psi_s(T^*)$	1	1	1	0.86	0.81	0.86	0.85	0.86	0.81	0.8
$\rho_s(1\%)$	0.25	0.25	0.25	0.24	0.23	0.18	0.078	0.028	0.0089	0.0026
$\Psi_s(1\%)$	0.03	0.03	0.038	0.12	0.33	0.8	1	1	1	1
$\rho_s(0.1\%)$	0.25	0.25	0.25	0.25	0.24	0.22	0.15	0.061	0.02	0.0059
$\Psi_s(0.1\%)$	0.013	0.013	0.017	0.053	0.16	0.46	0.94	1	1	1

Table 5: Optimal threshold for Expansion1 model. Similar to Table S4.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
$f_{1/2}$	0.51	0.48	0.42	0.23	0.040	0.0025	0.00035	4.9e-005	2.3e-006	5.0e-007
T^*	0.99	0.99	0.94	0.36	0.072	0.012	9.6e-005	7.6e-006	1.5e-006	5e-007
$\rho_s(T^*)$	0.25	0.25	0.21	0.16	0.18	0.18	0.18	0.17	0.16	0.15
$\Psi_s(T^*)$	1	1	1	0.87	0.8	0.86	0.84	0.86	0.82	0.86
$\rho_s(1\%)$	0.25	0.25	0.25	0.25	0.23	0.18	0.073	0.024	0.0075	0.0026
$\Psi_s(1\%)$	0.031	0.031	0.041	0.12	0.36	0.84	1	1	1	1
$\rho_s(0.1\%)$	0.25	0.25	0.25	0.25	0.24	0.21	0.12	0.042	0.013	0.0045
$\Psi_s(0.1\%)$	0.018	0.018	0.024	0.068	0.21	0.56	0.99	1	1	1

Table 6: Optimal threshold for Expansion2 model. Similar to Table S4.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
$f_{1/2}$	0.54	0.51	0.47	0.30	0.11	0.011	0.00055	4.5e-005	3.3e-006	5.0e-007
T^*	0.99	0.99	0.94	0.47	0.15	0.036	0.0025	0.00027	2.9e-005	8.4e-007
$\rho_s(T^*)$	0.25	0.25	0.21	0.15	0.15	0.18	0.18	0.17	0.17	0.18
$\Psi_s(T^*)$	1	1	0.98	0.86	0.73	0.79	0.82	0.83	0.9	0.84
$\rho_s(1\%)$	0.25	0.25	0.25	0.25	0.24	0.22	0.12	0.037	0.013	0.0039
$\Psi_s(1\%)$	0.019	0.019	0.024	0.071	0.21	0.57	0.99	1	1	1
$\rho_s(0.1\%)$	0.25	0.25	0.25	0.25	0.25	0.25	0.21	0.1	0.04	0.012
$\Psi_s(0.1\%)$	0.0064	0.0063	0.0081	0.024	0.073	0.23	0.62	0.99	1	1

Table 7: Optimal threshold for Europe model. Similar to Table S4.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
$f_{1/2}$	0.53	0.50	0.47	0.30	0.12	0.064	0.038	0.0023	5.1e-006	6.0e-007
T^*	0.99	0.99	0.99	0.5	0.23	0.15	0.0076	0.0053	0.0011	0.00021
$\rho_s(T^*)$	0.25	0.25	0.21	0.14	0.1	0.056	0.21	0.2	0.2	0.17
$\Psi_s(T^*)$	1	1	1	0.86	0.77	0.72	0.17	0.44	0.78	0.91
$\rho_s(1\%)$	0.25	0.26	0.26	0.25	0.22	0.2	0.19	0.16	0.1	0.036
$\Psi_s(1\%)$	0.0023	0.0025	0.003	0.0086	0.023	0.066	0.18	0.48	0.96	1
$\rho_s(0.1\%)$	0.25	0.26	0.25	0.25	0.25	0.24	0.25	0.25	0.2	0.097
$\Psi_s(0.1\%)$	0.00081	0.00083	0.001	0.0031	0.0095	0.029	0.093	0.27	0.76	1

Table 8: Optimal threshold for Finland model. Similar to Table S4.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
$f_{1/2}$	0.50	0.48	0.42	0.24	0.048	0.010	0.0054	0.0027	0.00015	1.2e-005
T^*	0.99	0.99	0.96	0.33	0.076	0.033	0.017	0.0016	0.0011	0.00038
$\rho_s(T^*)$	0.25	0.25	0.21	0.17	0.18	0.13	0.074	0.17	0.17	0.15
$\Psi_s(T^*)$	1	1	1	0.85	0.76	0.86	0.79	0.26	0.55	0.8
$\rho_s(1\%)$	0.25	0.25	0.26	0.25	0.24	0.19	0.095	0.048	0.022	0.0048
$\Psi_s(1\%)$	0.012	0.013	0.016	0.047	0.14	0.34	0.52	0.75	0.97	1
$\rho_s(0.1\%)$	0.25	0.25	0.25	0.25	0.24	0.23	0.2	0.2	0.18	0.08
$\Psi_s(0.1\%)$	0.00065	0.00065	0.00083	0.0026	0.008	0.024	0.066	0.2	0.51	0.97

Table 9: Optimal threshold for Iceland model. Similar to Table S4.

5.1 Heterogeneity of Alleles - Hypomorphs

Our basic two class model focuses only on neutral and null alleles. We ignore the possibility of hypomorphic alleles with effect size and selection coefficients lying between neutral and null alleles. We can refine the model to include a third class of hypomorphic alleles. This three-class model has

1. Null alleles with selection coefficient s and effect size $1+\lambda$ as before. with fraction α at birth.
2. Hypomorphic alleles with selection coefficient δs and effect size $1+\lambda\delta$ for some $0<\delta<1$ and fraction β at birth.
3. Neutral alleles with selection coefficient 0 and no effect, with frequency $1-\alpha-\beta$ at birth.

The above model assumes direct selection - that is, the proportional decrease in effect size for hypomorph alleles (compared to null alleles) is the same as the proportional decrease in the selective coefficient.

We can compute the effect of adding hypomorphs alleles on power. We consider two scenarios:

1. **Perfect knowledge.**
   In this idealized scenario, we assume that we can perfectly recognize whether a missense allele is null, hypormorphic or neutral. In this case, we can simply calculate the contribution to the likelihood ratio of hypomorphic alleles:

$$E[2\mathcal{LLR}_H] = 4n\frac{\beta\mu_M}{\delta s}g(\delta\lambda). \tag{5.6}$$

   The proportional contribution of missense alleles, relative to disruptive alleles, becomes:

$$\begin{aligned}\frac{E[2\mathcal{LLR}_M]}{E[2\mathcal{LLR}_D]} &= \frac{E[2\mathcal{LLR}_{M,null}+2\mathcal{LLR}_H]}{E[2\mathcal{LLR}_D]}\\ &= \frac{\alpha\mu_M}{\mu_D}\Big[1+\frac{\beta}{\alpha}\frac{g(\delta\lambda)}{\delta g(\lambda)}\Big].\end{aligned} \tag{5.7}$$

   Here, the first term in the brackets represents the contribution of null missense alleles and the second term represents hypomorphic missense alleles. For small values of λ, we have $g(\lambda)\sim\frac{\lambda^2}{2}$, and the additional relative contribution of hypomorphs is $\sim\frac{\beta}{\alpha}\delta$. For large effect size λ, we have $g(\lambda)\sim\lambda\log(\lambda)$ and the additional contribution of hypomorphs is $\sim\frac{\beta}{\alpha}[1+\frac{\log(\delta)}{\log(\lambda)}]$. In both cases, the relative additional contribution to power is insensitive to the selective coefficient (provided that the assumption of strong selection holds).

   Table S10 below shows the relative contribution of hypomorphic alleles in this idealized case. We considered our typical gene with $\alpha=0.25$ and $\frac{\mu_M}{\mu_D}=7.3$. We assumed that hypomorphic missense alleles are born at twice the rate of null missense alleles - that is, $\beta=0.5$. [15].

We suppose that null alleles increase disease risk by either 3-fold or 11-fold ($\lambda = 2$ or 10) and that hypomorphic alleles decrease the effect size and selection coefficient by either 5-fold or 10-fold ($\delta = 0.2$ or 0.1). (For example, with $\lambda = 10$ and $\delta = 0.1$, hypomorphic alleles increase disease risk by $1 + \delta\lambda = 2$-fold.)

		Reduction in sample size by supplementing disruptive alleles with:		
λ	δ	Null missense only	Null and hypomorphic missense	Improvement
10	0.2	2.83-fold	4.27-fold	51.1%
10	0.1	2.83-fold	3.69-fold	30.5%
2	0.2	2.83-fold	3.83-fold	35.4%
2	0.1	2.83-fold	3.35-fold	18.7%

Table 10: The relative contribution of missense null and hypomorph alleles for different values of λ and δ under perfect knowledge. The relative contribution is shown here for the European population.

With perfect knowledge, hypomorphic alleles can contribute usefully when the effect size is large.

2. **Imperfect knowledge.**
In this realistic scenario, we assume that we cannot distinguish whether a missense allele is null, hypormorphic or neutral. Instead, we must select a single optimal frequency threshold T* and count only those alleles with frequency $\leq T^*$. One challenge is that hypomorphic and null missense alleles have different selection coefficients and hence different allele frequencies, and thus a single threshold will be sub-optimal in terms of capturing these two groups of alleles. In addition, the two classes have different effect sizes, and thus the observed effect size will represent an average value (overweighted toward hypomorphic alleles, because they are more frequent due to a higher birth rate and weaker selection), leading to further loss of power.

We computed the total effect of hypomorphic alleles for the same parameters as above, for the European population, and for different selection coefficients. Figure S15 shows the results for various threshold values T, for intermediate effect size ($1 + \lambda = 4$). Table S11 shows the maximal contribution, achieved when using the optimal threshold T^*. For the reasons discussed above, the proportional contribution of hypomorphic alleles to power is much smaller than in the idealized case. In the examples shown, they typically reduce the required sample size by less than 10%.

		Reduction in sample size by supplementing disruptive alleles with:		
λ	δ	Null missense only	Null and hypomorphic missense	Improvement
10	0.2	$1.44 - 1.59$-fold	$1.58 - 1.79$-fold	$7.1 - 13.5\%$
10	0.1	$1.44 - 1.59$-fold	$1.48 - 1.63$-fold	$2.1 - 3.9\%$
2	0.2	$1.25 - 1.35$-fold	$1.31 - 1.44$-fold	$4.0 - 7.3\%$
2	0.1	$1.25 - 1.35$-fold	$1.26 - 1.37$-fold	$1.1 - 1.7\%$

Table 11: The relative contribution of missense null and hypomorph alleles for different values of λ and δ, in the situation of imperfect knowledge, where we choose the threshold T^* maximizing power. The relative contribution is shown here for the European population. The relative contribution depends on the selection coefficient s (see Figure S15). For each value of λ and δ, we display the range of relative contributions, when varying s in the range $10^{-4} - 10^{-1}$. The relative contributions of hypomorph alleles are tiny, unless the effect size λ and the proportionality constant δ are large.

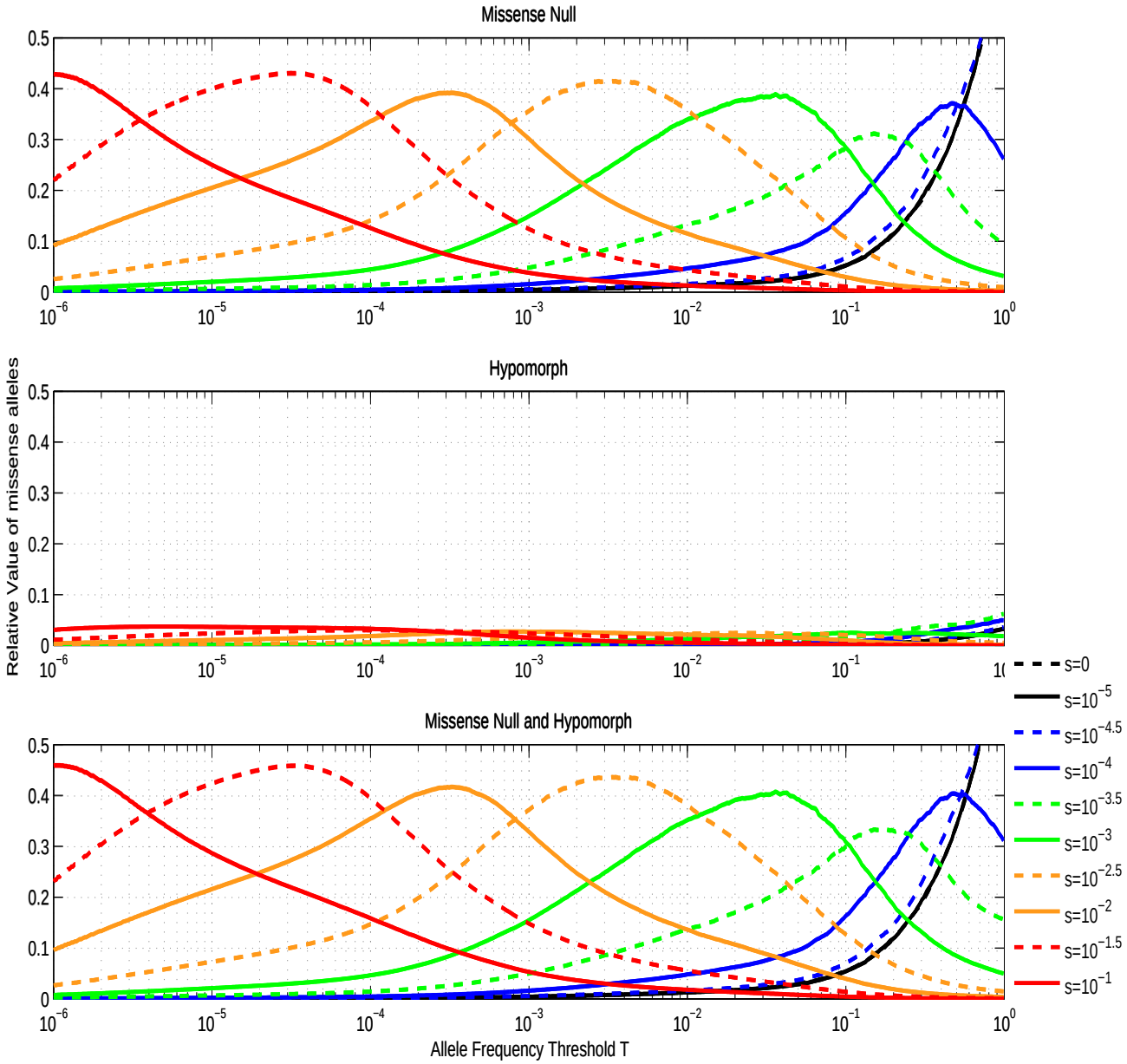

Figure 15: Increase in power from using both null and hypomorph missense alleles. The figures show the contribution of missense alleles to power, compared to the contribution of disruptive alleles, as function of the threshold T for the European population, and with the parameters, $1 + \lambda = 4, \alpha = 0.25, \beta = 0.5, \delta = 0.1$. The top plot shows the contribution of only null missense alleles. The middle plot shows the contribution of hypomorph missense alleles. The bottom plot shows the contribution of all missense alleles. Overall, the relative additional contribution of hypomorph alleles to power is $< 10\%$ of the contribution of missense null alleles.

