The American Journal of Human Genetics, Volume *94*

Supplemental Data

Simulation of Finnish Population History, Guided by Empirical Genetic Data, to Assess Power of Rare-Variant Tests in Finland

Sophie R. Wang, Vineeta Agarwala, Jason Flannick, Charleston W.K. Chiang, David Altshuler, GoT2D Consortium, and Joel N. Hirschhorn

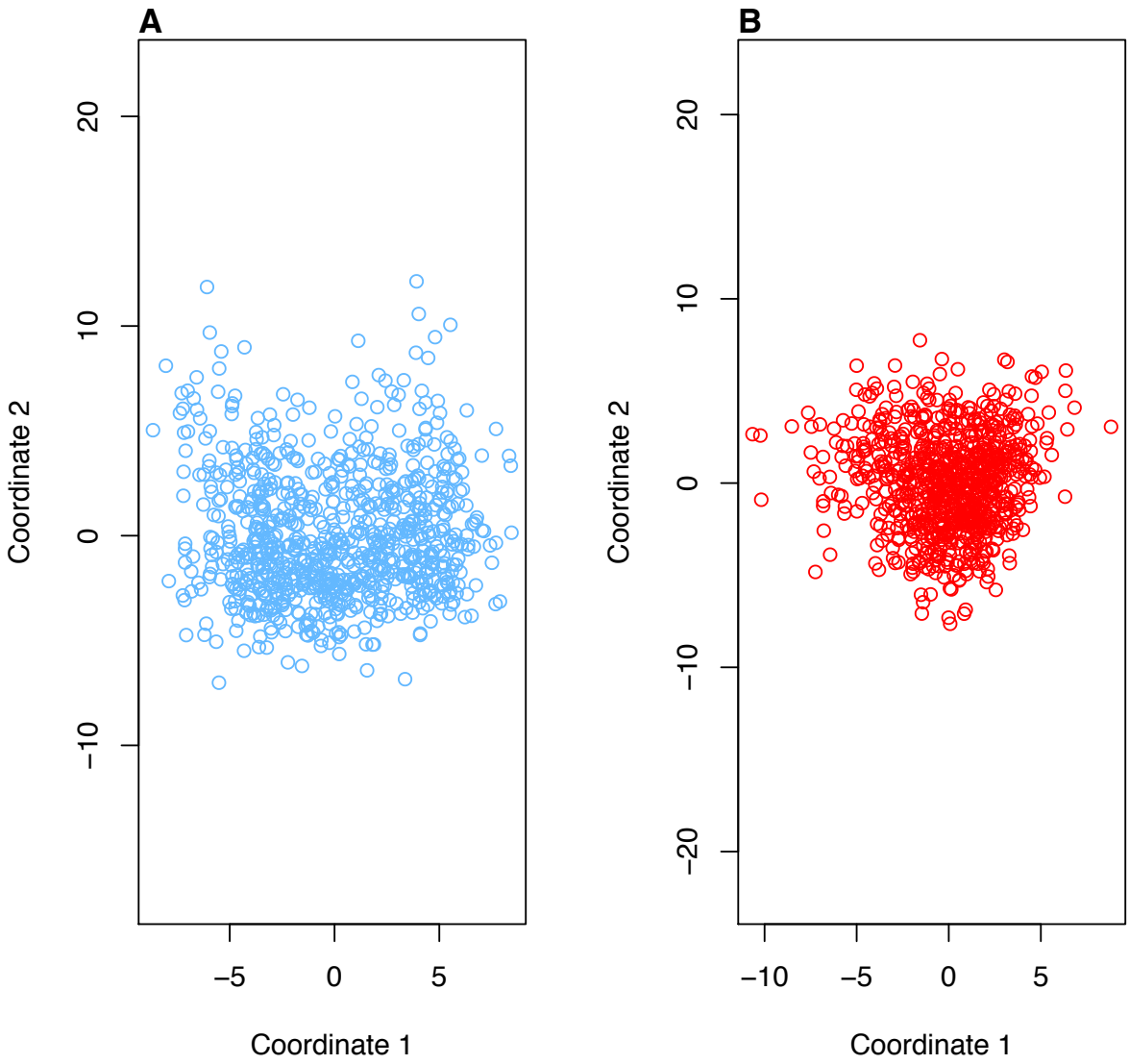

Figure S1. MDS plots of whole exome sequenced samples from GoT2D project. (A) Finns (n=843); (B) NFEs (n=820).

