1 Supplementary Material. Site-Frequency and Binomial Sampling.

Let us elaborate a bit more on the process by which the F_j were computed. The present and ancestral terms are clearly distinguished, and explicit formulas for binomially sampled Site-Frequency are given in the form that enables quick computations necessary for thorough investigation of the Likelihood surface.

1.1 Site-frequency Expectation at Present.

Let us predict how the site frequencies are expected to be distributed in the population today. In the following, μ is the per-site mutation rate. Time is measured in units of generations and has its origin at the event of a bottleneck sampling.

The Echo of Ancestral Mutations. For equilibrium ancestral population of size N_1, the expectation of the number of segregating sites with j representations of the mutant nucleotide is given by $(4N_1\mu)/j$ (e.g. (R1)). Binomial sampling of $N_b \leq N_1$ individuals from such a population at the bottleneck shifts the frequencies of segregating sites downwards, but does not change the absolute expected number of segregating sites with a given number of representations. For the sake of completeness, this well-known result is also derived in Supplementary material. At time τ, the expected number of ancestral sites with frequency q is

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

Accumulated Modern Mutations. At each time step (generation), an influx of new $2N(t)\mu$ singleton mutations is introduced. The cumulative contribution of these mutations to the site-frequency expectation at time τ can be approximated by the following

$$E_m(q,\tau) = \sum_{t=1}^{\tau} 2N_b\mu e^{\gamma t}\phi\left(q, \frac{1-e^{-(\tau-t)\gamma}}{\gamma} | \frac{1}{2N_b e^{\gamma t}}, N_b e^{\gamma t}\right),$$

where $2N_b e^{\gamma t}$ is the number of singleton mutations, introduced at time t.

1.2 Binomial Sampling.

In a sample of N_s chromosomes of length 63000 bps, we expect to observe $63000 \cdot F_j$ segregating sites with polymorphic multiplicity j, where $F_j =$

$F_j^{(a)} + F_j^{(m)}$, the sum of ancestral and modern segregating sites. Binomial sampling of a function $f(q)$ gives

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

For the ancestral and modern sites, respectively

$$\begin{aligned}
F_j^{(a)} &= 8\frac{N_1}{N_b}\mu \cdot \sum_{i=1}^{\infty} \frac{2i+1}{i(i+1)} \cdot e^{-\frac{i(i+1)}{4N_b} \cdot \frac{1-e^{-\gamma\tau}}{\gamma}} \cdot K_i \cdot G_{ij}, \\
F_j^{(m)} &= 4\mu \cdot \sum_{i=1}^{\infty} \frac{2i+1}{i(i+1)} \cdot T_i \cdot G_{ij}, \\
K_i &\equiv \sum_{k=1}^{2N_b-1} \left(1 - \frac{k}{2N_b}\right) C_{i-1}^{3/2}\left(1 - \frac{k}{N_b}\right), \\
T_i &\equiv \sum_{t=1}^{\tau} \left(1 - \frac{1}{2N_b e^{\gamma t}}\right) \cdot C_{i-1}^{3/2}\left(1 - \frac{1}{N_b e^{\gamma t}}\right) \cdot e^{-\frac{i(i+1)}{4N_b} \cdot \frac{e^{-t\gamma} - e^{-\tau\gamma}}{\gamma}}, \\
G_{ij} &\equiv \int_0^1 \binom{N_s}{j} q^j (1-q)^{N_s - j} C_{i-1}^{3/2}(1-2q) dq.
\end{aligned}$$

Note that the formulas above are given such that G_{ij} does not depend on the parameters of the demographic model and therefore can be precomputed. In such way, computing a point in the space of the four demographic model parameters is reduced to a computation of vectors T and K, followed by matrix multiplication. An additional benefit of such an approach, as opposed to a straightforward binomial sampling of the whole formula is that the convergence of the sums over the Gegenbauer polynomials is much quicker. To insure accuracy of our estimates, we summed 400 terms in the series, whereas prior analysis and numerical experiments indicated that 200 would have been enough for convergence.

References

[1] W. Ewens, "Mathematical Population Genetics I. Theoretical Introduction," 2nd edition, Springer 2004.

[2] M. Kimura, "Solution of a process of random genetic drift with a continuous model," Proc. Natl. Acad. Sci. USA 41: 144–150.