6 RVAS Strategy 3: Enriching for Null Missense Alleles

One might seek to enrich for null missense alleles by using computational programs, such as PolyPhen-2 [46], SIFT [47], and MutationTaster [48], that predict which missense changes are "damaging" - based on the biochemical nature of the substitution and evolutionary conservation of the site (but not allele frequency). A common approach is to use such programs to filter the alleles and analyze only the damaging alleles.

We can summarize the accuracy of such programs with two terms: (i) the false positives rate $\gamma_{neutral}$ (the probability that a neutral missnese allele will be incorrectly classified as null allele), and (ii) the true positives rate γ_{null} (the probability that a null missnese allele will be correctly classified as a null allele). We next calculate the gain in power as a function of the parameters $\gamma_{neutral}$ and γ_{null} ,where we combine the strategy from the previous section (of enriching further for null alleles by using only alleles below a frequency threshold) with the strategy of using functional predictions.

Considering only alleles with frequency below threshold T, we define $\Psi_s^{(\gamma_{neutral},\gamma_{null})}(T)$ to be the cumulative allele frequency of null alleles predicted to be damaging, and $\rho_s^{(\gamma_{neutral},\gamma_{null})}(T)$ to be the fraction of null alleles among all damaging alleles. The values are given by

$$\begin{aligned}\rho_s^{(\gamma_{neutral},\gamma_{null})}(T) &= \frac{\rho_s(T)\gamma_{null}}{\rho_s(T)\gamma_{null} + (1-\rho_s(T))\gamma_{neutral}} \\ \Psi_s^{(\gamma_{neutral},\gamma_{null})}(T) &= \gamma_{null}\Psi_s(T).\end{aligned} \tag{6.1}$$

To compute power, we can simply plug these values in eq. (5.3), instead of $\rho_s(T^*)$ and $\Psi_s(T^*)$, giving:

$$\frac{E[\mathcal{LLR}_{missense}^{(\gamma_{neutral},\gamma_{null})}(T^*)]}{E[\mathcal{LLR}_D]} \sim \frac{\mu_M \alpha \Psi_s^{(\gamma_{neutral},\gamma_{null})}(T^*) g(\rho_s^{(\gamma_{neutral},\gamma_{null})}(T^*)\lambda)}{\mu_D \rho_s^{(\gamma_{neutral},\gamma_{null})}(T^*) g(\lambda)}. \tag{6.2}$$

We can then search for optimal threshold maximizing power in the above equation, which we denote as $T^*(\gamma_{neutral}, \gamma_{null})$. This threshold is different from the optimal threshold T^* for the case in which we lack functional prediction (see previous section), and will typically be higher, allowing us to capture more missense null alleles, while still maintaining them at high proportion.

There is a simple formula relating the proportion of null alleles ρ_s, without functional predictions to the proportion of null alleles $\rho_s^{(\gamma_{neutral},\gamma_{null})}(T)$ with functional prediction,

$$\frac{\rho_s^{(\gamma_{neutral},\gamma_{null})}(T)}{1-\rho_s^{(\gamma_{neutral},\gamma_{null})}(T)} = \frac{\gamma_{null}}{\gamma_{neutral}} \frac{\rho_s(T)}{1-\rho_s(T)}. \tag{6.3}$$

In Figure S16 we show the advantage of using functional prediction.

If we have perfect prediction, ($\gamma_{neutral} = 0, \gamma_{null} = 1$), we get $\rho_s^{(0,1)}(T) = 1$, $\forall T \in [0,1]$, and the above ratio in eq. (6.3) diverges to infinity. In this case, the optimal threshold will be $T^* = 1$, taking all null missense alleles, and simply get (using $\Psi_s(1) = 1$),

$$\frac{E[\mathcal{LLR}^{(0,1)}_{missense}(1)]}{E[\mathcal{LLR}_D]} \sim \frac{\mu_M \alpha}{\mu_D}. \tag{6.4}$$

For a typical gene with $\frac{\mu_M}{\mu_D} = 7.3$ and $\alpha = 0.25$, we get that missense contribute roughly 1.8-fold as much information as disruptives, thus in this ideal case we can decrease the sample size by roughly 2.8-fold.

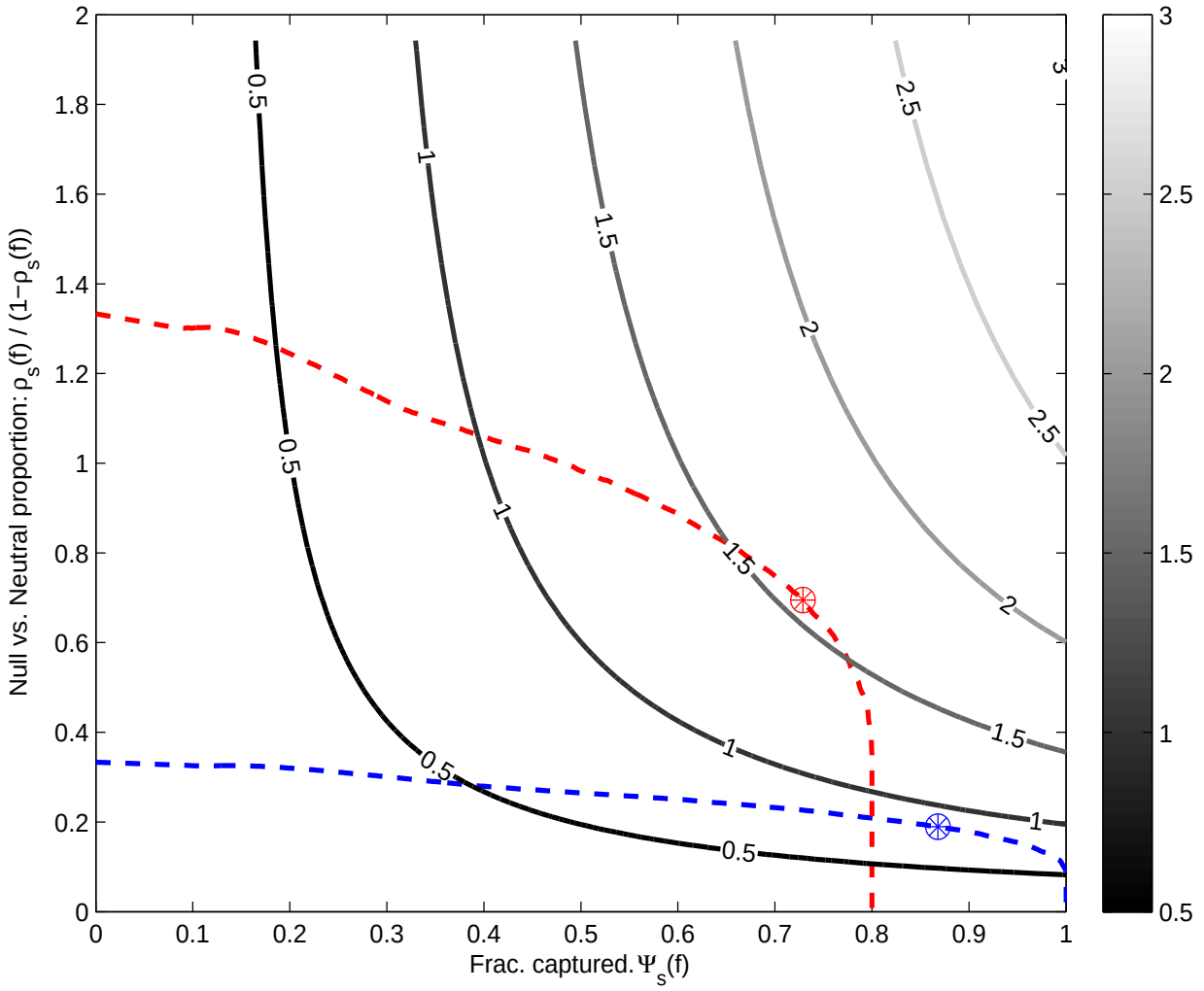

Figure 16: Power-gain from functional prediction of missense null alleles for the equilibrium population. As in Figure S14, the contour plots represent iso-power curves. The x-axis displays the fraction of null alleles captured, and the y-axis represent the proportion of null alleles among all missense alleles. Varying the threshold T corresponds to moving along the colored curves. The blue curve shows results without functional prediction (strategy 2) and the red curve shows results with a functional prediction with $\gamma_{null} = 0.8, \gamma_{neutral} = 0.2$. The circled dots on each curve represent the values obtained for the optimal threshold. When using functional prediction, we see a shift in the feasible curve, and thus an almost 2-fold increase in relative contributions, from 0.85 to 1.56. We have used $1 + \lambda = 4$, $\alpha = 0.25$, and $s = 0.01$.

7 Effect of Thresholding in a Finite Sample (with *LDLR* Example)

The burden analyses in Sections 5 and 6 assume an idealized case, in which we know the precise population frequency of each allele based on a large population survey. In this case, we can filter alleles based on their frequency, and enrich for null alleles, with $\rho(T)$ of missense alleles below a threshold T being null.

Here, we consider the situation where we don't know the allele frequencies based on a large population survey, but have only the sample data from a case-control or case-unaffected study with a total on n individuals. In this case, we can filter alleles based on their frequency in the sample.

Given a threshold T, we can retain alleles that have at most $\lfloor 2nT \rfloor$ occurrences in the sample. We denote $\rho_s(T;n)$ to be the expected fraction of nulls when we consider only alleles with frequency below T in a *sample* of n individuals. We have:

$$\rho_s(T;n) = \frac{\int_{x=0}^{1} \alpha f_s \psi_s(x) B(\lfloor 2nT \rfloor, 2n, x) dx}{\int_{x=0}^{1} [\alpha f_s \psi_s(x) + (1-\alpha) f_0 \psi_0(x)] B(\lfloor 2nT \rfloor, 2n, x) dx} \tag{7.1}$$

where we denote by $B(k,n,p)$ the cumulative binomial distribution (minus the value at zero):

$$B(k,n,p) = \sum_{i=1}^{k} \binom{2n}{i} p^i (1-p)^{n-i}. \tag{7.2}$$

For example, if we take only singleton alleles (i.e. setting $T = \frac{1}{2n}$) we get,

$$\rho_s(\frac{1}{2n};n) = \frac{\int_{x=0}^{1} \alpha f_s \psi_s(x) x(1-x)^{2n-1} dx}{\int_{x=0}^{1} [\alpha f_s \psi_s(x) + (1-\alpha) f_0 \psi_0(x)] x(1-x)^{2n-1} dx} \tag{7.3}$$

When we set the frequency threshold based on cases, we need to be more careful with the choice of threshold, since the frequency in cases is higher, $(1+\lambda)f$ for null alleles with population frequency f. If we set the same threshold for cases and population controls, i.e. take all alleles with frequency at least T in either cases or controls, (assuming equal number of cases and controls), then

$$\begin{aligned}
\rho_s(T;n,\lambda) \approx & \int_{x=0}^{1} \alpha f_s \psi_s(x) \Big[B\big(\lfloor 2nT \rfloor, 2n, x\big) + B\big(\lfloor 2nT \rfloor, 2n, x(1+\lambda)\big) \\
& - B\big(\lfloor 2nT \rfloor, 2n, x\big) B\big(\lfloor 2nT \rfloor, 2n, x(1+\lambda)\big) \Big] dx \\
& \Big/ \int_{x=0}^{1} \Big\{ \alpha f_s \psi_s(x) \Big[B\big(\lfloor 2nT \rfloor, 2n, x\big) + B\big(\lfloor 2nT \rfloor, 2n, x(1+\lambda)\big) \\
& - B\big(\lfloor 2nT \rfloor, 2n, x\big) B\big(\lfloor 2nT \rfloor, 2n, x(1+\lambda)\big) \Big] \\
& + (1-\alpha) f_0 \psi_0(x) \Big[2B\big(\lfloor 2nT \rfloor, 2n, x\big) - B\big(\lfloor 2nT \rfloor, 2n, x\big)^2 \Big] \Big\} dx.
\end{aligned} \tag{7.4}$$

The fraction of nulls, ρ_s, thus depends on λ; large values of λ increase ρ_s.

7.1 *LDLR* Example

To illustrate the results in the paper (including the issues with finite samples), we use an example taken from an exome-wide RVAS of Early-Onset Myocardial Infarction (EOMI Study).

Briefly, several of the authors of this paper are involved in a large-scale exome-wide RVAS for EOMI, including analysis of 2743 cases and 2465 controls. The study subjects were taken from the National Heart Lung and Blood Institute's Exome Sequencing Project (NHLBI ESP) EOMI Study (1027 cases, 946 controls) and the Italian Atherosclerosis, Thrombosis and Vascular Biology (ATVB) study (1716 cases, 1519 controls) funded by the National Human Genome Research Institute. In the NHLBI ESP study, cases were individuals with MI at age ≤ 50 for males and age ≤ 60 for females and controls were individuals of advanced age (≥ 60 for men and ≥ 70 for women) who had not suffered Myocardial Infarction (MI). The Italian ATVB Study ascertained 1716 survivors of acute MI before the age of 45 from 125 intensive care units across Italy between $1998 - 2001$. Controls were individuals without a personal history of thrombotic disease and were matched to cases for age, gender, and geographic location. The EOMI study employed the commonly used definition of rare variants as being alleles with frequency $\leq 1\%$ in the general population.

The full results from the EOMI Study will be published elsewhere. Here, we focus only on the results for the *LDLR* gene (coding region of 2583 base pairs, including the stop codon).

The results (Table S12) show strong enrichment for disruptive alleles ($1 + \lambda_D = 18.1$) but much weaker enrichment for missense alleles ($1 + \lambda_M = 1.5$). The low enrichment seen for missense alleles is explained under our two-class model by the fact that null alleles in *LDLR* are under strong selection ($s \approx 10^{-1.7}$). Given such strong selection, the optimal threshold is very low: $T^* = 10^{-4}$. With the much higher threshold of $T = 1\%$, missense null alleles are swamped by missense neutral alleles. The relative risk for missense alleles with frequency $\leq 1\%$ is expected to be $1 + \lambda_M = 1.3$, which is very close to the observed value of 1.5. For the optimal threshold of $T^* = 10^{-4}$, the relative risk would be expected to be higher: $1 + \lambda_M = 3.8$ (although this is still lower than seen with disruptive alleles, which is a pure class of null alleles).

There is a practical problem, however, with using the optimal threshold $T^* = 10^{-4}$. Given the relatively small sample size (and no additional population survey data), we can't tell which alleles actually have population frequency $\leq 10^{-4}$. The best alternative is to study only the singleton alleles - that is, the alleles that appear exactly once in the sample.

For the EOMI study, counting singletons turns out to be almost as good as if we could identify precisely the alleles with frequency $\leq 10^{-4}$. (The expected value of $1 + \lambda_M$ is 3.7 in the former case vs 3.8 in the latter; Table S12). This is because the EOMI study just happens to involve $\sim 10^4$ chromosomes (more precisely, $2 \times (2743 + 2465) = 10,420$) - that is, singletons correspond to a frequency of $\sim 10^{-4}$ in the sample.

If the EOMI study involved a much smaller sample, the use of singletons would be less effective at enriching for null alleles. With only 1000 chromosomes, the expected relative risk is 2.1. With only 100 chromosomes, the expected relative risk is only 1.4. (Table S12). Figure S17 shows how the proportion of nulls among missense alleles varies with sample size.