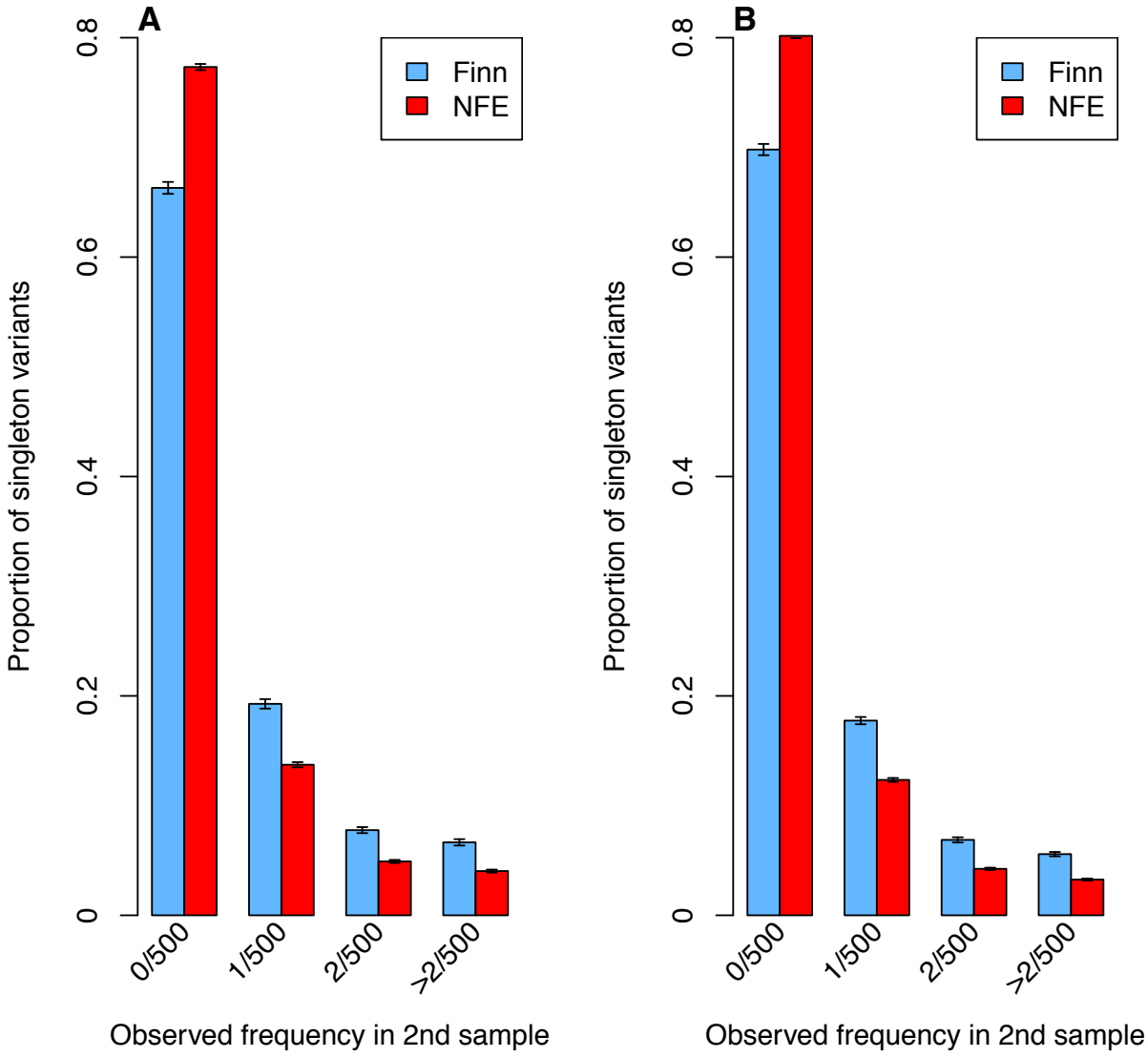

Figure S2. Singleton variants in a population of Finns are more likely to be seen again in another population of Finns. For the set of singleton variants ascertained from a random sample of 250 individuals, we assessed the proportion (y-axis) and the frequencies (x-axis) of these variants observed in a second sample of 250 individuals. The analysis was done in synonymous (A) and missense (B) variants separately.