[3] M. Kimura, "Diffusion Models in Population Genetics," J. Appl. Prob., Vol. 1, No. 2, 177–232. Dec., 1964.

2. Supplementary Material. Methods and Figure Legends.

Estimation of the combined 'equivalent' number of neutral sites. Synonymous and non-coding polymorphisms have very similar spectra of minor allele frequency and were pooled for analysis as neutral standard. Only sites that have been successfully sequenced in at least 1400 chromosomes were retained for further analysis. Using relatively frequent polymorphisms (minor allele detected 10 or more times) we estimated that application of such 1400 cut-off leads to a ~10% reduction in the effective length of sequenced coding region. (44 out of 49 common SNPs were successfully sequenced in 1400 or more chromosomes).

Using context-dependent mutation matrix described in Asthana et al. (S1) we calculated that 2.15 *de novo* missense mutations correspond to 1 *de novo* synonymous mutation in genes from an experimental re-sequencing dataset. It means that only 1 out of 3.15 *de novo* mutations in the coding regions is synonymous. Total number of sites in coding regions sequenced by Ahituv et al. (8) is 60,373. Correspondingly, the effective length of synonymous mutations in the resequencing dataset that we used is approximately 17 kilobases: L_{syn} = 60,373*0.90*(1/3.15)=17250

Total number of SNPs identified in the non-coding regions of the experimental dataset is 2.66 times more than the number of synonymous coding SNPs (473 vs 178). Thus, the total number of neutral sites we should model to represent pooled non-coding and synonymous variation is equal approximately 63 kilobases: 17250 + 17250*2.66 = 63,088

Estimation of the 'equivalent' number of sites all mutations at which leads to amino acid change. We calculated number of sites that should be simulated to model experimental data on missense substitutions as the difference between total length of sequenced coding region and the 'effective' number of synonymous sites: Lmissen= 60,373*0.90 – 17250 = 37,085

Gene length. The median length of 18418 non-redundant human RefSeq proteins is 423 a.a. and the average length is 560 a.a. (the largest isoform was counted for alternatively spliced genes). Thus, we used 1500 nucleotides as the length of “the average gene”.

Software. Estimation of the demographic history parameters was implemented as a MatLab script. The simulation of the evolving human population and simulation of the resequencing study were written in C++.

SOM References

S1 Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S (2007) Analysis of sequence conservation at nucleotide resolution. *PLoS Comput Biol* 3:e254.

SOM Figure legends

Table S1 Demographic histories

Table S2 Sequencing studies of MC4R gene in individuals with extreme values of body mass index. *) Number of lean individuals has been calculated as number of individuals below 15% BMI percentile from unbiased population of 4068 individuals (4068*0.15 = 610.2)

Figure S1 Agreement of allele frequency spectrum predicted by analytical approach with results of forward evolutionary simulation.

Figure S2 Agreement of the experimental allele frequency spectra with the spectra predicted by analytical approach.

Figure S3 Population history model – long term constant population size is followed by a bottleneck and subsequent exponential population growth. The model has four parameters and limited to the European population: N_1 – ancestral population size; N_b – bottleneck population size; N_f – final population size; τ– time of the population expansion since the bottleneck.

Figure S4 Predicted number of SNPs. (A) Predicted fractions of ancient and recent alleles among detected SNPs (B) Predicted number of SNPs to be detected in 1400 sequenced chromosomes calculated using 1) maximum likelihood demographic history based on *Ahituv et al.* dataset and 2) joined maximum likelihood demographic history based on *Ahituv et al.* and SeattleSNPs dataset