Allelic-Class	Cases	Controls	$\Psi_s(T)$	$\rho_s(T)$	$1+\lambda$
n	2743	2465			
Disruptive:					
Observed	20	1			18.1
Expected					18.1
Missense($f \leq 1\%$):					
Observed	172	102			1.5
Expected (population)			100.0%	2.0%	1.3
Expected ($n = 5208$)			100.0%	2.0%	1.3
Missense($f \leq 0.01\%$):					
Observed	37	19			1.8
Expected (population)			90.9%	16.6%	3.8
Expected ($n = 5208$, singletons)			86.8%	16.0%	3.7
Expected ($n = 500$, singletons)			100.0%	6.5%	2.1
Expected ($n = 50$, singletons)			100.0%	2.3%	1.4
Missense-polyphen($f \leq 1\%$):					
Observed	89	32			2.5
Expected (population)			100.0%	7.7%	2.3
Expected ($n = 5208$, singletons)			100.0%	7.7%	2.3
Missense-polyphen-Singletons($f \leq 0.01\%$):					
Observed	19	6			2.8
Expected (population)			90.9%	44.4%	8.6
Expected ($n = 5208$, singletons)			86.8%	39.0%	7.7

Table 12: Values observed in an association test for *LDLR* with EOMI. The table shows observed counts and expected values of $\rho_s(T)$ and $1+\lambda$. 'Expected (population)' refers to the situation when allele frequencies are known precisely. 'Expected ($n =$.., singletons)' refers to the situation when the sample size is n and only singletons are counted. Calculations assume that the true relative risk for null alleles is $1+\lambda = 18.1$ (based on the observed value for disruptive alleles), the selection coefficient is $s = 10^{-1.7}$ (based on the observed frequency of disruptive alleles), $\alpha = 0.25$, the population follows the European model, and the functional prediction model misclassifies 20% of null and 20% of neutral alleles.

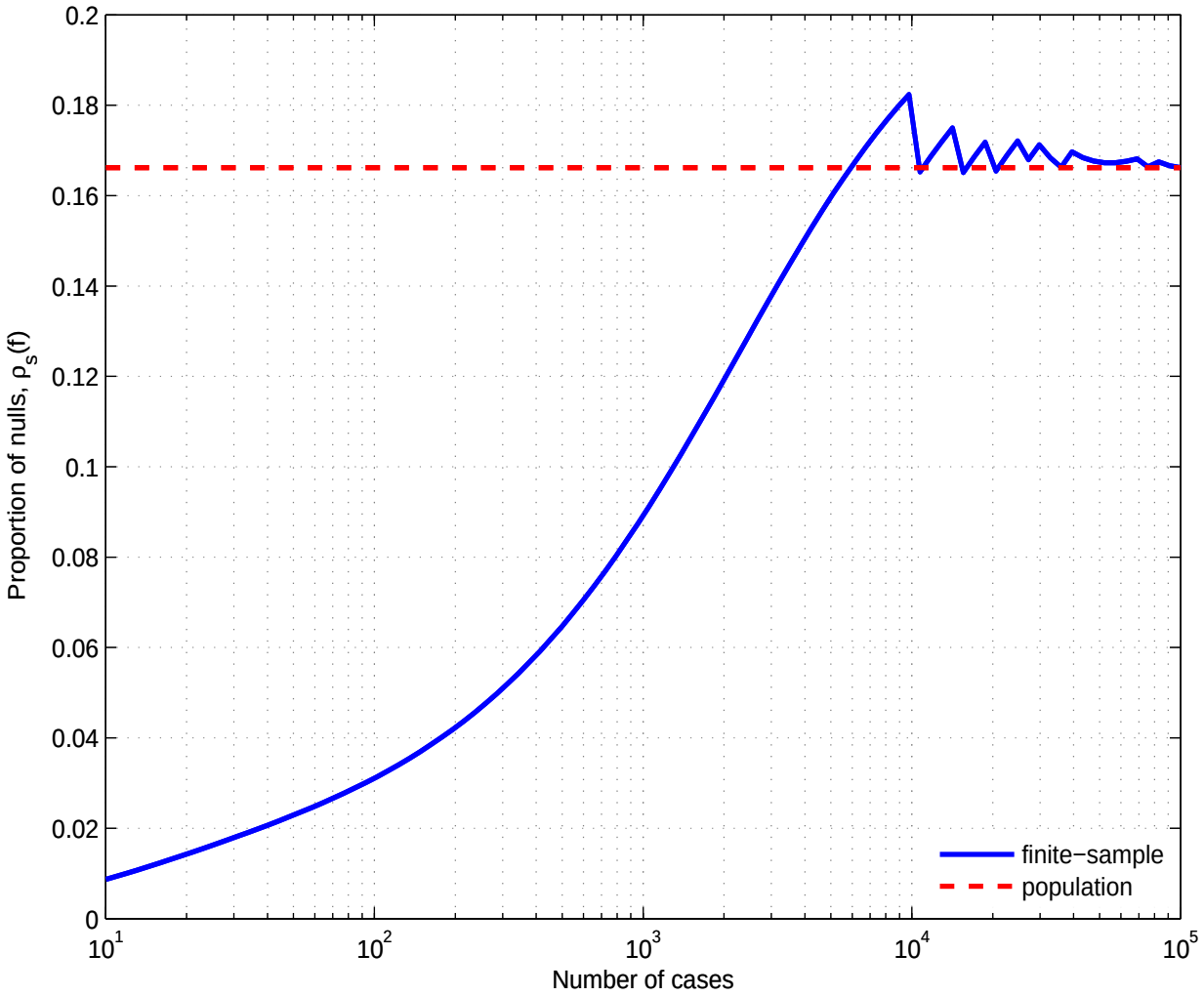

Figure 17: Proportion $\rho_s(f)$ of missense alleles that are null, as a function of the number n of cases, for the EOMI study with assumptions as in Table S12. The red dashed line shows the proportion of missense alleles that are null when allele frequencies are known precisely and one uses the optimal threshold $T^* \sim 10^{-4}$. The blue curve shows the proportion of missense alleles that are null if allele frequencies are not known precisely and one instead uses (i) singletons, if $J = 0$ and (ii) only alleles with J counts, if $J > 0$, where $J = \lfloor 2nT^* \rfloor$. The proportion of nulls reaches the dashed red curve when the number of cases is ~ 5000, corresponding to $10,000$ chromosomes. For smaller sample sizes, the proportion of missense alleles that are null is much smaller (necessitating a much larger sample size).

8 RVAS Strategy 4: Isolated Populations

Another potential approach is to exploit the inherent variation in combined allele frequencies across different populations and genes, with the hope that, for certain genes associated with disease, the CAF will happen to be large in the population tested, making it possible to detect association in a smaller sample. To study this variation, we computed the CAF for various populations using simulations, as described in Section 1.3. Specifically, we recorded the CAF for each gene in each individual simulation, to obtain an overall distribution of the CAF for different populations, shown in Figure 3 in the main text for all null alleles ($\mu_{null} = 5 \times 10^{-6}$).

Here, we describe results for both null alleles ($\mu_{null} = 5 \times 10^{-6}$) and disruptive alleles ($\mu_D = 1.7 \times 10^{-6}$). For large populations without bottlenecks, the distribution is in general tightly concentrated around the mean value, as shown in Tables S15-S18. For such populations we can thus largely ignore variations in the CAF, and simply use the mean value (Table S2). However, for populations that have experienced recent severe bottlenecks, the variation in CAF is much larger, as shown in Tables S21- S24 and in Figures S18-S22.

Figures S18-S22 show results for (i) all null alleles (as shown in Figure 3 in the main text) and (ii) disruptive alleles. Models with severe and recent bottlenecks (Finland and Iceland) show substantially higher variation in the CAF, compared to the Expansion1 model with a large population and no severe bottleneck, whereas the behaviour for Europe is intermediate.

The distributions depend on the bottleneck size, number of generations since expansion, mutation rate of the allelic class and selection coefficient. With a tighter bottleneck (e.g., 100 chromosomes in Finland vs. 1000 chromosomes in Iceland), the probability that a rare variant passes through the bottleneck is lower but those rare alleles that are lucky enough to pass jump to a higher frequency ($1\% vs 0.1\%$). Consequently, the proportion of genes at intermediate deviations (e.g., 5- to 10-fold) is often higher in Iceland, but the proportion of genes at high deviations (e.g., 30-fold or more) is typically higher in Finland. The curves also depend on the absolute mutation rate, and thus differ for all null alleles vs. disruptive alleles.

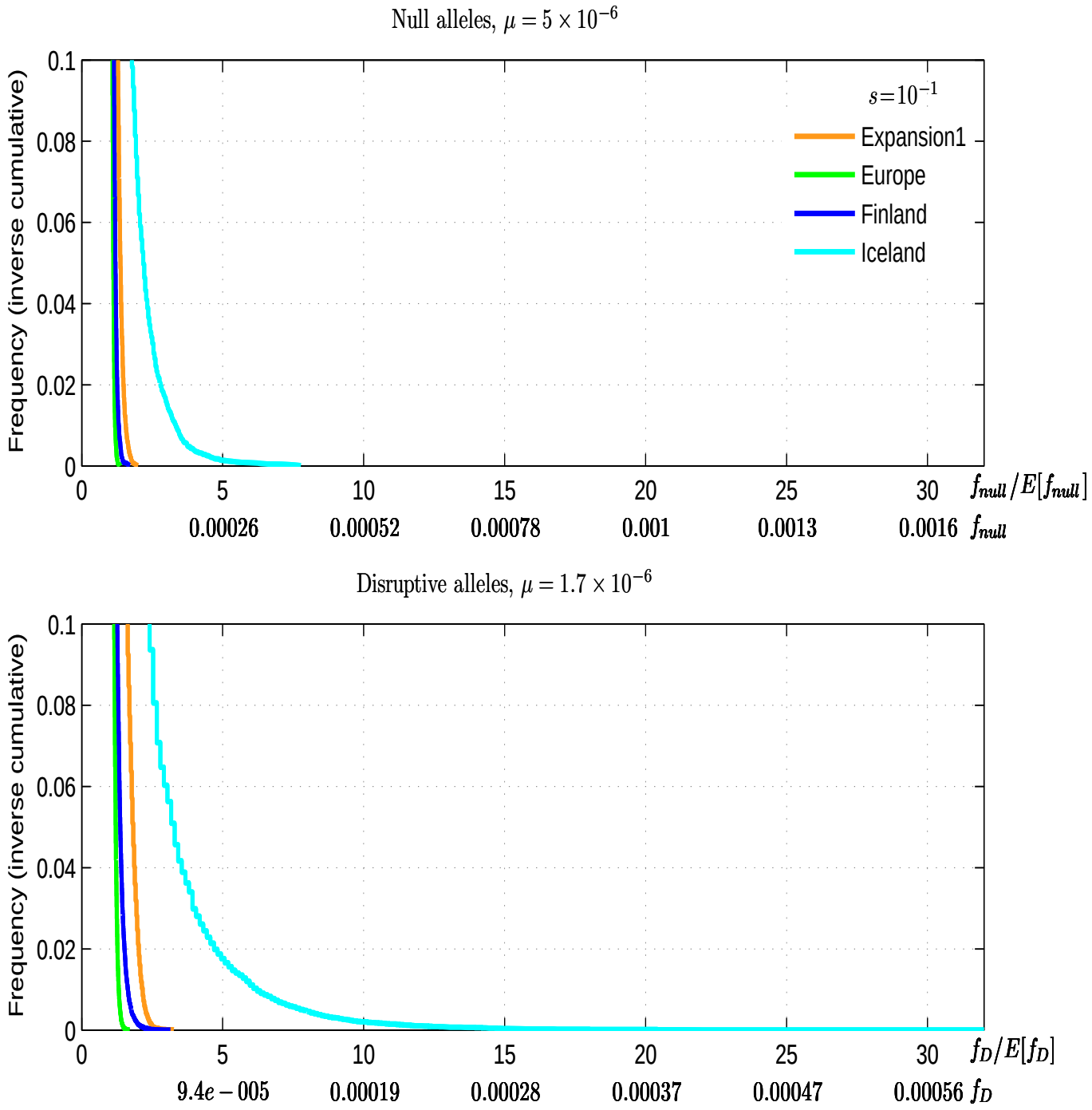

Figure 18: The inverse cumulative distribution of the Combined Allele Frequency (CAF) for $s = 10^{-1}$. Obtained by simulating 50,000 genes for each demographic model. The top panel represents null alleles ($\mu_{null} = 5 \times 10^{-6}$), as in Figure 3 in the main text, and the bottom panel represents disruptive alleles ($\mu_D = 1.7 \times 10^{-6}$). The x-axis shows both the CAF itself (f_{null}, f_D), and the CAF normalized to have mean 1 ($f_{null}/E[f_{null}]$, $f_D/E[f_D]$). The figures show the right tail of the distributions. While the mean CAF is essentially the same for all models (see Figure 1), different models show different distributions.

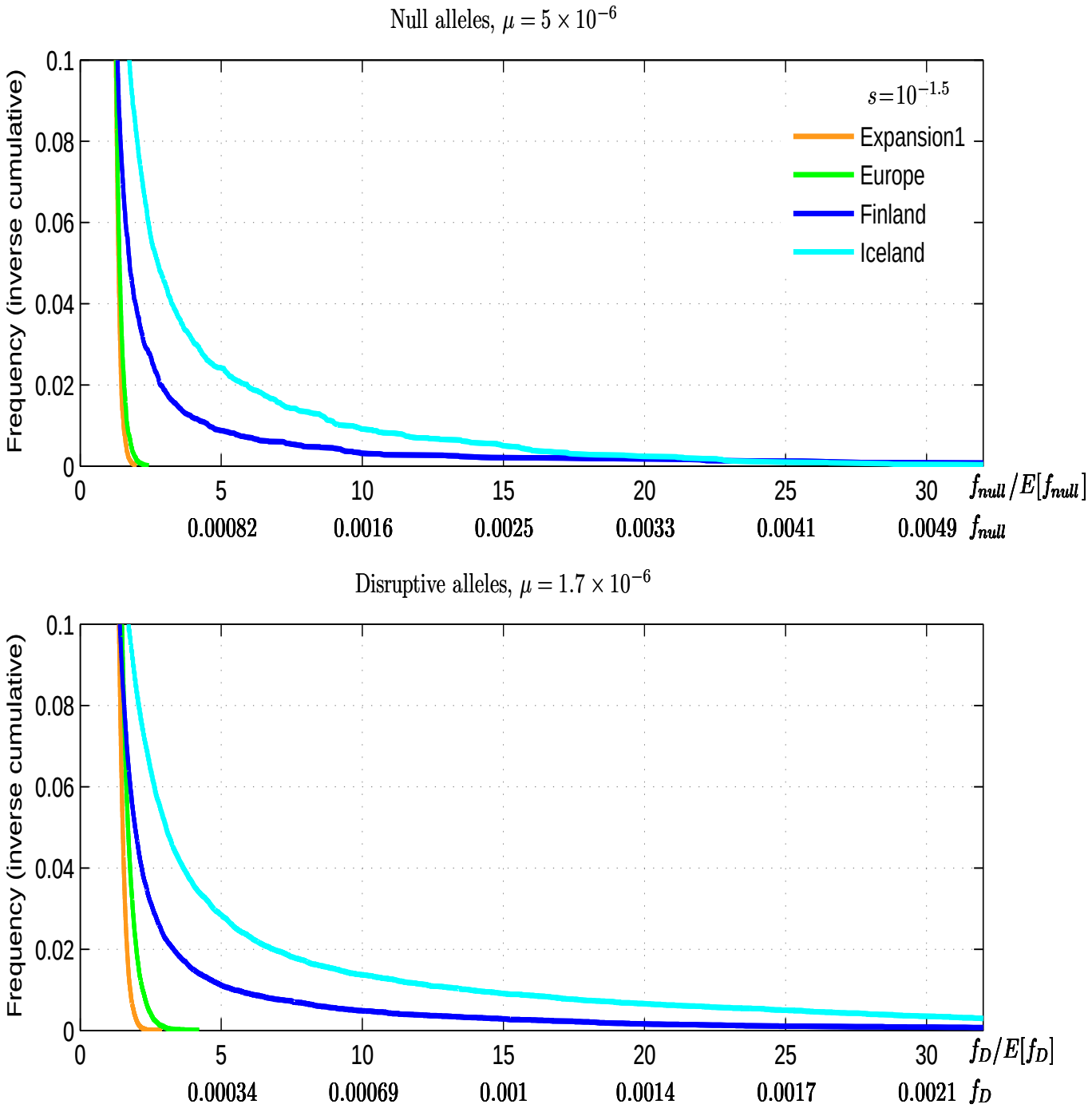

Figure 19: The inverse cumulative distribution of the Combined Allele Frequency (CAF) for $s = 10^{-1.5}$. Similar to Figure S18.