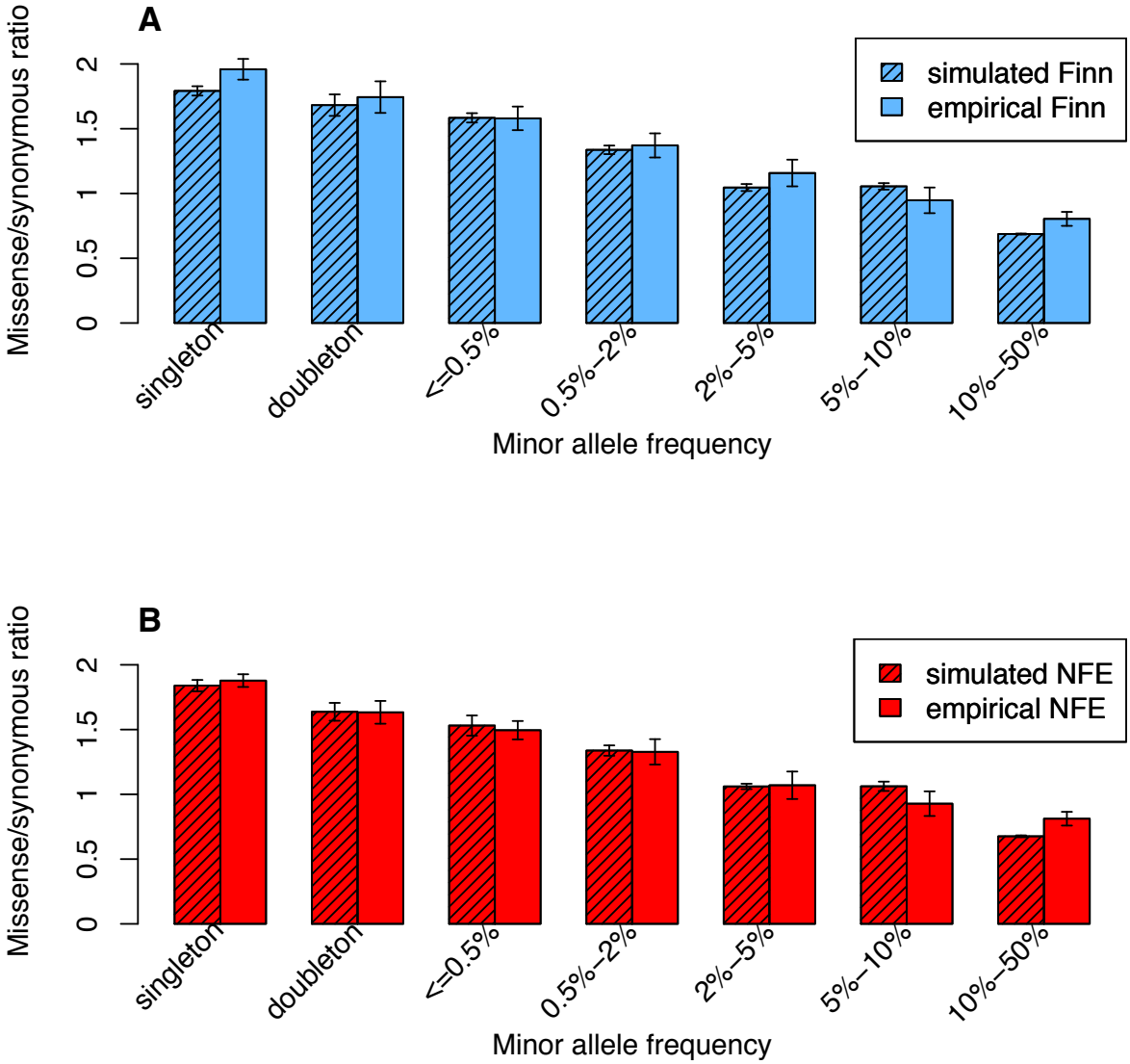

Figure S3. Agreement of empirical missense/synonymous ratios with the modeled ratios.