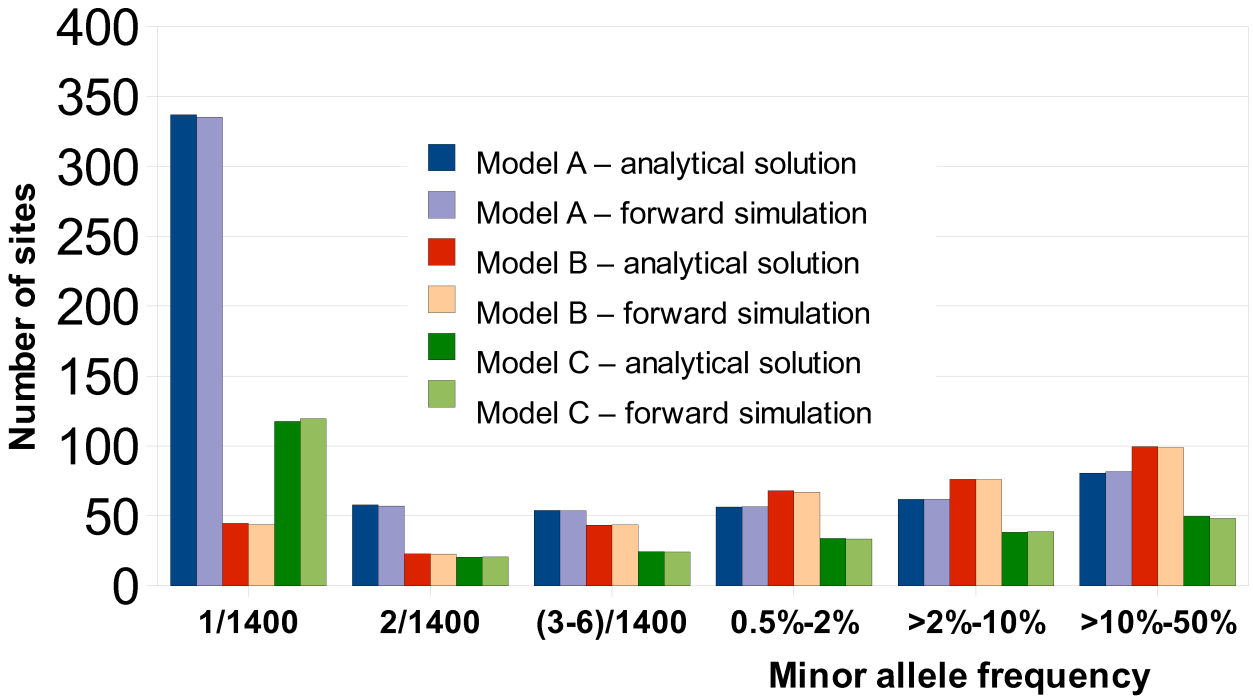
Figure S5 Distribution of selection coefficients associated with *de novo* missense mutations deduced for different demographic histories.

Figure S6 Dependence of power estimates on parameters. (*A*) Dependence of power on the assumed shift in quantitative trait distribution mean caused by a functional mutation. (*B*) Dependence of power on gene length. (*C*) Dependence of power on selection coefficient cut-off for functional mutations. Mutations with selection coefficient below cut-off value were assumed to have no effect on quantitative phenotype in our model.

Demographic history ID	Normalized Likelihood	Ancestral population size	Bottleneck population size	Modern population size	Duration of exponential growth (generations)	Description
1	1	8100	7900	900000	370	Maximum likelihood demography
2	0.17	7000	6200	800000	410	Smallest ancestral population size
3	0.18	8000	4800	500000	510	Smallest modern population size and longest growth time
4	0.16	8500	2000	800000	510	Smallest bottleneck population size and longest growth time
5	0.14	8500	7600	2000000	310	Largest modern population time and shortest growth time
6	0.22	9000	9000	1100000	360	Largest ancestral population size, no bottleneck