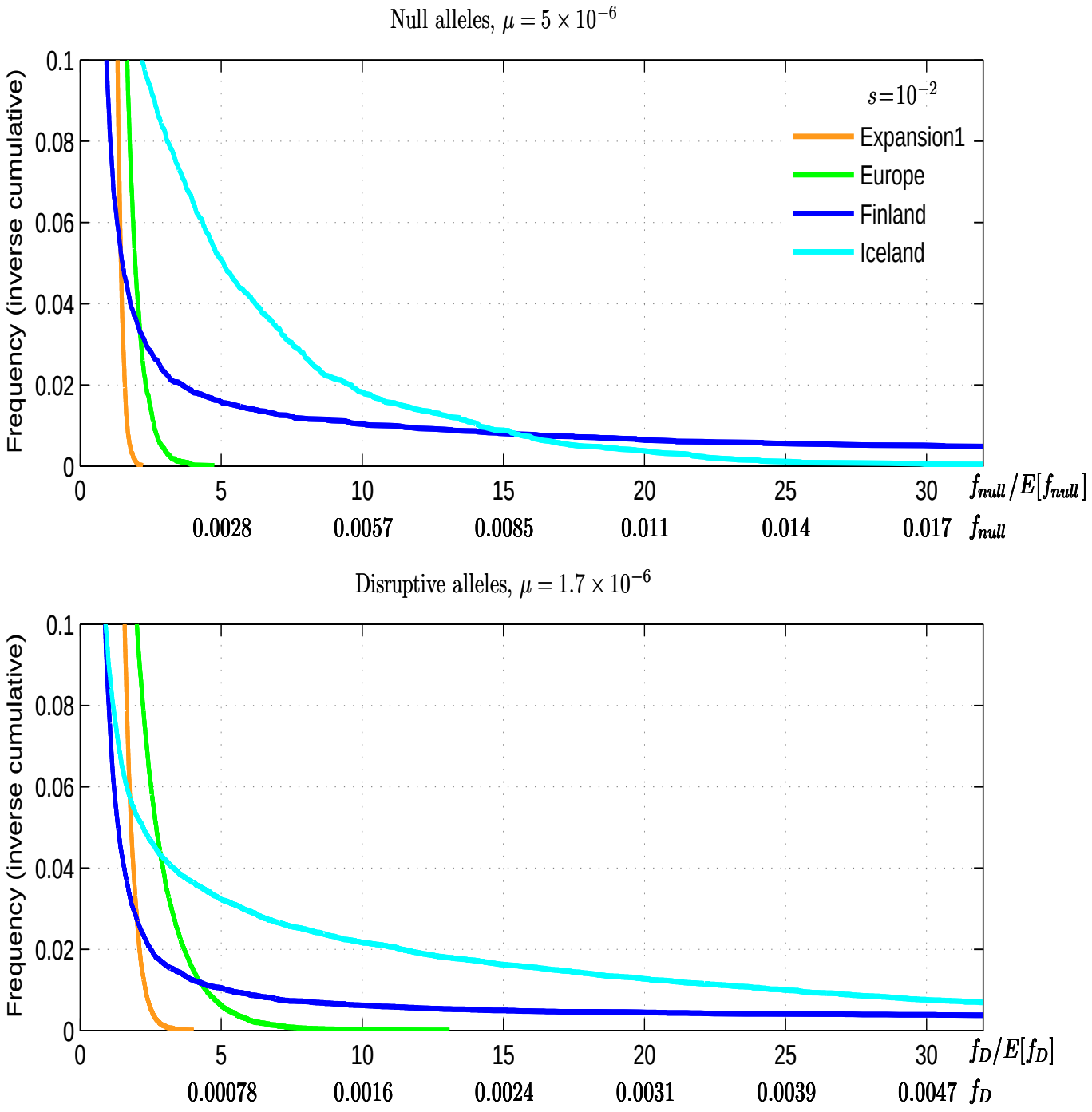

Figure 20: The inverse cumulative distribution of the Combined Allele Frequency (CAF) for $s = 10^{-2}$. Similar to Figure S18.

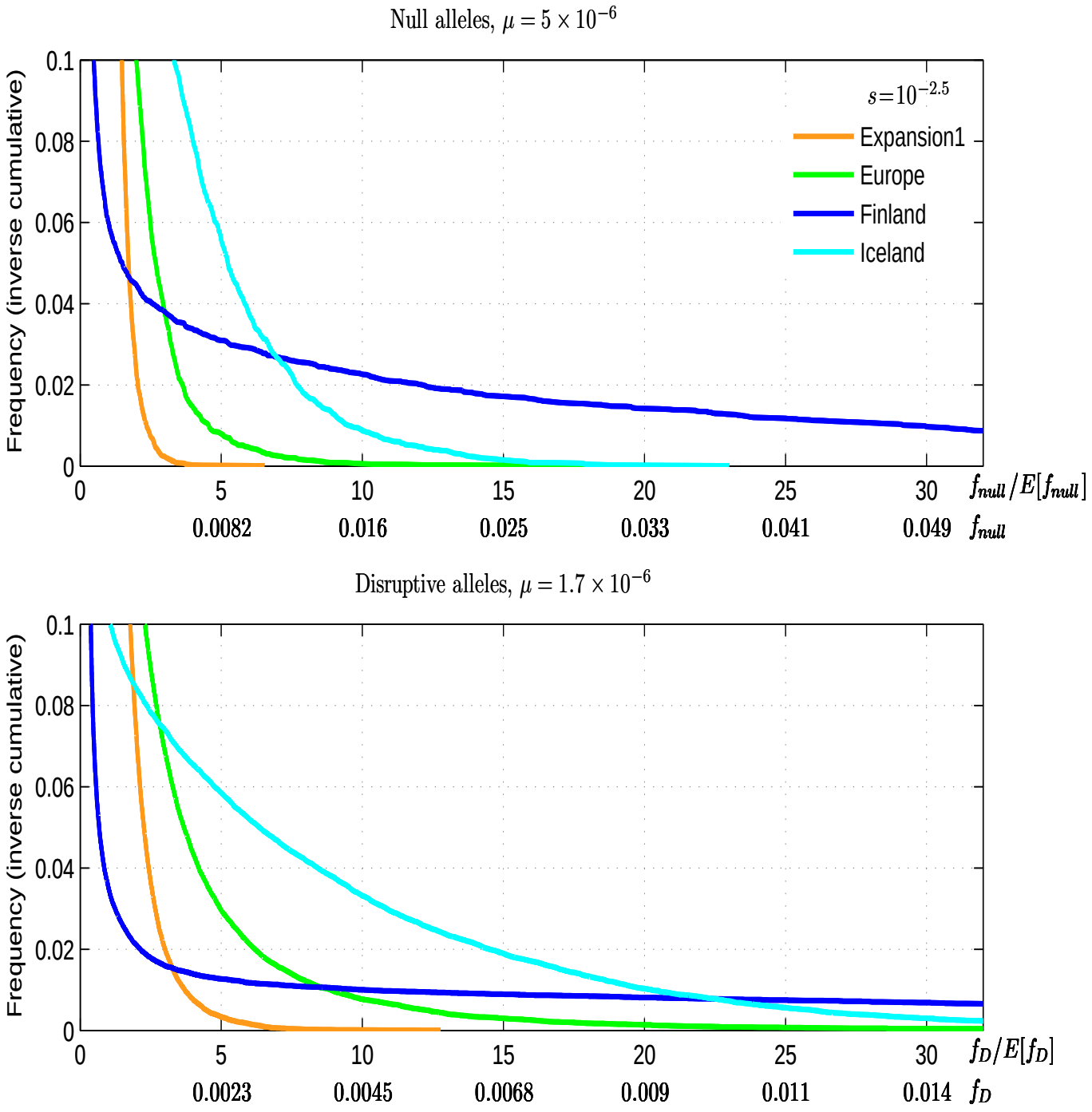

Figure 21: The inverse cumulative distribution of the Combined Allele Frequency (CAF) for $s = 10^{-2.5}$. Similar to Figure S18.

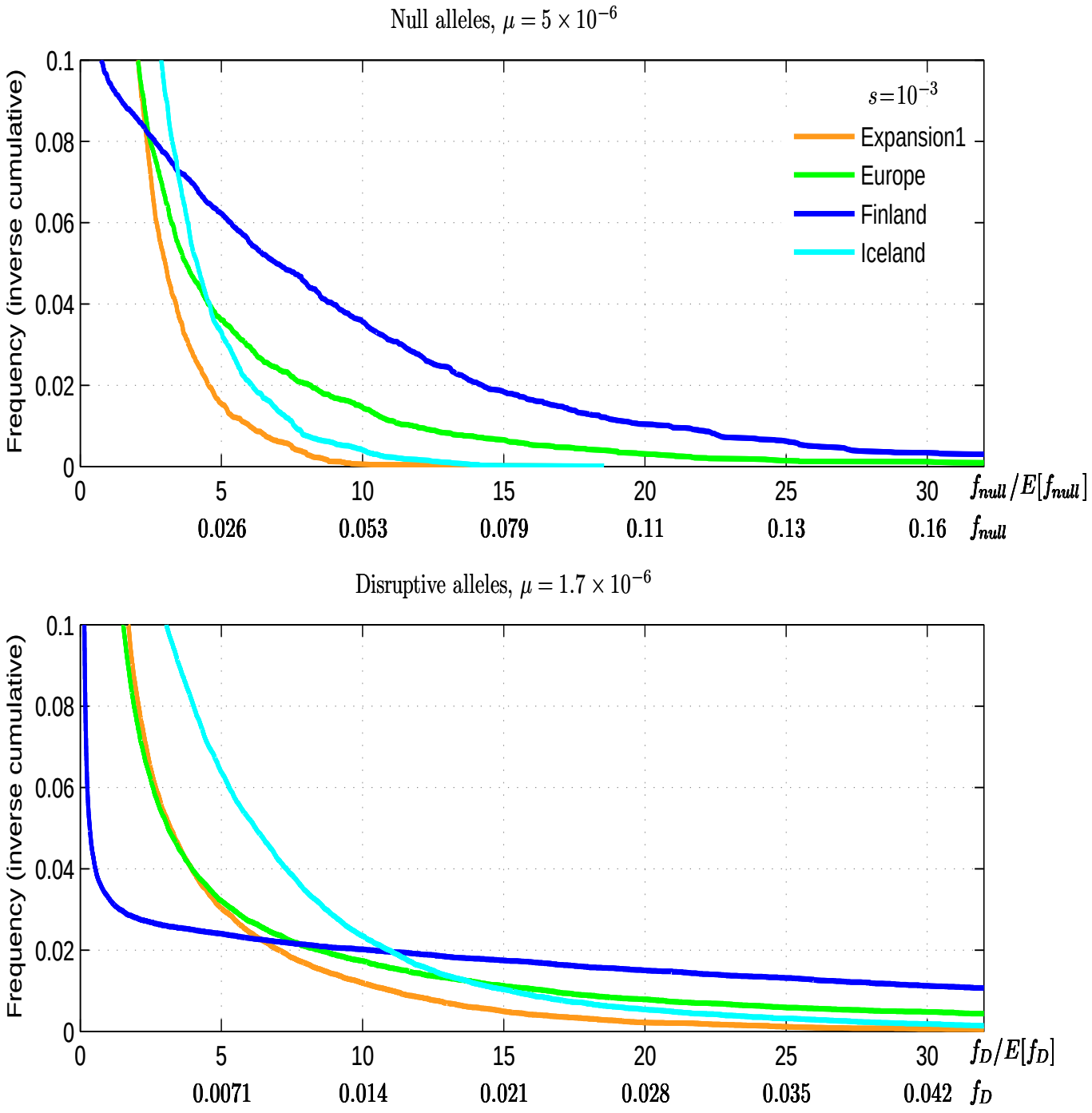

Figure 22: The inverse cumulative distribution of the Combined Allele Frequency (CAF) for $s = 10^{-3}$. Similar to Figure S18.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
Mean	0.076	0.068	0.053	0.024	0.0056	0.0015	0.00049	0.00011	6.4e-005	2.2e-005
St.d.	0.2	0.18	0.16	0.086	0.023	0.0065	0.0021	0.00053	0.00024	7.7e-005
1%	0	0	0	0	0	0	0	0	0	0
5%	0	0	0	0	0	0	0	0	0	0
10%	0	0	0	0	0	0	0	0	0	0
25%	0	0	0	0	0	0	0	0	0	0
50%	0	0	0	0	0	0	0	0	0	0
75%	0.019	0.013	0.007	0.0012	0.0002	0	0	0	0	0
90%	0.27	0.23	0.16	0.054	0.0095	0.0021	0.00085	0.00015	0.0002	0.0001
95%	0.55	0.5	0.38	0.15	0.033	0.0082	0.0027	0.00055	0.0004	0.0002
99%	0.93	0.9	0.82	0.47	0.12	0.033	0.01	0.0026	0.0012	0.0004

Table 13: The mean, standard deviation, and different quantiles of the distribution of CAF for disruptive alleles ($\mu_D = 1.7 \times 10^{-6}$) across different simulations for a population at equilibrium. Values were obtained by simulating $50,000$ realizations corresponding to $50,000$ different simulated genes (see Section 1.3).

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
Mean	0.22	0.19	0.15	0.073	0.018	0.0052	0.0015	0.00065	0.0002	5.7e-005
St.d.	0.33	0.31	0.26	0.15	0.041	0.012	0.0035	0.0013	0.0004	0.00012
1%	0	0	0	0	0	0	0	0	0	0
5%	0	0	0	0	0	0	0	0	0	0
10%	0	0	0	0	0	0	0	0	0	0
25%	0.0013	0.00075	0.00035	0.0001	5e-005	0	0	0	0	0
50%	0.051	0.039	0.025	0.0086	0.0019	0.00045	0.00015	0.0001	5e-005	0
75%	0.32	0.27	0.19	0.071	0.017	0.0048	0.0013	0.0007	0.0002	5e-005
90%	0.73	0.66	0.51	0.23	0.057	0.015	0.0045	0.0019	0.00065	0.0002
95%	0.93	0.88	0.73	0.39	0.091	0.027	0.0083	0.0031	0.001	0.0003
99%	1.4	1.3	1.1	0.74	0.2	0.056	0.017	0.0064	0.002	0.0006

Table 14: The mean, standard deviation, and different quantiles of the distribution of CAF for all null alleles ($\mu_D = 5 \times 10^{-6}$) across different simulations for a population at equilibrium. Values were obtained by simulating $50,000$ realizations corresponding to $50,000$ different simulated genes (see Section 1.3).