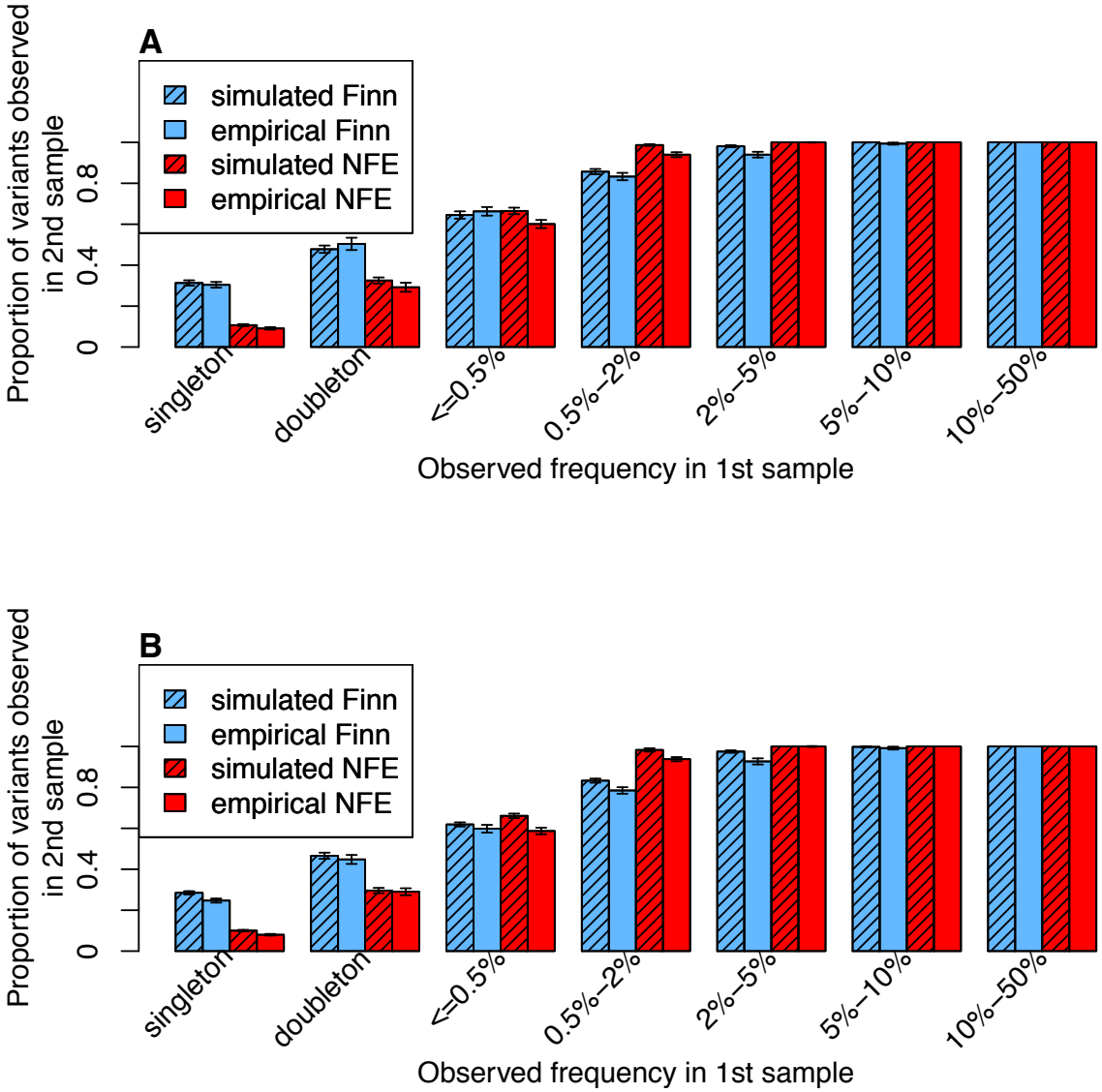

Figure S4. Allele sharing between the Finns and the NFEs, comparing simulated data and empirical data. For the set of variants ascertained from the first sample, we assessed their frequencies (x-aixs) in the first sample and the proportion (y-axis) of these variants observed in a second sample. For results in Finns, the first sample is 843 Finns and the second sample is 820 NFEs; for results in NFEs, the first sample is 820 NFEs and the second sample is 843 Finns. The analysis was done in synonymous variants (A) and missense variants (B) separately.