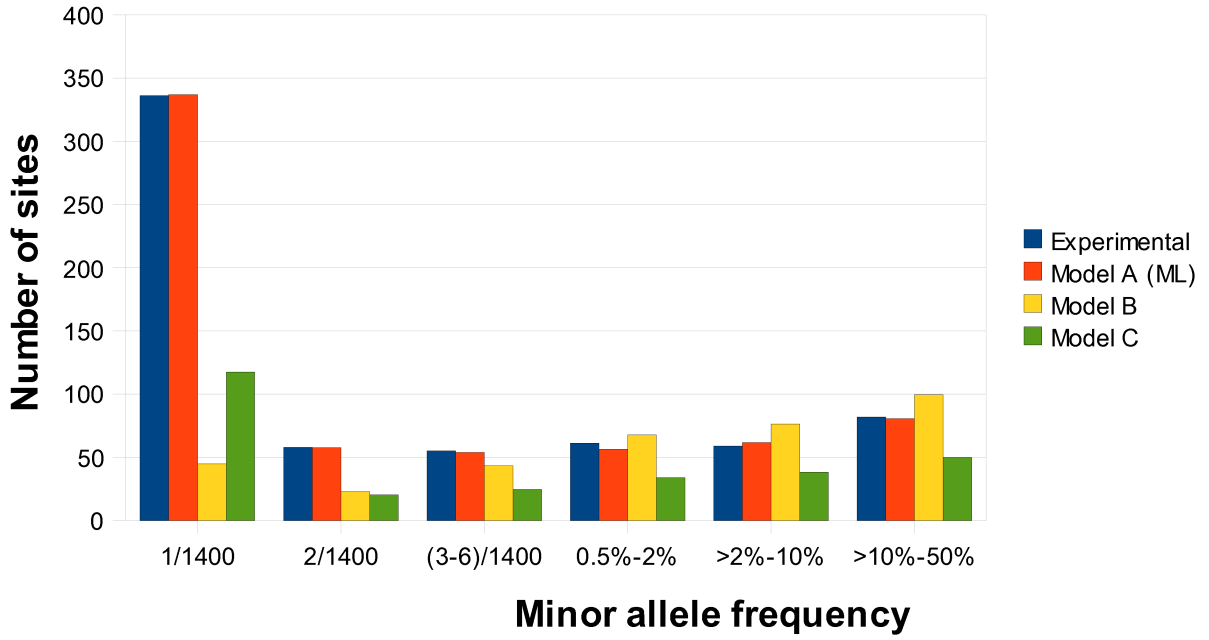
Study	Population	Number of obese individuals	Number of rare missenses in obese individuals	Number of lean individuals	Number of rare missenses in lean individuals
Ahituv et al., 2007	Whites, Canada	379	8	378	2
Hinney et al., 2006	Whites, Germany	1003	11	610*	1
Larsen et al., 2005	Whites, Denmark	750	9	706	0
Hinney et al., 2003	Whites, Germany	808	15	231	0
Total		2940	43	1925	3

Figure S1 Agreement of allele frequency spectrum predicted by analytical approach with results of forward evolutionary simulation.



Demographic history	*Ancestral population size*	*Bottleneck population size*	*Final population size*	*Duration of exponential growth*
Model A (best fit to experimental data)	8,100	7,900	900,000	370 generations
Model B (stable population size)	10,000	10,000	10,000	-
Model C	5,000	5,000	500,000	100 generations

Figure S2 Agreement of the experimental allele frequency spectra with the spectra predicted by analytical approach.



Demographic history	*Ancestral population size*	*Bottleneck population size*	*Final population size*	*Duration of exponential growth*
Model A (best fit to experimental data)	8,100	7,900	900,000	370 generations
Model B (stable population size)	10,000	10,000	10,000	-
Model C	5,000	5,000	500,000	100 generations

Figure S3. Population history model – long term constant population size is followed by a bottleneck and subsequent exponential population growth. The model has four parameters and limited to the European population: N_1 – ancestral population size; N_b – bottleneck population size; N_f – final population size; τ– time of the population expansion since the bottleneck.

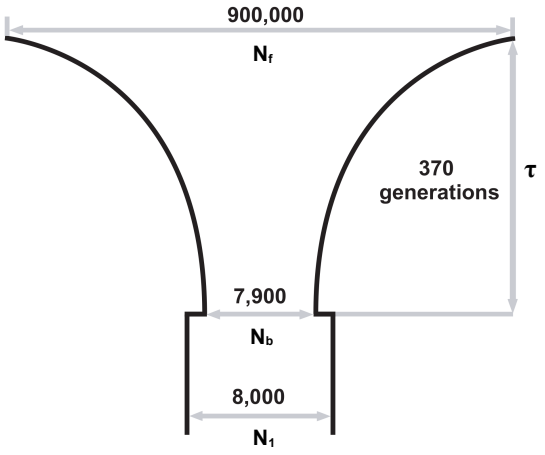


Figure S4 Predicted number of SNPs. (A) Predicted fractions of ancient and recent alleles among detected SNPs (B) Predicted number of SNPs to be detected in 1400 sequenced chromosomes calculated using 1) maximum likelihood demographic history based on *Ahituv et al.* dataset and 2) joined maximum likelihood demographic history based on *Ahituv et al.* and SeattleSNPs dataset