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
Mean	0.076	0.07	0.054	0.024	0.0054	0.0015	0.00047	0.00016	8.2e-005	1.8e-005
St.d.	0.19	0.18	0.15	0.084	0.018	0.0031	0.00033	7.1e-005	2.4e-005	7.5e-006
1%	0.00019	0.00019	0.00019	0.00018	0.00017	0.00014	9.6e-005	4.9e-005	3.4e-005	6.5e-006
5%	0.00033	0.00032	0.00032	0.0003	0.00028	0.00023	0.00015	6.9e-005	4.6e-005	8.5e-006
10%	0.00044	0.00043	0.00043	0.0004	0.00036	0.00029	0.00018	8.2e-005	5.3e-005	9.5e-006
25%	0.00074	0.00072	0.00071	0.00065	0.00057	0.00044	0.00026	0.00011	6.6e-005	1.2e-005
50%	0.0015	0.0015	0.0014	0.0012	0.00099	0.00072	0.00038	0.00015	8.1e-005	1.6e-005
75%	0.018	0.014	0.0087	0.0036	0.0021	0.0012	0.00057	0.0002	9.7e-005	2.2e-005
90%	0.27	0.23	0.16	0.051	0.0084	0.0025	0.00083	0.00026	0.00011	2.9e-005
95%	0.54	0.5	0.38	0.14	0.026	0.0046	0.001	0.0003	0.00012	3.2e-005
99%	0.92	0.91	0.82	0.46	0.09	0.015	0.0017	0.00038	0.00015	3.8e-005

Table 15: Variation in disruptive alleles combined allele frequency f_D for the Expansion1 model. Similar to Table S13.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
Mean	0.22	0.2	0.16	0.07	0.017	0.0052	0.0016	0.00049	0.00015	4.9e-005
St.d.	0.33	0.31	0.26	0.14	0.032	0.0057	0.00062	0.00012	2.9e-005	1.1e-005
1%	0.0017	0.0017	0.0016	0.0016	0.0015	0.0012	0.00068	0.00026	9e-005	2.7e-005
5%	0.0024	0.0024	0.0023	0.0022	0.0019	0.0015	0.00086	0.00031	0.0001	3.2e-005
10%	0.003	0.003	0.0029	0.0026	0.0023	0.0018	0.00096	0.00034	0.00011	3.5e-005
25%	0.0057	0.0054	0.0049	0.0039	0.0032	0.0023	0.0012	0.0004	0.00012	4.1e-005
50%	0.052	0.041	0.029	0.011	0.0054	0.0033	0.0015	0.00048	0.00014	4.9e-005
75%	0.32	0.29	0.2	0.064	0.016	0.0055	0.0019	0.00056	0.00016	5.6e-005
90%	0.73	0.68	0.53	0.21	0.044	0.011	0.0024	0.00065	0.00018	6.3e-005
95%	0.95	0.9	0.75	0.36	0.073	0.016	0.0027	0.00071	0.0002	6.8e-005
99%	1.4	1.3	1.1	0.7	0.15	0.031	0.0038	0.00082	0.00022	7.7e-005

Table 16: Variation in null alleles combined allele frequency f_{null} for the Expansion1 model. Similar to Table S14.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
Mean	0.076	0.068	0.054	0.023	0.0052	0.0014	0.00045	0.00019	6.3e-005	1.8e-005
St.d.	0.2	0.18	0.16	0.083	0.018	0.0032	0.00055	0.00014	2.5e-005	2e-006
1%	7.1e-005	7.2e-005	7.2e-005	6.9e-005	6.7e-005	6.3e-005	5.2e-005	5.1e-005	3.2e-005	1.4e-005
5%	0.00011	0.00011	0.00011	0.00011	9.9e-005	8.9e-005	6.9e-005	6.4e-005	3.6e-005	1.5e-005
10%	0.00016	0.00016	0.00016	0.00015	0.00013	0.00012	8.5e-005	7.4e-005	3.9e-005	1.6e-005
25%	0.00039	0.00038	0.00036	0.00032	0.00027	0.00021	0.00013	9.7e-005	4.5e-005	1.7e-005
50%	0.0014	0.0014	0.0012	0.00099	0.00072	0.0005	0.00026	0.00015	5.6e-005	1.8e-005
75%	0.018	0.014	0.0093	0.0043	0.0023	0.0013	0.00053	0.00024	7.3e-005	2e-005
90%	0.27	0.23	0.16	0.047	0.0089	0.0031	0.001	0.00038	9.5e-005	2.1e-005
95%	0.55	0.49	0.38	0.14	0.024	0.0055	0.0015	0.00048	0.00011	2.2e-005
99%	0.93	0.9	0.83	0.45	0.088	0.015	0.0027	0.00073	0.00015	2.3e-005

Table 17: Variation in disruptive alleles combined allele frequency f_D for the Expansion2 model. Similar to Table S13.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
Mean	0.23	0.19	0.15	0.072	0.017	0.0048	0.0016	0.00057	0.00017	5e-005
St.d.	0.33	0.3	0.25	0.14	0.03	0.006	0.0011	0.00028	3.5e-005	3.3e-006
1%	0.00064	0.00063	0.00059	0.00054	0.00045	0.00035	0.00027	0.00019	0.00011	4.3e-005
5%	0.0013	0.0013	0.0011	0.001	0.00079	0.00059	0.0004	0.00024	0.00013	4.5e-005
10%	0.002	0.002	0.0018	0.0015	0.0011	0.00082	0.00051	0.00028	0.00013	4.6e-005
25%	0.0057	0.0051	0.0043	0.0032	0.0022	0.0014	0.00079	0.00037	0.00015	4.8e-005
50%	0.051	0.04	0.028	0.012	0.0052	0.0027	0.0013	0.00051	0.00017	5e-005
75%	0.34	0.27	0.19	0.067	0.016	0.0056	0.002	0.00072	0.00019	5.2e-005
90%	0.74	0.64	0.52	0.22	0.045	0.011	0.003	0.00095	0.00022	5.4e-005
95%	0.95	0.85	0.72	0.37	0.074	0.016	0.0037	0.0011	0.00024	5.6e-005
99%	1.4	1.2	1.1	0.73	0.15	0.03	0.0057	0.0015	0.00028	5.8e-005

Table 18: Variation in null alleles combined allele frequency f_{null} for the Expansion2 model. Similar to Table S14.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
Mean	0.071	0.064	0.051	0.023	0.0055	0.0014	0.00041	0.00016	7.4e-005	1.7e-005
St.d.	0.19	0.18	0.16	0.091	0.028	0.0069	0.00082	0.00014	2.7e-005	2.1e-006
1%	5.3e-005	5.3e-005	5.3e-005	5.2e-005	5e-005	4.7e-005	4.4e-005	4.4e-005	3.3e-005	1.2e-005
5%	7.1e-005	7.2e-005	7.2e-005	7e-005	6.5e-005	6e-005	5.2e-005	5.3e-005	4e-005	1.4e-005
10%	9.2e-005	9.2e-005	9.2e-005	8.8e-005	8e-005	7.3e-005	6e-005	6e-005	4.5e-005	1.5e-005
25%	0.00018	0.00017	0.00017	0.00016	0.00014	0.00012	8.4e-005	7.7e-005	5.5e-005	1.6e-005
50%	0.00058	0.00056	0.00052	0.00043	0.00034	0.00026	0.00017	0.00011	6.9e-005	1.7e-005
75%	0.005	0.0039	0.003	0.0017	0.0011	0.00072	0.00041	0.00019	8.7e-005	1.8e-005
90%	0.26	0.22	0.15	0.029	0.0044	0.0021	0.00095	0.00032	0.00011	2e-005
95%	0.54	0.49	0.38	0.15	0.016	0.0043	0.0015	0.00044	0.00012	2.1e-005
99%	0.93	0.91	0.84	0.5	0.14	0.026	0.0037	0.00073	0.00016	2.2e-005

Table 19: Variation in disruptive alleles combined allele frequency f_D for the European model. Similar to Table S13.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
Mean	0.21	0.19	0.15	0.068	0.019	0.0055	0.0016	0.00065	0.00017	4.7e-005
St.d.	0.33	0.3	0.27	0.15	0.051	0.013	0.0015	0.00032	3.7e-005	3.4e-006
1%	0.00029	0.00029	0.00028	0.00027	0.00025	0.00025	0.00027	0.0002	0.00012	3.9e-005
5%	0.0005	0.0005	0.00046	0.00043	0.00038	0.00036	0.00037	0.00027	0.00013	4.1e-005
10%	0.00076	0.00076	0.00069	0.00061	0.00053	0.00048	0.00046	0.00031	0.00013	4.2e-005
25%	0.002	0.0019	0.0017	0.0013	0.001	0.00088	0.00069	0.00042	0.00015	4.4e-005
50%	0.02	0.014	0.0091	0.0039	0.0025	0.0019	0.0012	0.00058	0.00017	4.7e-005
75%	0.31	0.26	0.18	0.046	0.0088	0.0044	0.002	0.0008	0.00019	4.9e-005
90%	0.71	0.64	0.55	0.24	0.048	0.011	0.0032	0.0011	0.00022	5.1e-005
95%	0.93	0.87	0.78	0.4	0.11	0.02	0.0043	0.0013	0.00024	5.2e-005
99%	1.4	1.3	1.1	0.74	0.26	0.064	0.0072	0.0017	0.00029	5.5e-005

Table 20: Variation in null alleles combined allele frequency f_{null} for the European model. Similar to Table S14.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
Mean	0.073	0.065	0.053	0.024	0.0052	0.0014	0.00046	0.00016	5.8e-005	1.8e-005
St.d.	0.2	0.18	0.16	0.091	0.027	0.01	0.0051	0.0016	0.00014	3.7e-006
1%	2.7e-005	2.7e-005	2.7e-005	2.7e-005	2.7e-005	2.7e-005	2.6e-005	2e-005	2.1e-005	1.1e-005
5%	3.4e-005	3.4e-005	3.3e-005	3.3e-005	3.3e-005	3.2e-005	3.2e-005	2.5e-005	2.4e-005	1.2e-005
10%	3.8e-005	3.8e-005	3.8e-005	3.7e-005	3.7e-005	3.7e-005	3.6e-005	2.8e-005	2.7e-005	1.3e-005
25%	5e-005	5e-005	5e-005	4.8e-005	4.7e-005	4.6e-005	4.6e-005	3.6e-005	3.2e-005	1.5e-005
50%	7.9e-005	7.7e-005	7.6e-005	7.1e-005	6.6e-005	6.4e-005	6.2e-005	5.2e-005	4.1e-005	1.7e-005
75%	0.00062	0.00035	0.00024	0.00015	0.00011	0.0001	9.5e-005	8.2e-005	5.6e-005	2e-005
90%	0.27	0.22	0.16	0.047	0.00037	0.0002	0.00017	0.00014	8.2e-005	2.2e-005
95%	0.55	0.48	0.39	0.15	0.024	0.0005	0.00032	0.00022	0.00011	2.4e-005
99%	0.93	0.9	0.84	0.49	0.14	0.047	0.0044	0.0009	0.00034	2.9e-005

Table 21: Variation in disruptive alleles combined allele frequency f_D for the Finland model. Similar to Table S13.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
Mean	0.21	0.2	0.15	0.069	0.017	0.0055	0.0019	0.00051	0.00016	5.5e-005
St.d.	0.33	0.31	0.27	0.15	0.049	0.022	0.01	0.0032	0.00031	6.1e-006
1%	0.00014	0.00014	0.00014	0.00014	0.00013	0.00013	0.00012	0.0001	7.1e-005	4.3e-005
5%	0.00017	0.00017	0.00017	0.00017	0.00016	0.00015	0.00014	0.00012	8e-005	4.6e-005
10%	0.0002	0.0002	0.0002	0.00019	0.00017	0.00017	0.00016	0.00013	8.7e-005	4.7e-005
25%	0.00029	0.0003	0.00027	0.00024	0.00022	0.00021	0.00019	0.00015	0.0001	5e-005
50%	0.035	0.031	0.0051	0.00049	0.00032	0.00028	0.00025	0.0002	0.00012	5.4e-005
75%	0.3	0.28	0.19	0.064	0.0012	0.0005	0.00039	0.00028	0.00015	5.8e-005
90%	0.71	0.67	0.53	0.23	0.061	0.004	0.00089	0.00047	0.00021	6.2e-005
95%	0.91	0.87	0.76	0.39	0.11	0.038	0.0028	0.00076	0.00027	6.5e-005
99%	1.4	1.3	1.1	0.77	0.24	0.12	0.054	0.0054	0.00071	7.2e-005

Table 22: Variation in null alleles combined allele frequency f_{null} for the Finland model. Similar to Table S14.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
Mean	0.077	0.069	0.054	0.024	0.0054	0.0014	0.00044	0.00014	6.3e-005	1.4e-005
St.d.	0.2	0.18	0.16	0.085	0.02	0.0044	0.0017	0.00072	0.00021	2.1e-005
1%	0	0	0	0	0	0	0	0	0	0
5%	0	0	0	0	0	0	0	0	0	0
10%	4e-006	4e-006	4e-006	4e-006	4e-006	2e-006	2e-006	0	2e-006	0
25%	1.8e-005	1.8e-005	1.6e-005	1.4e-005	1.2e-005	1e-005	8e-006	8e-006	8e-006	4e-006
50%	0.00015	0.00012	9.4e-005	6.4e-005	4.4e-005	3e-005	2.2e-005	2e-005	2e-005	8e-006
75%	0.02	0.016	0.011	0.0051	0.0016	0.00016	6.8e-005	5e-005	4.8e-005	1.8e-005
90%	0.28	0.23	0.16	0.048	0.012	0.0042	0.00049	0.00012	0.00011	3.6e-005
95%	0.55	0.49	0.37	0.14	0.028	0.0084	0.0029	0.00031	0.0002	5e-005
99%	0.93	0.9	0.83	0.46	0.098	0.021	0.0089	0.0034	0.00081	9.6e-005

Table 23: Variation in disruptive alleles combined allele frequency f_D for the Iceland model. Similar to Table S13.

s	0	10^{-5}	$10^{-4.5}$	10^{-4}	$10^{-3.5}$	10^{-3}	$10^{-2.5}$	10^{-2}	$10^{-1.5}$	10^{-1}
Mean	0.22	0.2	0.16	0.073	0.018	0.0054	0.0017	0.00054	0.00013	5.5e-005
St.d.	0.33	0.3	0.27	0.15	0.034	0.0086	0.0034	0.0014	0.00025	3.6e-005
1%	3.9e-005	3.6e-005	3.4e-005	2.8e-005	2.2e-005	1.8e-005	1.4e-005	1.2e-005	1.2e-005	6e-006
5%	0.0001	9.8e-005	9.2e-005	7.6e-005	6e-005	4.6e-005	3.2e-005	2.3e-005	2e-005	1e-005
10%	0.00026	0.00024	0.00018	0.00014	9.6e-005	6.6e-005	4.6e-005	3.2e-005	2.8e-005	1.8e-005
25%	0.0057	0.0052	0.0037	0.0019	0.00042	0.00015	8.2e-005	6.4e-005	4.2e-005	3.2e-005
50%	0.053	0.044	0.028	0.014	0.0058	0.0016	0.0002	0.00013	7e-005	4.8e-005
75%	0.31	0.27	0.2	0.07	0.019	0.0072	0.0015	0.0003	0.00013	7e-005
90%	0.73	0.67	0.55	0.22	0.051	0.015	0.0055	0.0012	0.00022	9.8e-005
95%	0.92	0.88	0.78	0.37	0.084	0.022	0.0087	0.0027	0.00035	0.00012
99%	1.4	1.3	1.2	0.74	0.17	0.041	0.016	0.0076	0.0012	0.00018

Table 24: Variation in null alleles combined allele frequency f_{null} for the Iceland model. Similar to Table S14.