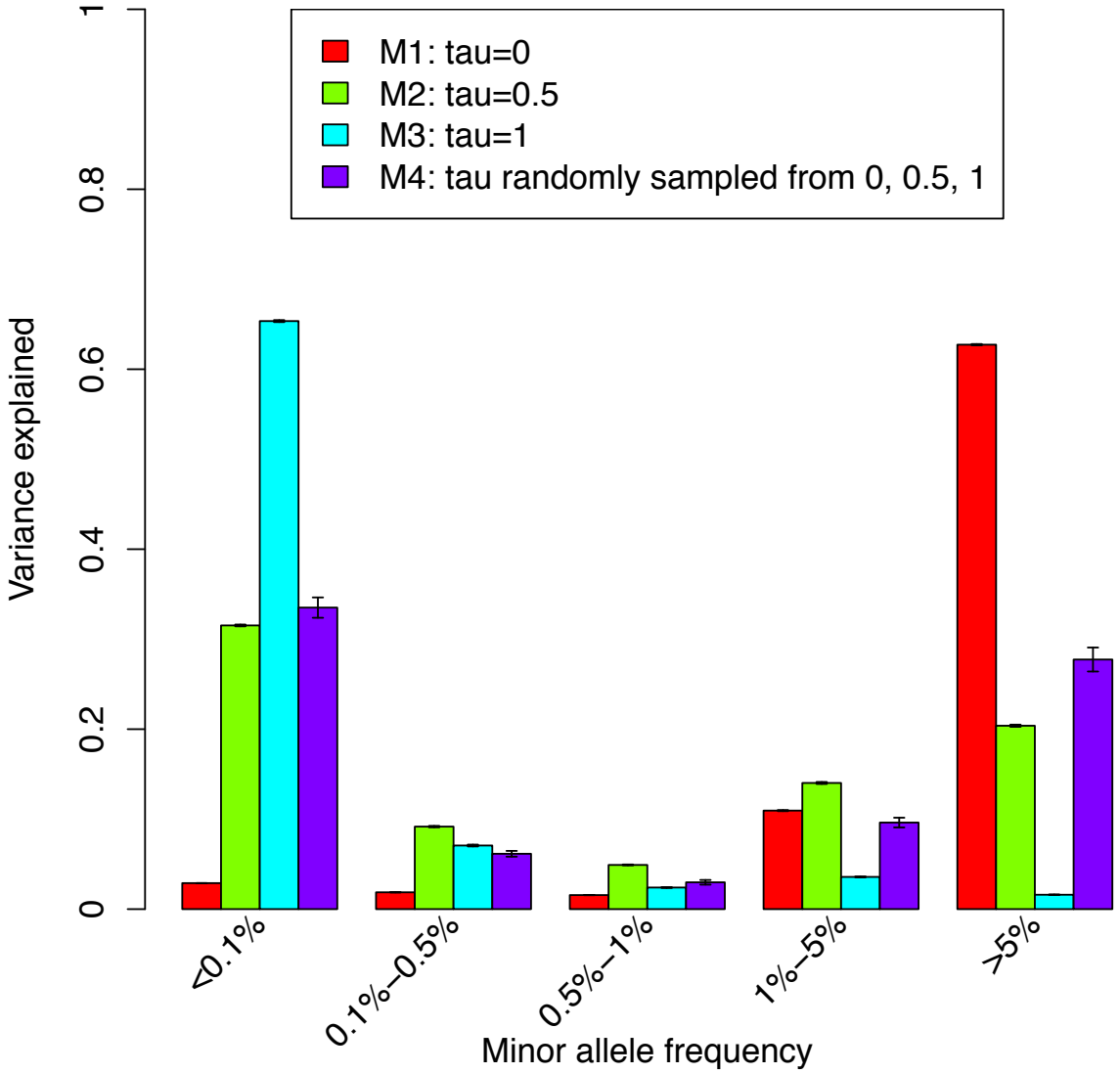

Figure S5A. Variance explained by variants within different frequency ranges under four different disease models.