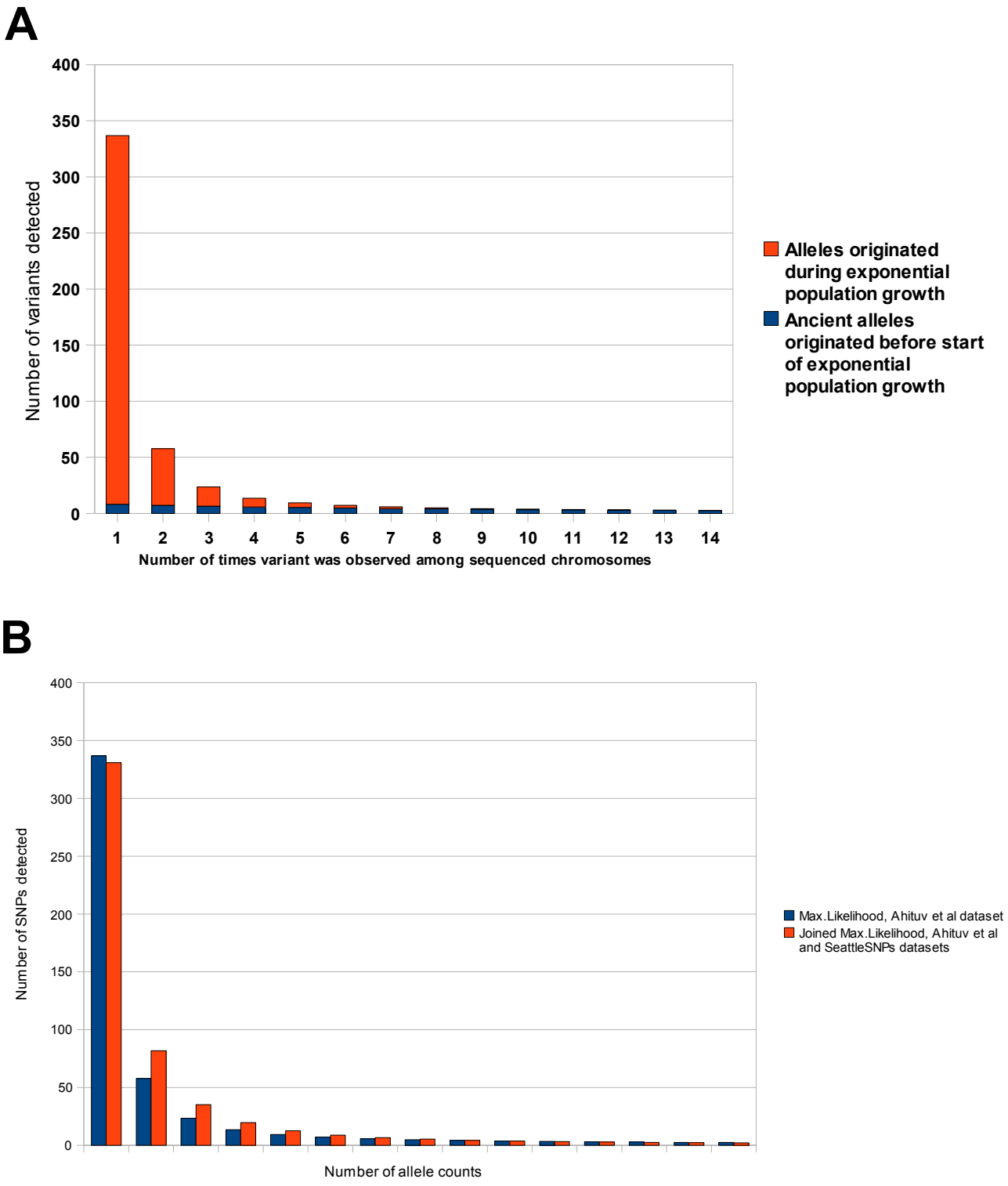


Figure S5 Distribution of selection coefficients associated with *de novo* missense mutations deduced for different demographic histories