9 RVAS Strategy 5: Using Gene Sets

Another attractive strategy is to move beyond single genes and study gene-sets. By aggregating signal from multiple genes, we may potentially enhance the association signal and improve power. Consider a simple approach for testing association with gene-sets, where we test for burden of disruptive alleles in all genes in the set in cases vs. controls. Consider a set S with m genes, with disruptive CAF $f_1, .., f_m$ and effect sizes $\lambda_1, .., \lambda_m$ on disease. We define the total cumulative frequency in the set $f_S = \sum_i f_i$, and the average excess effect size for the set is $\lambda_{avg} = \frac{\sum_i f_i \lambda_i}{f_S}$.

Then, the sample size required to detect association is similar to eq. (3.6) for one gene, with the frequency f replaced by $\sum_i f_i$, and with the effect size λ replaced by $\beta\lambda$, where β is the (weighted) fraction of genes having an effect, $\beta \equiv \frac{\sum_i f_i 1_{\{\lambda_i = \lambda\}}}{\sum_i f_i}$, giving:

$$n_{a,b} = \frac{\nu_{a,b}}{4[(1+\lambda_{avg})f_S log(1+\lambda_{avg}) + (1-(1+\lambda_{avg})f_S) log\frac{1-(1+\lambda_{avg})f_S}{1-f_S}]}. \tag{9.1}$$

Compared to a single gene with cumulative frequency f_C and effect size λ, the ratio of the two sample sizes is,

$$\frac{n_{a,b}(f_C, \lambda)}{n_{a,b}} = \frac{[(1+\lambda)f_C log(1+\lambda) + (1-(1+\lambda)f_C) log\frac{1-(1+\lambda)f_C}{1-f_C}]}{[(1+\lambda_{avg})f_C log(1+\lambda_{avg}) + (1-(1+\lambda_{avg})f_S) log\frac{1-(1+\lambda_{avg})f_S}{1-f_S}]}. \tag{9.2}$$

Whether a gene set increase or decreases power depends on the extent to which it is enriched with genes having large effect sizes.

Similarly, the gain in power due to missense alleles can be obtained by modifying eq. (5.3) accordingly. The study of gene-sets present a tradeoff - including more genes may increase the allele frequency, but may decrease the average effect size. Using gene-set based on prior biological knowledge (e.g. genes in a certain pathway, sharing a same function, or GWAS-hits) may allow increased power.

If genes have different selection coefficients, effect sizes, and gene lengths, the situation becomes more complicated. For example, it will no longer be optimal to filter alleles in different genes by a single threshold. In this case, it is still possible to use gene-specific thresholds, and combine the alleles obtained from different genes.

10 RVAS Strategy 6: *De novo* Mutations

Another strategy is to study only *de novo* disruptive mutations. At a first glance, using *de novo* alleles might seem counter productive, because their frequency in the population is so low. However, this strategy may make sense in certain cases - namely when searching for genes with huge effect sizes.

One advantage of *de novo* alleles is that we do not need to estimate the allele frequency, f_C, from population surveys: the frequency of *de novo* alleles depends only on is determined by the mutation rate (and not the selection coefficient), which can be reliably estimated from genomic parameters, as discussed above.

Consider a gene with disruptive mutation rate μ_D, and effect size λ. The sample size needed to detect *de novo* alleles is obtained simply by replacing f with μ_D in eq. (3.7), giving,

$$n_{a,b}(de\ novo) \approx \frac{\nu_{a,b}}{4\mu_D g(\lambda)}. \tag{10.1}$$

The average mutation rate of disruptive alleles in our typical gene is $\mu_D \approx 1.7 \times 10^{-6}$, although there is considerable variation in rates between genes. Longer genes with higher overall mutation rates will tend to have more *de novo* alleles, and will be slightly easier to detect.

In Figure S23, we plot the sample size needed to detect burden of *de novo* mutations, for a typical gene. The sample size is huge unless the effect size is also huge. For our typical gene, more than $100,000$ cases must be studied for *de novo* mutations to detect a gene in which disruptive alleles increase risk by 20-fold. In the case cited (autism with several mental disabilities), the effect size is 300-fold.

One benefit of using *de novo* alleles is that the proportion of null alleles will be higher (compared to taking all missense alleles), and is insensitive to the selection coefficient - thus using *de novo* alleles may be particularly attractive for genes under strong selection.

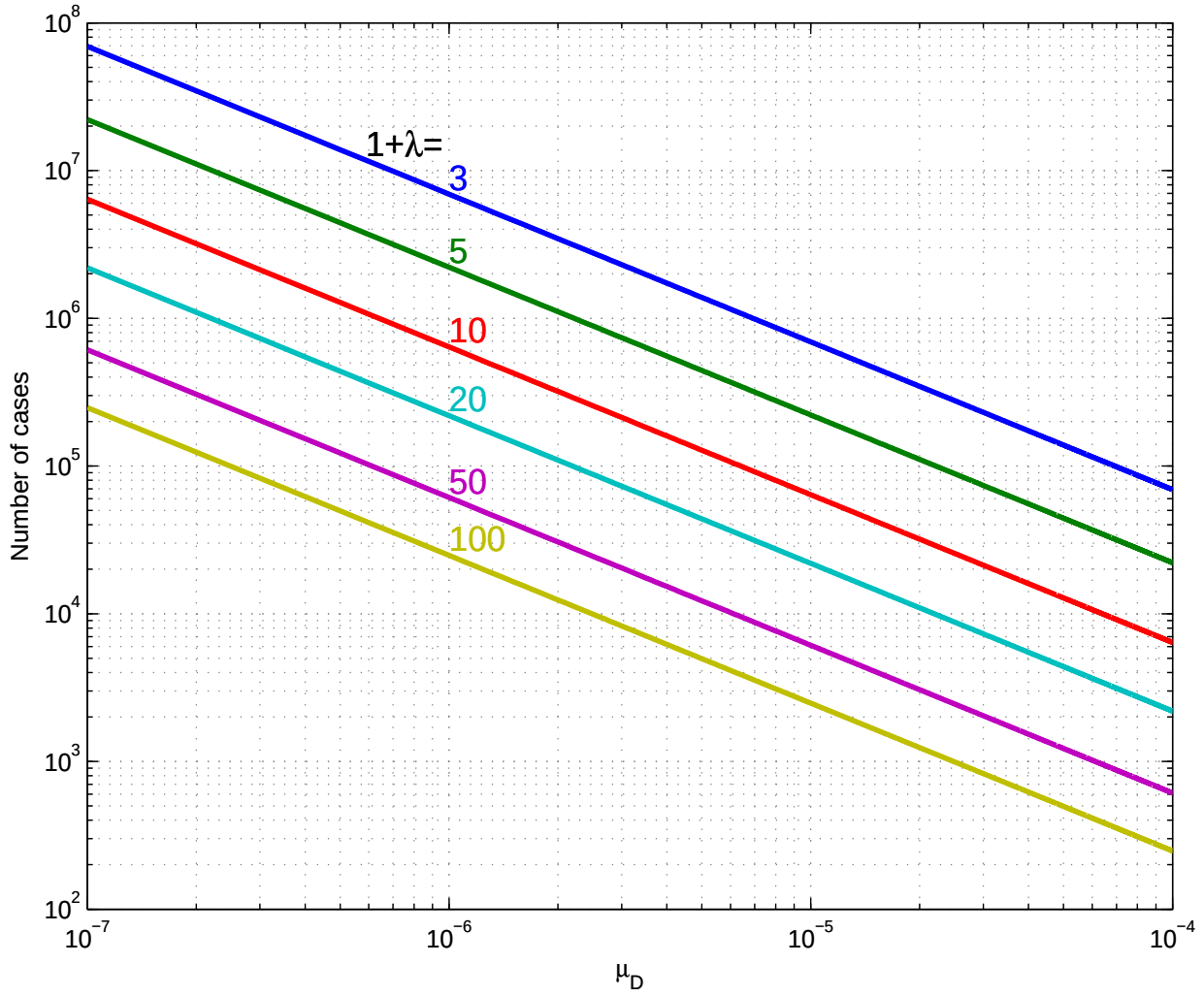

Figure 23: The number of samples (cases) required to detect *de novo* alleles as function of effect size $1 + \lambda$ (shown as different colored curves), and the overall genic mutation rate μ_D (x-axis). We require 90% power and $0.05/20,000$ p-value cutoff. Power increases with both effect size and mutation rate. Overall, the sample size is huge, unless effect size is strong and mutation rate is high. For example, for an average gene ($\mu_D = 1.7 \times 10^{-6}$), and with $1 + \lambda = 20$, the sample size is over $100,000$ cases.

11 Prospects for RVAS in Non-coding Regions

It is natural to want to extend RVAS to non-coding regions, especially in light of the observation that the majority of known CVAS hits are non-coding. However, non-coding regions provide a major challenge for RVAS. The main issue is which alleles to aggregate. The density of functional alleles in non-coding regions is expected to be lower than in coding regions, and the effect size of functional alleles is also expected to be lower.

If we let α_{NC} be the fraction of bases encoding functional DNA elements in a non-coding region that is aggregated for RVAS, the effect size of these functional alleles will be diluted by a factor of α_{NC}. We can use the same formulations from the two-class model describing missense alleles to study non-coding alleles as well, including power calculations. If we take all alleles below a threshold T^* we get, similarly to eq. (5.3),

$$\frac{E[\mathcal{LLR}_{non-coding}(T^*)]}{E[\mathcal{LLR}_{non-coding-functional}]} \sim \frac{\alpha_{NC}\Psi_s(T^*)[(1+\rho_s(T^*)\lambda)\log(1+\rho_s(T^*)\lambda)-\rho_s(T^*)\lambda]}{\rho_s(T^*)[(1+\lambda)\log(1+\lambda)-\lambda]} \tag{11.1}$$

where $E[\mathcal{LLR}_{non-coding-functional}$ would be the aggregate effect size in an hypothetical case where we know precisely the functional non-coding variants, and $E[\mathcal{LLR}_{non-coding}(T^*)]$ would be the aggregate effect size we use in practice, where we take all alleles with frequency below T^* in an entire non-coding region. The resulting inflation in sample size will be between $\frac{1}{\alpha_{NC}}$ and $\frac{1}{\alpha_{NC}^2}$, depending on the effect size λ.

11.1 Multiple Hypothesis Burden for Genome-Wide Testing

When performing association tests across the entire genome, the burden of multiple hypothesis testing is also increased relative to the situation of testing only genes. Suppose that we wish to achieve an overall type-1 error of $a = 0.05$. When testing for association across $20,000$ individual genes, one can simply apply a Bonferroni correction to set the required nominal significance level at $a/20,000 = 2.5 \times 10^6$.

When testing for association in overlapping windows across the entire genome, a different approach is needed to calculate the required nominal significance level. Suppose that we study overlapping windows W_i, each of length w base-pairs, and calculate for each window a test statistic X_i with a Normal distribution under the null hypothesis of no association, and we are interested in one-sided tests, i.e. detecting X_i's values which are unusually large (Such a statistic arises naturally when testing for excess of rare alleles observed in cases relative to the expected number based on a large set of controls). To achieve genome-wide significance, we need to consider the distribution of $X = \max_i X_i$ under the null, where the maximum is taken across the entire human genome. The total number of different tests is thus $\sim 3 \times 10^9$ (equal to the number of positions in the human genome, because a window can start at each nucleotide). Tests corresponding to non-overlapping windows

can be considered independent (because of linkage disequilibrium among *rare* variants), while tests corresponding to overlapping windows are clearly correlated, with the correlation proportional to the extent of overlap. The effective number of independent tests is thus somewhere between $\frac{3\times 10^9}{w}$ and 3×10^9.

Considering the situation of family-based linkage analysis, Lander and Botstein [49] showed that the nominal significance level corresponding to a desired genome-wide significance level can be calculated in terms of the largest deviation of an Orenstein-Uhlenbeck diffusion process [50]. Briefly, an Orenstein-Uhlenbeck process describes a stationary random process $Y(t)$ for $t \in [0, T]$ such that (i) $Y(t)$ is a standard normal random variable for all t and (ii) the local correlation of $Y(t)$ and $Y(t+x)$ is $r(x) = e^{-|x|}$. According to [51] (Theorem 12.2.9., page 232), the distribution of $M(T) = \max_{t\in[0,T]} Y(t)$, the maximum value of the process across an interval of length T, is asymptotically

$$P[M(T) > u] \sim Tu\phi(u) \sim Tu^2[1 - \Phi(u)] \tag{11.2}$$

where $\phi(u)$ is the standard Gaussian probability density function, and $\Phi(u)$ is the standard Gaussian cumulative distribution function.

In the case of RVAS, the statistics X_i together approximately follow an Orenstein-Uhlenbeck distribution (after rescaling the x-axis by a factor of $\frac{1}{w}$ and ignoring the division of the genome into 23 chromosomes). The local correlation function $r(x) = \max\Big(0, 1 - \frac{|x|}{w}\Big)$ is triangular rather than exponential, but can be reasonably approximated by the exponential $e^{-|x|/w}$. (One can also directly study the triangular covariance function; see for example [52].)

With this approximation, the large deviation theory for the Orenstein-Uhlenbeck process implies that the cumulative distribution of X is given by

$$P(X > x) \approx \frac{3 \times 10^9}{w} x^2 \Big(1 - \Phi(x)\Big). \tag{11.3}$$

(This formula follows from eq. (11.2) above. It is equivalent to Proposition 2 in [49], where the formula is expressed in terms of chi-square, rather than Gaussian, test statistics; the two formulations are equivalent by a simple square transformation).

With $w = 4$ kilobases and $a = 0.05$, we have following equation,

$$0.05 = 750,000 x^2 \Big(1 - \Phi(x)\Big). \tag{11.4}$$

The solution (obtained numerically) is $x \sim 5.9$. This corresponds to a genome-wide significance level of $(1 - \Phi(x)) \approx 1.9 \times 10^{-9}$, which is much more stringent than the value 2.5×10^{-6} required to control for testing $20,000$ genes.

The sample size when testing using a Gaussian test-statistic (Z-score) is proportional to $\left(\Phi^{-1}(a) + \Phi^{-1}(b)\right)^2$, where a is the desired type-1 error, b is the desired type-2 error, and Φ^{-1} is the inverse of the standard Gaussian cumulative distribution function (Z-value). For the case of 50% power ($b = 0.5$ and $\Phi^{-1}(b) = 0$) and family-wise error rate of 0.05, we have $\left(\Phi^{-1}(a) + \Phi^{-1}(b)\right)^2 \approx (-4.6 + 0)^2 = 20.8$ when testing $20,000$ genes and $\left(\Phi^{-1}(a) + \Phi^{-1}(b)\right)^2 \approx (-5.9 + 0)^2 = 34.7$ when scanning the entire genome. Thus, 1.66-fold more samples are required when scanning the genome in windows of 4 kb than when testing $20,000$ genes. The corresponding numbers for 90% power ($b = 0.1$ and $\Phi^{-1}(b) = -1.3$) are 34.2 and 51.4, corresponding to a 1.5-fold increase in sample size.

12 Estimating the Parameters s and α in the Two-class Model

In our analyses, we have assumed that we know the (gene-specific) parameters s and α in our two-class model. In practice, these parameters are unknown and must be estimated. We provide estimators for s and α, and discuss their shortcomings - which are especially serious in the case of α.