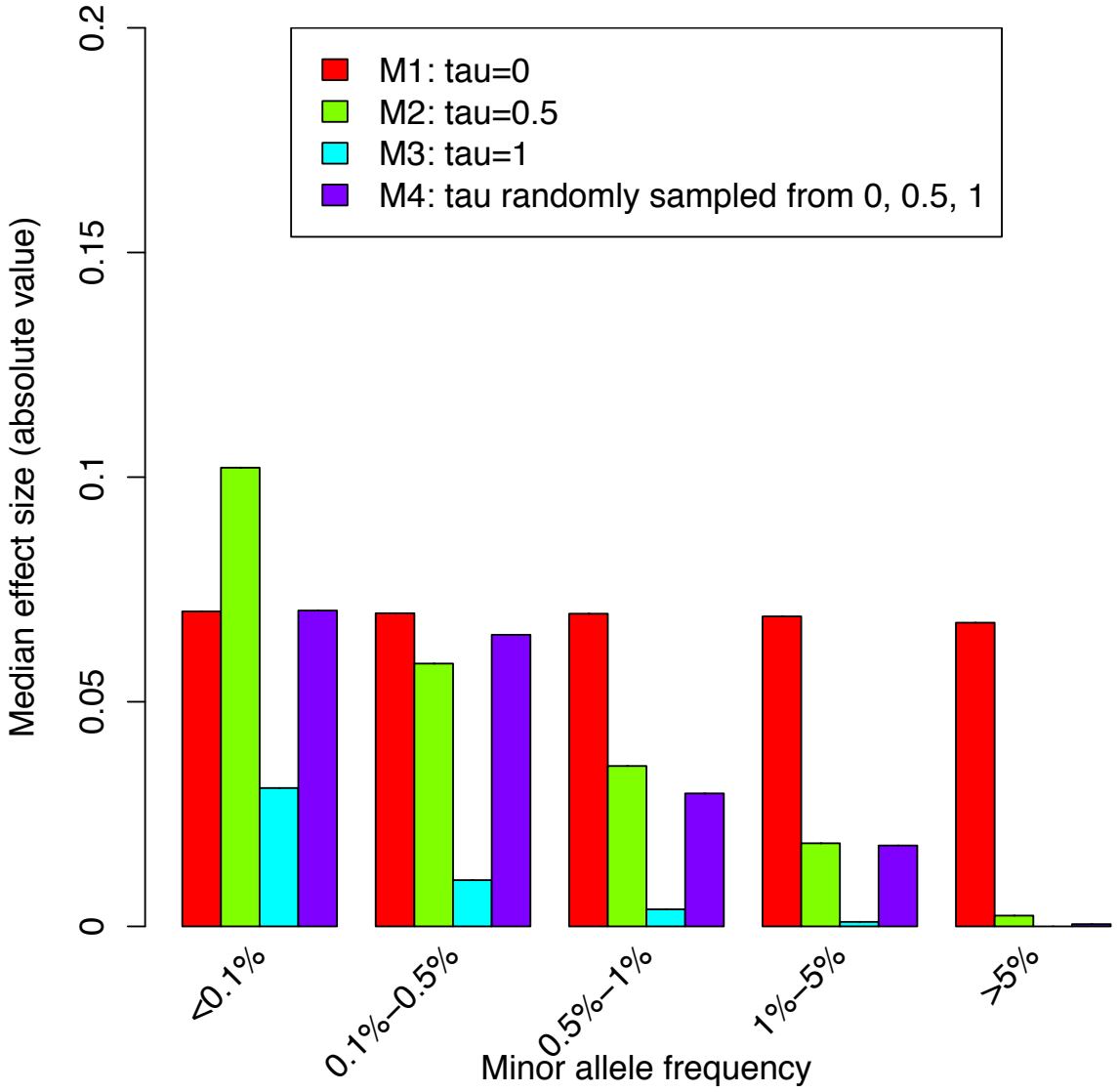

Figure S5B. Median effect size of variants within different frequency ranges under four different disease models. Absolute values of effect sizes are taken for the analysis. Rare variants contributed the most under M3 as shown in Figure S5A. However, the median effect size of rare variants under M3 is not as big as under M2 or M4 due to the wider distribution of effect sizes under M3.