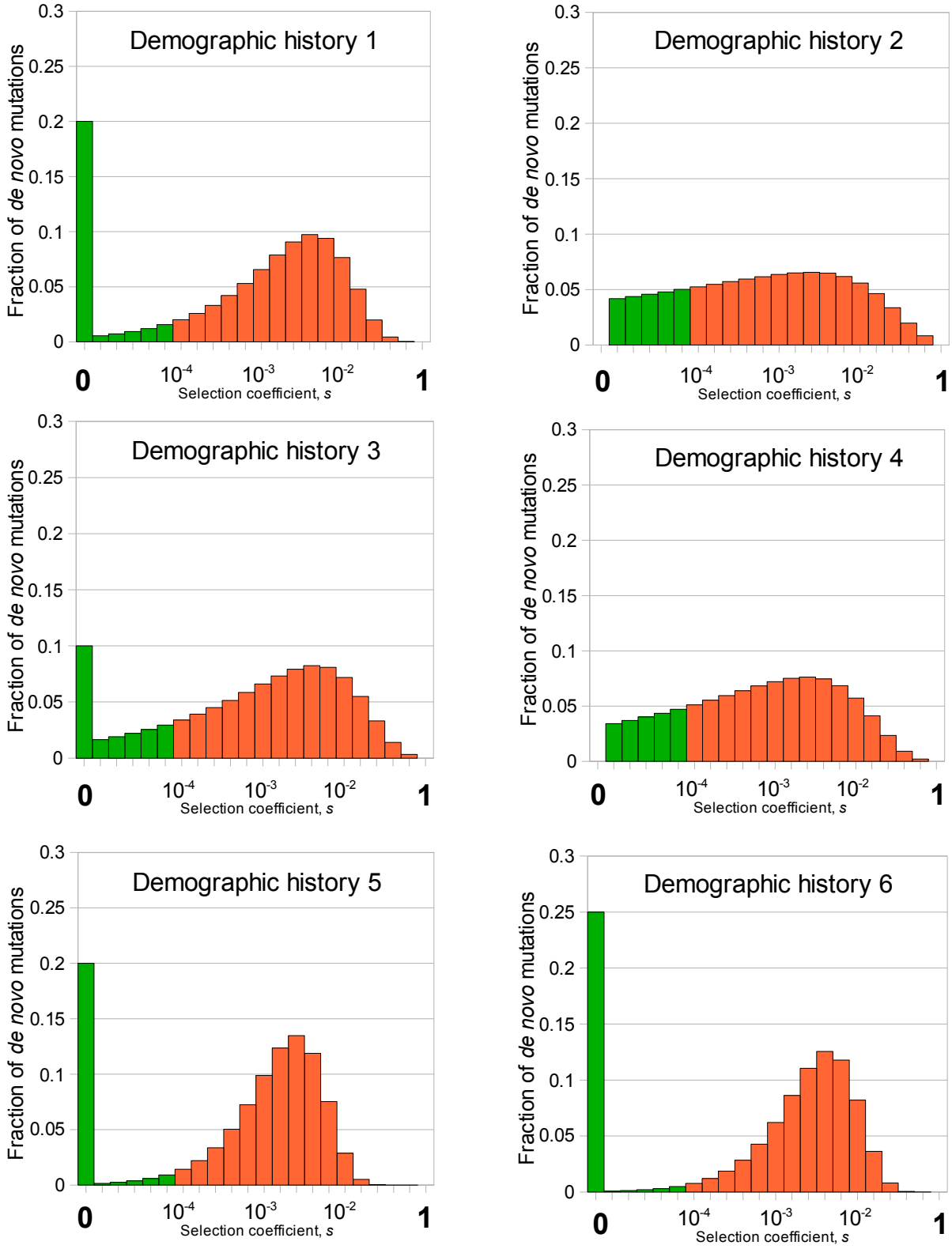


Figure S6 Dependence of power estimates on parameters. (*A*) Dependence of power on the assumed shift in quantitative trait distribution mean caused by a functional mutation. (*B*) Dependence of power on gene length. (*C*) Dependence of power on selection coefficient cut-off for functional mutations. Mutations with selection coefficient below cut-off value were assumed to have no effect on quantitative phenotype in our model.