12.1 Estimating s

We can estimate s from the frequency f_D of disruptive alleles. For strong selection, we can use equation (1.14) to obtain:

$$\hat{s} = \frac{\mu_D}{f_D} \tag{12.1}$$

To know if the estimator is useful, we need to understand its variance. There are two sources of variation:

1. Sampling noise in our estimation of the true value of f_D in the population. This variation can be made arbitrarily small by increasing the sample size.
2. Stochastic noise in the realized value of f_D around its expected value (given s), resulting from random events during population history (e.g., two genes with the same s might reach different allele frequencies due to stochastic variation in the random births of new alleles). This variation cannot be reduced by increasing the sample size, but rather is determined by population history.

Suppose that we use equation (12.1) to infer the selection coefficient s, based on the observed frequency f_D of disruptive alleles. Let $s(x, y)$ be the value of s, such that $Prob(\hat{s} \geq x) = y$. We can define a 95% confidence interval for s around $\hat{s}$ to be $[s(\hat{s}, 0.025), s(\hat{s}, 0.975)]$ and a "median" value to be $s(\hat{s}, 0.50)$.

Figure S24 shows that confidence interval and median value for various values of $\hat{s}$ for the European population. For $\hat{s} > 7 \times 10^{-3}$, the confidence interval is acceptably tight for our purposes - a factor of $\sim$9-fold between the upper and lower bound. By contrast, it is hard to obtain good estimates when selection is weak: the size of the confidence interval (relative to the median) blows up. Fortunately, we care primarily about the case of strong selection.

One way to improve the estimator would be to combine estimates from multiple unrelated populations.

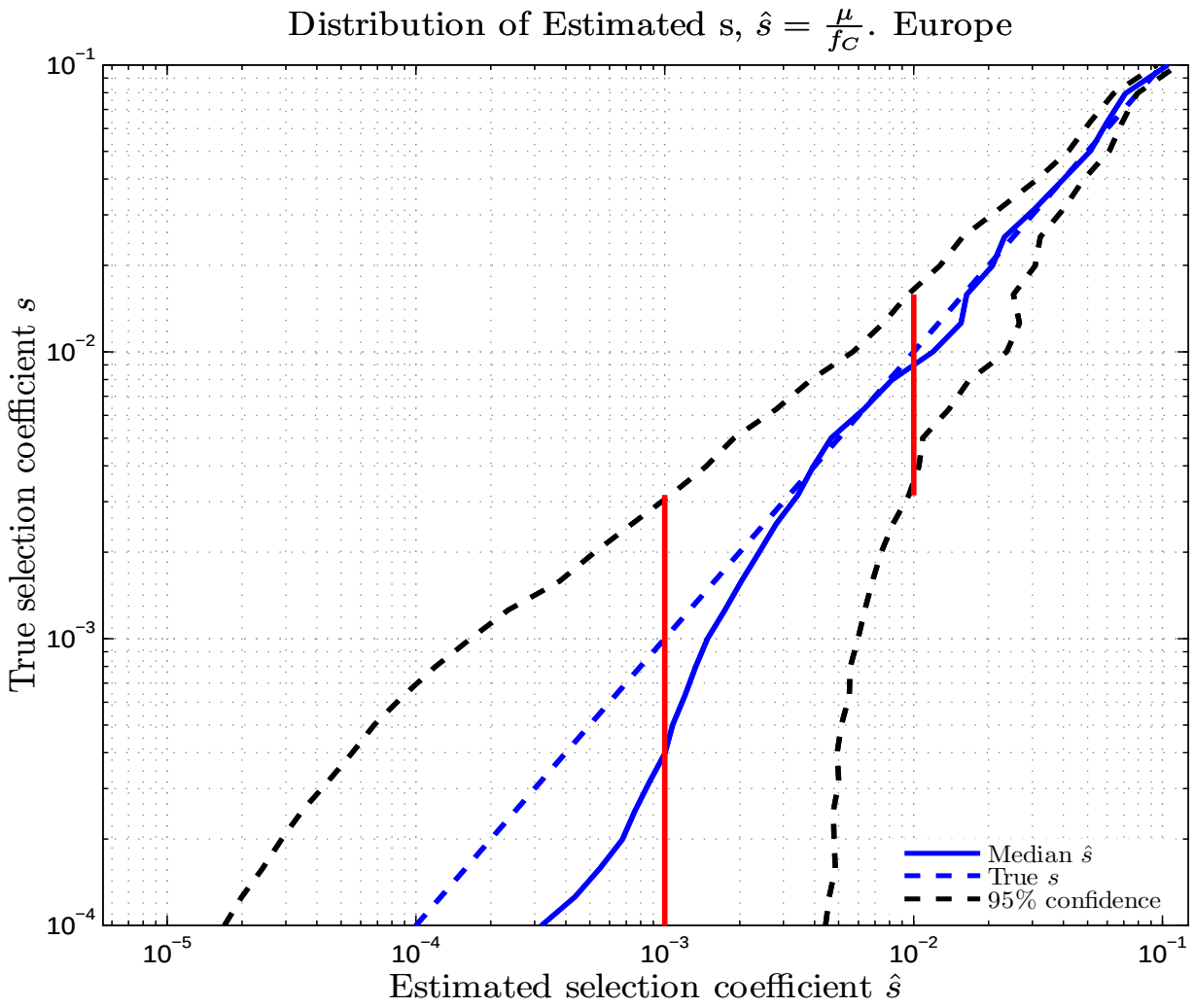

Figure 24: Confidence intervals around estimated value of s. The y-axis shows true values of s, while the x-axis shows values of $\hat{s}$. Given an estimate $\hat{s}$, the curves show the upper confidence limit $s(\hat{s}, 0.975)$, the lower confidence limit $s(\hat{s}, 0.025)$, and the median $s(\hat{s}, 0.50)$. Values obtained from $8,000$ simulations for the European population model, assuming $\mu_G = 1.7 \times 10^{-6}$. Note: The estimator $\hat{s} = \frac{\mu_D}{f_D}$ is biased upwards. Although $E[f_D] = \frac{\mu_D}{s}$, we have $s < E[\frac{\mu_D}{f_D}]$ by Jensen's inequality applied to the function $\frac{1}{x}$, giving $E[\frac{1}{x}] > \frac{1}{E[x]}$.

12.2 Estimating α

One approach is to estimate α based on population data by comparing f_M, the observed CAF for missense alleles (f_M), to f_{M^*}, the CAF that would be expected if all missense alleles were neutral.

We have:

$$f_M = f_{M,neutral} + f_{M,null} \tag{12.2}$$

For a population at equilibrium and strong selection, we can write (by Proposition 2):

$$\begin{aligned} E[f_M] &\approx (1-\alpha) f_{M^*} + \alpha \frac{\mu_M}{s} \\ &\approx (1-\alpha) 4N_{eq}\mu_M + \alpha \frac{\mu_M}{s} \end{aligned} \tag{12.3}$$

where N_{eq} is the equilibrium population size for the ancestral human population.

Moreover, this approximation also holds, in expectation, for modern populations (because, as shown above and in the main text, the CAF is essentially unchanged by recent human demography).

We can neglect the term $f_{M,null} \approx \alpha \frac{\mu_M}{s}$ in eq. (12.3) because

$$\frac{(1-\alpha)4N_{eq}\mu_M}{\alpha(\mu_M/s)} = \frac{1-\alpha}{\alpha} 4N_{eq}s \gg 1 \tag{12.4}$$

provided that selection is strong. (For very weak selection, one could include the term.)

We can thus solve eq. (12.3) for α to obtain

$$\alpha \approx 1 - \frac{E[f_M]}{4N_{eq}\mu_M}. \tag{12.5}$$

This provides an estimator $\hat{\alpha}$:

$$\hat{\alpha} = 1 - \frac{f_M}{4N_{eq}\mu_M}. \tag{12.6}$$

The variation in the estimator $\hat{\alpha}$ depends on the variation in the CAF for neutral missense alleles.

As for s above, there are two sources of variation: (i) sampling noise, which can be decreased by increasing sample size and (ii) stochastic variation in population history, which cannot be reduced by increasing sample size.

As shown in Tables S13-S24, the coefficient of variation in the CAF under neutrality is large: the standard deviation is roughly equal to the mean. This inherently limits the precision in estimates of α.

Another approach is to estimate α based on evolutionary comparison, using the estimator:

$$\hat{\alpha}_{evo.} = \frac{dN}{dS} \tag{12.7}$$

where dN denotes the number of non-synonymous (missense) substitutions per non-synonymous site, and dS denotes the number of synonymous (silent) substitutions per synonymous site.

If our two-class model is strictly true, this estimator should perform well - because evolution will largely reject null alleles and allow neutral alleles. Thus, the "proportional deficit" in missense alleles (relative to silent alleles) will be an accurate reflection of α.

A problem arises, however, if there are also hypomorphic alleles (between null and neutral). Evolution will reject hypomorphic alleles when $s \gg \frac{1}{N_{eff}}$, where N_{eff} is the effective population size of the species. For species with large population sizes, evolution will tend to reject alleles under weak selection. For example, evolution will tend to reject alleles with $s = 10^{-4}$ in a population with 10^5 individuals.

In this case, the estimator based on evolution may overestimate the true proportion α of newborn mutations that are truly null. Still, it is still useful to have an upper bound on α.

Improved methods for estimating s and α would be useful.

13 Estimating Variance Explained by Rare Variants

A major open question is the extent to which rare variants contribute to the heritability of complex diseases. To answer this question, we must be able to estimate the heritability explained by rare variants found in RVAS. Here, we provide the simple formula for variance explained for dichotomous trait. We then apply this to two studies of rare variants in two diseases, and calculate the variance explained by different classes of alleles.

13.1 Calculating Variance Explained

For a disease with prevalence π and an individual allele with frequency f_i and effect size λ_i, the contribution to heritability for a disease trait is,

$$V_i = \frac{2f_i}{1-2f_i}\frac{\pi}{1-\pi}\lambda_i^2. \tag{13.1}$$

Therefore, for a class of null alleles with CAF f_{null} and effect size λ, the total variance explained by the class when f_{null} is small is,

$$V \approx 2f_{null}\frac{\pi}{1-\pi}\lambda^2. \tag{13.2}$$

13.2 Sample Size Required to Detect Loci Explaining a Given Proportion of the Phenotypic Variance

We can calculate the number of samples needed to detect genes that explain a given proportion of the phenotypic variance of a dichotomous trait. The power to detect an association of a class of alleles is directly related to the total phenotypic variance explained by these alleles. Consider a gene with classes of disruptive (D) and (all) *null* alleles, with combined frequencies f_D and f_{null}, respectively, and effect size λ.

By taking eq. (13.2) for the variance explained by all null alleles, and eq. (3.7) for the sample size required when testing for disruptive alleles we get,

$$n_{a,b} \approx \left[NCP^*(a,b)\frac{f_{null}}{f_D}\frac{\pi}{1-\pi}\right]\frac{\lambda^2}{g(\lambda)}\frac{1}{V}. \tag{13.3}$$

Therefore, for a set of alleles with combined variance explained V, the sample size is proportional to $\frac{\lambda^2}{g(\lambda)}$. The function $\frac{\lambda^2}{g(\lambda)}$ is monotonically increasing with λ. If the effect size λ is small, $\frac{\lambda^2}{g(\lambda)} \approx 2$ and the sample size is not sensitive to the effect size. For large values of λ, the sample size increases as we increase λ and keep the variance explained fixed (by decreasing the allele frequencies). The maximum possible sample size occurs for the largest possible effect size, which corresponds alleles

causing full penetrance, i.e. $\pi(1+\lambda) = 1$ or $\lambda = \frac{1-\pi}{\pi}$. In this case we get $g(\lambda) = g(\frac{1-\pi}{\pi}) = \frac{\pi-1-\log(\pi)}{\pi}$ and

$$n_{a,b}^{(MAX)}(V) \approx NCP^*(a,b)\frac{\pi-1}{\log(\pi)+1-\pi}\Big[1+\alpha\frac{\mu_M}{\mu_D}\Big]\frac{1}{V} \tag{13.4}$$

where we recall that a fraction α of missense alleles at birth is null, hence $\frac{f_{null}}{f_D} = 1 + \alpha\frac{\mu_M}{\mu_D}$. For example, suppose that we want to identify genes explaining at least 1% of the phenotypic variance (with significance thresholds of $a = 0.05/20,000$ and $b = 0.1$, and taking $\alpha = 0.25$ and $\frac{\mu_M}{\mu_D} = 7.3$). We have $\frac{\pi-1}{\log(\pi)+1-\pi} = 0.27, 0.46$ and 0.64 for $\pi = 0.01, 0.05, 0.10$. These correspond to sample sizes of 2775, 4707, and 6503 cases, respectively. Once we attain these sample sizes, we ar thus reasonably confident that we will have identified all genes explaining at least 1% of the phenotypic variance.

13.3 Estimating heritability explained by rare variants for two traits

For two recent RVAS studies that detected association with candidate genes, we calculate the proportion of phenotypic variance explained by rare variants in the genes. In each study, the authors designated the class of alleles that they believe to be deleterious - based on frequency and either (i) genetic and evolutionary data (in the first case) or (ii) biochemical data (in the second case). In calculating the proportion of the phenotypic variance explained, we use this class as the class of null alleles.

13.3.1 Blood Pressure [Ji et al 2008]

Ji et al. [45] studied association between blood pressure and rare variants in three genes (*SCL12A3*, *SCL12A1*, and *KCNJ1*) in which strong mutations cause Bartter's or Gittleman's syndrome, rare recessive diseases in which patients have extremely low blood pressure.

The authors defined (i) rare variants as being those with population frequency $\leq$ 1% and (ii) deleterious rare variants as being the subset of rare variants occurring at sites that are evolutionarily conserved and/or are known to cause Bartter's or Gittleman's syndrome. For deleterious rare variants, the total counts are 27, 11 and 12 and the allele frequency is 0.43%, 0.17% and 0.19% for *SLC12A3*, *SLC12A1* and *KCNJ1*, respectively.

The authors studied 3126 individuals from the general population with respect to blood pressure and rare variants in the three genes.

We converted blood pressure into a set of dichotomous traits, by considering individuals with blood pressure in the top $10\%, 20\%, 30\%, 40\%$ and 50% of the sample as well as the bottom $10\%, 20\%, 30\%, 40\%$ and 50% of the sample (based on Table 3 from their paper). Our results are shown in Table S25.

Rare variants in the three gene together explain $\approx 0.1\% - 0.5\%$ of the phenotypic variance in the dichotomous traits, depending on the percentile chosen. The average proportion explained by each gene is thus $0.04\% - 0.17\%$. The proportion of the phenotypic variance explained is similar if blood pressure is treated as a quantitative trait.