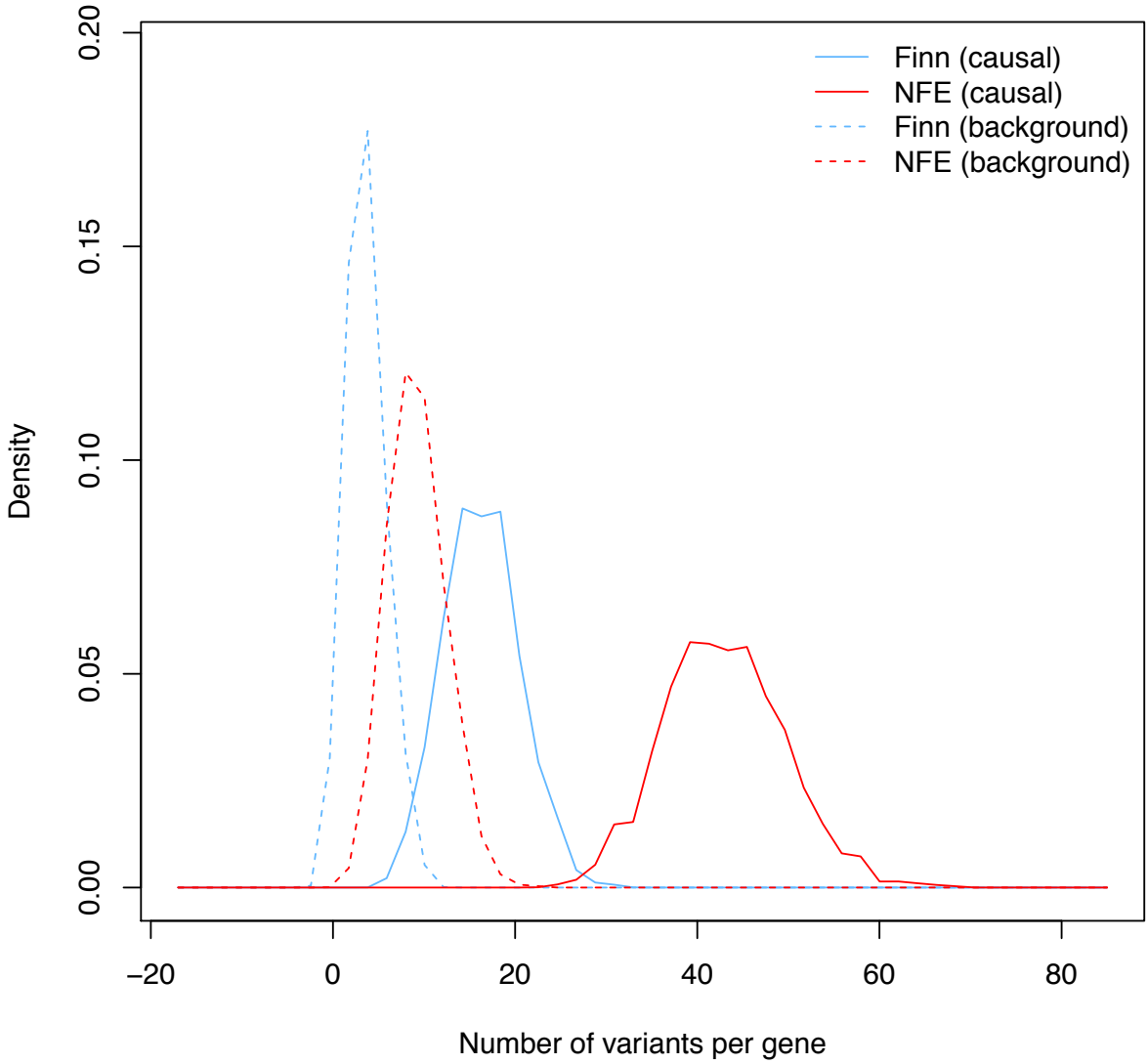

Figure S6. Number of causal variants (solid lines) or background variants (dashed lines) with MAF below 5% per gene, in either 30,000 Finns or 30,000 NFEs.

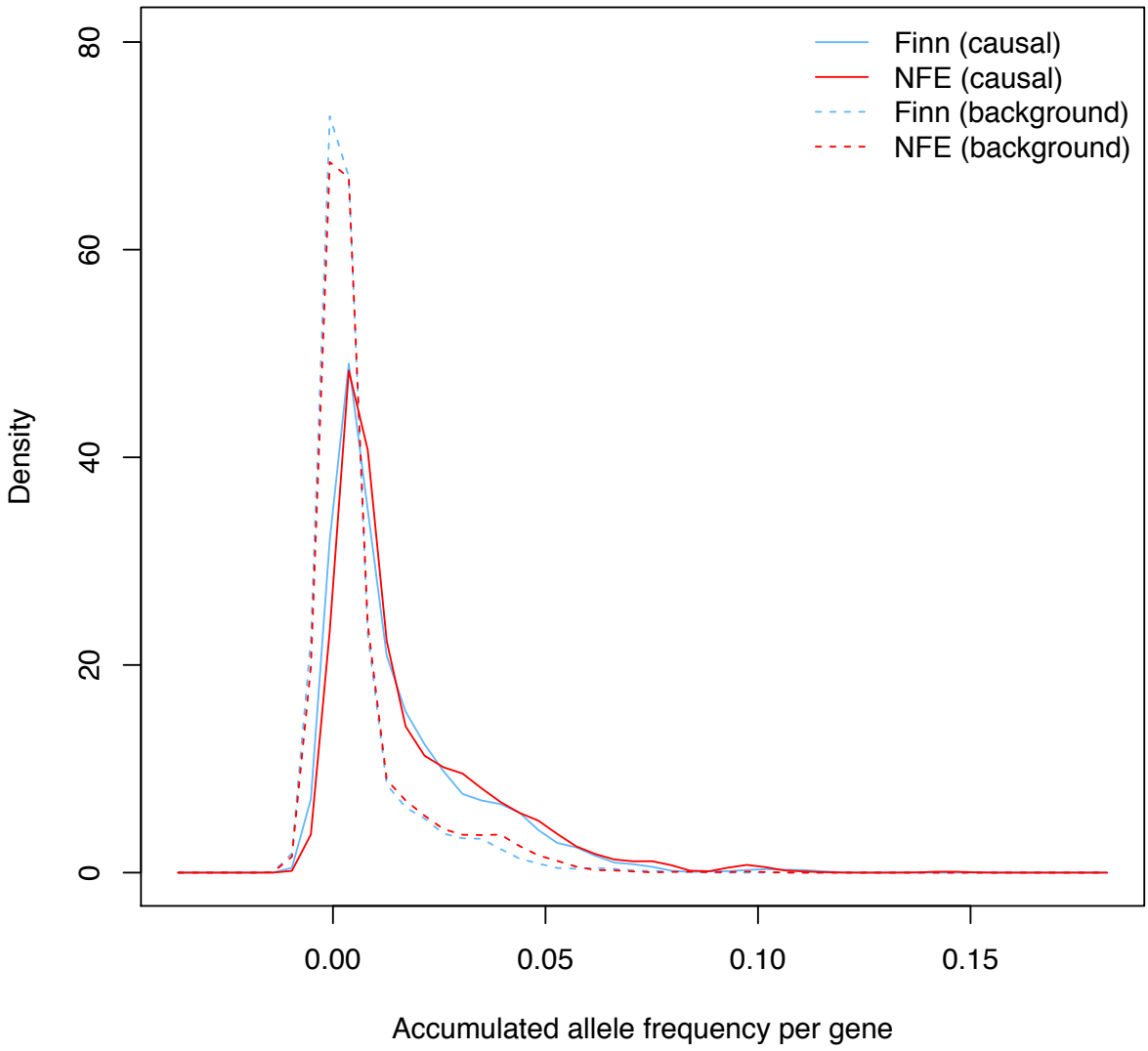

Figure S7. Accumulated allele frequency of causal variants (solid lines) or background variants (dashed lines) with MAF below 5% per gene, in either 30,000 Finns or 30,000 NFEs.