Gene	Cases	Unaffecteds	$1+\lambda$	V
n	313 (top 10%)	2813 (bottom 90%)		
SLC12A3 (CAF=0.43%)	1	26	1.08	0.051%
SLC12A1 (CAF=0.17%)	0	11	1.13	0.056%
KCNJ1 (CAF=0.19%)	0	12	1.15	0.074%
All 3 Genes (CAF=0.79%)	1	49	1.11	0.17%
n	626 (top 20%)	2500 (bottom 80%)		
SLC12A3	3	24	1.12	0.052%
SLC12A1	1	10	1.16	0.035%
KCNJ1	1	11	1.18	0.051%
All 3 Genes	5	45	1.15	0.13%
n	939 (top 30%)	2187 (bottom 70%)		
SLC12A3	3	24	1.28	0.16%
SLC12A1	1	10	1.33	0.085%
KCNJ1	1	11	1.35	0.11%
All 3 Genes	5	45	1.31	0.35%
n	1252 (top 40%)	1874 (bottom 60%)		
SLC12A3	5	22	1.37	0.18%
SLC12A1	1	10	1.55	0.15%
KCNJ1	1	11	1.58	0.19%
All 3 Genes	7	43	1.46	0.5%
n	1453 (top 50%)	1673 (bottom 50%)		
SLC12A3	7	20	1.4	0.14%
SLC12A1	2	9	1.56	0.11%
KCNJ1	1	11	1.77	0.22%
All 3 Genes	10	40	1.52	0.43%

Table 25: Phenotypic variance explained by rare variants in the *SCL12A3*, *SCL12A1*, and *KCNJ1* genes for various dichotomous traits defined by blood pressure. The table shows the number of deleterious rare variants in 'cases', relative risk and proportion of variance explained. This page shows 'cases' at the top percentiles of blood pressure levels (hypertension). The next page shows 'cases' at the bottom percentiles of blood pressure levels (hypotension).

Gene	Cases	Unaffecteds	$1+\lambda$	V
n	313 (bottom 10%)	2813 (top 90%)		
SLC12A3 (CAF=0.43%)	3	24	1.12	0.0014%
SLC12A1 (CAF=0.17%)	2	9	1.85	0.028%
KCNJ1 (CAF=0.19%)	4	8	3.44	0.25%
All 3 Genes (CAF=0.79%)	9	41	1.83	0.12%
n	626 (bottom 20%)	2500 (top 80%)		
SLC12A3	6	21	1.12	0.0031%
SLC12A1	2	9	0.926	0.00047%
KCNJ1	5	7	2.15	0.12%
All 3 Genes	13	37	1.32	0.041%
n	939 (bottom 30%)	2187 (top 70%)		
SLC12A3	10	17	1.25	0.022%
SLC12A1	3	8	0.926	0.0008%
KCNJ1	6	6	1.72	0.083%
All 3 Genes	19	31	1.29	0.056%
n	1252 (bottom 40%)	1874 (top 60%)		
SLC12A3	13	14	1.21	0.026%
SLC12A1	5	6	1.16	0.0057%
KCNJ1	9	3	1.94	0.22%
All 3 Genes	27	23	1.37	0.15%
n	1673 (bottom 50%)	1453 (top 50%)		
SLC12A3	20	7	1.4	0.14%
SLC12A1	9	2	1.56	0.11%
KCNJ1	11	1	1.77	0.22%
All 3 Genes	40	10	1.52	0.43%

Table 19: Phenotypic variance explained by rare variants for various dichotomous traits defined by blood pressure. (cont.)

13.3.2 Type 2 Diabetes [Bonnenfond et al. 2012]

Bonnefond et al. [41] studied association between Type 2 Diabetes and rare variants in *MTNR1B*. By sequencing the genes in 2186 cases and 4804 controls, they found 40 non-synonymous variants associated with the disease. They performed biochemical assays to study the biochemical consequences of each variant on protein function.

Their strongest signal came from a class of 12 variants that they classified as "Loss of Function variants with $MAF < 0.1\%$". They observed 15 counts in cases and 13 counts in controls.

This corresponds to a frequency $f_{null} = 0.4\%$ and a relative risk of $1 + \lambda = 2.5$. The proportion of variance explained is 0.2%.

[Note: The authors reported a somewhat higher relative risk, based on using the KBAC method [30]. The KBAC method gives more weight to lower frequency alleles, based on the notion that they are more likely to be deleterious. The effect of applying the KBAC method to these data is to down-weight the influence of one specific missense allele ($L60R$, which appears in 5 cases and 6 controls). However, the KBAC does not seem appropriate in this case because the authors used biochemical assays to determine which alleles are non-functional: it is problematic to exclude alleles determined biochemically to be non-functional simply because they have higher frequency.]

14 Consequences and Limitations of our Framework

Our simple framework was chosen to maximize insight and intuition. However, it is important to address the limitations of the framework:

1. **Two-class model**
   We assumed above that missense mutations are either null or neutral. We ignored hypomorphic missense alleles, treating them as neutral for the purposes of our analysis. Hypomorphic missense alleles would be expected to contribute only modestly to the power of RVAS because (i) population genetic analyses indicate that the selection coefficients for typical hypomorphic alleles are in the range $10^{-3} - 10^{-4}$ [15, 53, 54], which corresponds primarily to common rather than rare variants, and (ii) the effect sizes for hypomorphic alleles are likely to be smaller, giving rise to smaller enrichment in cases. We tested this expectation by studying a three-class model involving null, hypomorphic and neutral alleles and confirmed that the contribution of hypomorphic alleles is indeed small (see Section 5.1). While likely to contribute little power, hypomorphic missense alleles may seriously skew estimates of effect size based on missense alleles. This consideration strengthens the case for basing estimates of effect size on disruptive alleles.

   We have also ignored the potential role of countervailing missense alleles, which act in the opposite direction as null alleles (for example, protective missense alleles in the case that null alleles are deleterious). While countervailing alleles are known to occur and may be highly informative, we expect that the most missense alleles will act in the same direction as null alleles. If desired, our framework could be extended to allow for a proportion of alleles to have countervailing effects. Countervailing alleles would tend to decrease the power of burden tests, because they are decrease the imbalance of missense alleles between cases and controls. "Dispersion tests" have been proposed as a way to simultaneously detect both risk-increasing and risk-decreasing alleles [27, 40]. While our simple two-class model may be appropriate for the purpose of discovering gene association, once a gene has been implicated by RVAS it will be important to consider carefully the possibility of hypomorphic and countervailing alleles.

2. **Constancy of selection**
   We assumed that the selection coefficient s has not changed over time. However, changes associated with civilization (such as abundant food, population density and differences in exposure to microorganisms) may have increased the prevalence of certain diseases (such as diabetes, hypertension and autoimmune diseases). Alleles that were once neutral or advantageous may thus have become deleterious in the past several thousand years. In such cases, the allele frequency distributions will still largely reflect the ancestral selection regime - unless the modern selection is extremely strong. For the median allele in a neutral distribution (minor allele frequency $\sim 25\%$) to become rare ($< 0.5\%$) requires nearly $4,000$ generations ($\sim 100,000$ years) if $s = 10^{-3}$ and nearly 400 generations ($\sim 10,000$ years) if $s = 10^{-2}$, but

only 40 generations ($\sim 1,000$ years) if $s = 10^{-1}$.

3. **Fixed frequency threshold**
   We considered burden tests that count alleles with frequency below a pre-specified (although gene-specific) threshold. Various authors have discussed ways to use frequency-based weights, to optimize the allele frequency threshold based on data, or to use variable thresholds [28–35].

4. **Unrelated individuals**
   We assumed that cases are unrelated. It may be possible to increase power by studying affected relatives from families with unusually high disease prevalence, because such families may harbor alleles with especially high penetrance, corresponding to high values of λ and greater power. There are two issues with this approach. First, the increase in power from larger must be balanced against the considerable challenge of ascertaining relatives from a very large collection of enriched families. Second, familial enrichment may reflect shared environment rather than shared genetics.

5. **Discrete traits**
   We considered only discrete traits - such as having a disease or being in the top 5% of the tail of the distribution for a quantitative trait. Most RVAS studies have been of this type. Various authors have considered the analysis of purely quantitative traits (e.g. [42, 45]).

6. **Role of environment**
   We ignore the role of environment, including genotype x environment interaction. Environmental factors cannot be readily studied in case-control studies. They are best explored in large prospective cohorts, with extensive information about exposure.

References

1. R.A. Fisher. *The genetical theory of natural selection*. Oxford University Press, 1930.
2. S. Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.
3. W.J. Ewens. *Mathematical population genetics: I. Theoretical introduction*, volume 27. Springer, 2004.
4. M. Kimura. Stochastic processes and distribution of gene frequencies under natural selection. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 20, pages 33–53. Cold Spring Harbor Laboratory Press, 1955.
5. M. Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232, 1964.
6. M. Kimura and J.F. Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49(4):725, 1964.
7. S.N. Evans, Y. Shvets, and M. Slatkin. Non-equilibrium theory of the allele frequency spectrum. *Theoretical Population Biology*, 71(1):109–119, 2007.
8. D. Živković and W. Stephan. Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. *Theoretical Population Biology*, 79(4):184–191, 2011.
9. R.C. Griffiths. The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theoretical Population Biology*, 64(2):241–251, 2003.
10. J.B.S. Haldane. A mathematical theory of natural and artificial selection. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 23, pages 607–615. Cambridge Univ Press, 1927.
11. A. Keinan and A.G. Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743, 2012.
12. H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.
13. A. Helgason, S. Siguróardóttir, J.R. Gulcher, R. Ward, and K. Stefánsson. mtdna and the origin of the icelanders: deciphering signals of recent population history. *The American Journal of Human Genetics*, 66(3):999–1016, 2000.
14. E. Jakkula, K. Rehnström, T. Varilo, O.P.H. Pietiläinen, T. Paunio, N.L. Pedersen, U. deFaire, M.R. Järvelin, J. Saharinen, N. Freimer, et al. The genome-wide patterns of variation expose significant substructure in a founder population. *The American Journal of Human Genetics*, 83(6):787–794, 2008.
15. A.R. Boyko, S.H. Williamson, A.R. Indap, J.D. Degenhardt, R.D. Hernandez, K.E. Lohmueller,

M.D. Adams, S. Schmidt, J.J. Sninsky, S.R. Sunyaev, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics*, 4(5):e1000083, 2008.

16. Y.B. Simons, M.C. Turchin, J.K. Pritchard, and G. Sella. The deleterious mutation load is insensitive to recent population history. *Nature Genetics, in press*.

17. J.R. Lupski, J.W. Belmont, E. Boerwinkle, and R.A. Gibbs. Clan genomics and the complex architecture of human disease. *Cell*, 147(1):32–43, 2011.

18. A. Siepel and D. Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution*, 21(3):468–488, 2004.

19. D.G. Hwang and P. Green. Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences*, 101(39):13994–14001, 2004.

20. P.F. Arndt and T. Hwa. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*, 21(10):2322–2328, 2005.

21. S. Asthana, M. Roytberg, J. Stamatoyannopoulos, and S. Sunyaev. Analysis of sequence conservation at nucleotide resolution. *PLoS Computational Biology*, 3(12):e254, 2007.

22. A.S. Kondrashov. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Human Mutation*, 21(1):12–27, 2003.

23. A. Kong, M.L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S.A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, et al. Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature*, 488(7412):471–475, 2012.

24. B.M. Neale, Y. Kou, L. Liu, A. Maayan, K.E. Samocha, A. Sabo, C-F. Lin, C. Stevens, L-S. Wang, V. Makarov, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397):242–245, 2012.

25. J.C. Cohen, A. Pertsemlidis, I.K. Kotowski, R. Graham, C.K. Garcia, and H.H. Hobbs. Low LDL cholesterol in individuals of african descent resulting from frequent nonsense mutations in *PCSK9*. *Nature Genetics*, 37(2):161–165, 2005.

26. J.C. Cohen, E. Boerwinkle, T.H. Mosley Jr, and H.H. Hobbs. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *New England Journal of Medicine*, 354(12):1264–1272, 2006.

27. B. Li and S.M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

28. B.E. Madsen and S.R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384, 2009.

29. C.R. King, P.J. Rathouz, and D.L. Nicolae. An evolutionary framework for association testing in resequencing studies. *PLoS Genetics*, 6(11):e1001202, 2010.

30. D.J. Liu and S.M. Leal. 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 Genetics*, 6(10):e1001156, 2010.

31. A.L. Price, G.V. Kryukov, P.I.W. de Bakker, S.M. Purcell, J. Staples, L-J. Wei, and S.R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838, 2010.

32. I. Ionita-Laza, J.D. Buxbaum, N.M. Laird, and C. Lange. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genetics*, 7(2):e1001289, 2011.

33. B.M. Neale, M.A. Rivas, B.F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S.M. Purcell, K. Roeder, and M.J. Daly. Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7(3):e1001322, 2011.

34. J.H. Sul, B. Han, D. He, and E. Eskin. An optimal weighted aggregated association test for identification of rare variants involved in common diseases. *Genetics*, 188(1):181–188, 2011.

35. M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

36. S. Basu and W. Pan. Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, 35(7):606–619, 2011.

37. N.O. Stitziel, A. Kiezun, S. Sunyaev, et al. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biology*, 12(9):227, 2011.

38. M. Ladouceur, Z. Dastani, Y.S. Aulchenko, C.M.T. Greenwood, and J. B. Richards. The empirical power of rare variant association methods: Results from sanger sequencing in 1,998 individuals. *PLoS Genetics*, 8(2):e1002496, 2012.

39. C.T. Johansen, J. Wang, M.B. Lanktree, H. Cao, A.D. McIntyre, M.R. Ban, R.A. Martins, B.A. Kennedy, R.G. Hassell, M.E. Visser, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature Genetics*, 42(8):684–687, 2010.

40. M.A. Rivas, M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C.K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, N. Burtt, et al. Deep resequencing of gwas loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genetics*, 43(11):1066–1073, 2011.

41. A. Bonnefond, N. Clément, K. Fawcett, L. Yengo, E. Vaillant, J.L. Guillaume, A. Dechaume, F. Payne, R. Roussel, S. Czernichow, et al. Rare *MTNR1B* variants impairing melatonin receptor 1b function contribute to type 2 diabetes. *Nature Genetics*, 44(3):297–301, 2012.

42. J.C. Cohen, R.S. Kiss, A. Pertsemlidis, Y.L. Marcel, R. McPherson, and H.H. Hobbs. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, 305(5685):869–872, 2004.

43. S. Romeo, W. Yin, J. Kozlitina, L.A. Pennacchio, E. Boerwinkle, H.H. Hobbs, and J.C. Cohen. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *The Journal of Clinical Investigation*, 119(1):70–79, 2009.

44. D. Diogo, F. Kurreeman, E.A. Stahl, K.P. Liao, N. Gupta, J.D. Greenberg, M.A. Rivas, B. Hickey, J. Flannick, B. Thomson, et al. Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from gwass contribute to risk of rheumatoid arthritis. *The American Journal of Human Genetics*, 92(1):15–27, 2013.

45. W. Ji, J.N. Foo, B.J. O'Roak, H. Zhao, M.G. Larson, D.B. Simon, C. Newton-Cheh, et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature Genetics*, 40(5):592–599, 2008.

46. I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, and S.R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, 2010.

47. P.C. Ng and S. Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003.

48. J.M. Schwarz, C. Rödelsperger, M. Schuelke, and D. Seelow. Mutationtaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, 7(8):575–576, 2010.

49. E.S. Lander and D. Botstein. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–199, 1989.

50. G.E. Uhlenbeck and L.S. Ornstein. On the theory of the brownian motion. *Physical Review*, 36(5):823–841, 1930.

51. M.R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes*. Springer Berlin Heidelberg New York, 1982.

52. D. Slepian. First passage time for a particular gaussian process. *The Annals of Mathematical Statistics*, 32(2):610–612, 1961.

53. G.V. Kryukov, L.A. Pennacchio, and S.R. Sunyaev. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *The American Journal of Human Genetics*, 80(4):727–739, 2007.

54. G.V. Kryukov, A. Shpunt, J.A. Stamatoyannopoulos, and S.R. Sunyaev. Power of deep, all-exon resequencing for discovery of human trait genes. *Proceedings of the National Academy of Sciences*, 106(10):3871–3876, 2009.