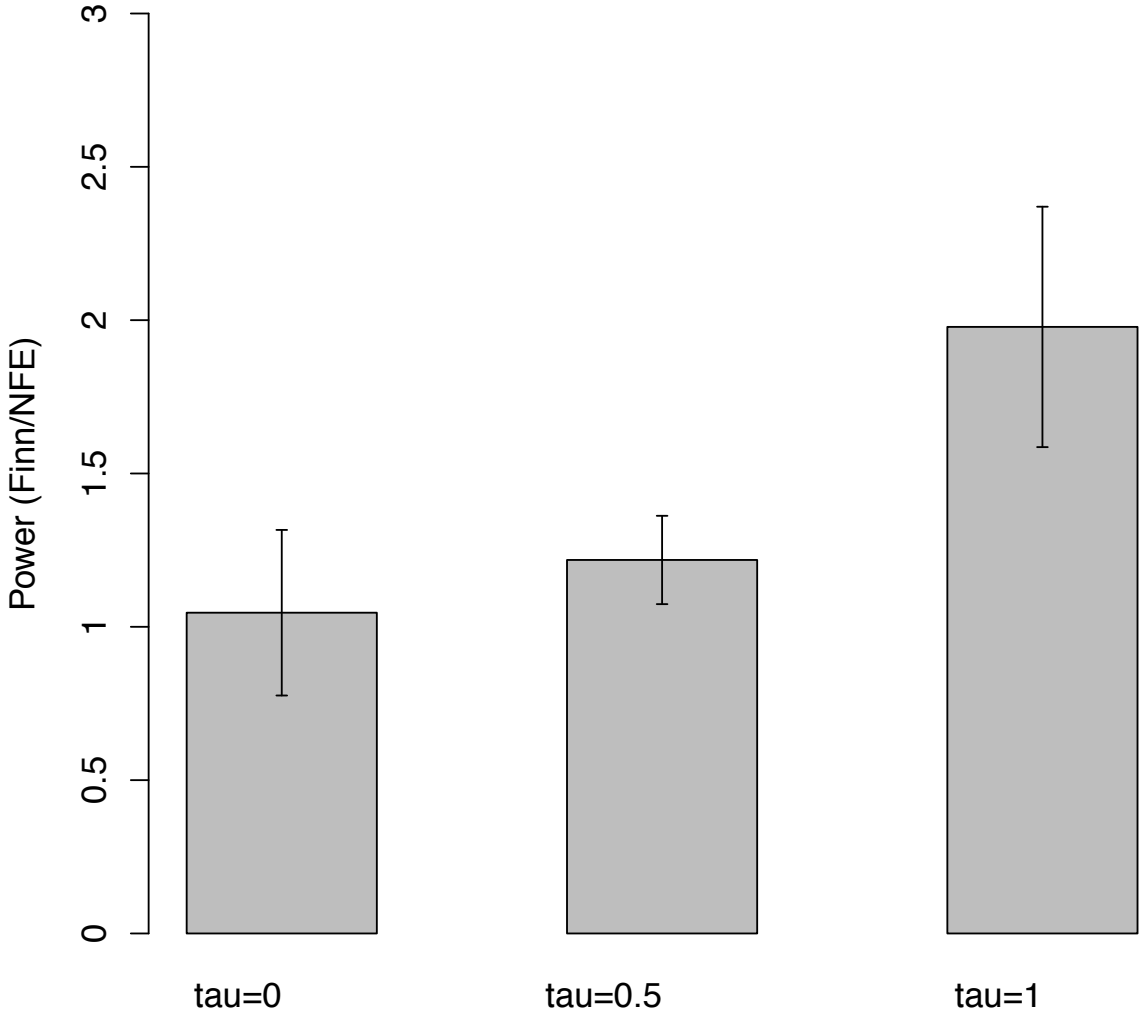

Figure S8. Power difference between using the Finns and the NFEs for genes of different τ values under M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene). Shown here is the result for SKAT-O test and the sample size is 30,000. The biggest power gain in the Finns is seen among genes with τ value of 1 (almost doubling in power, paired t-test p value < 0.01).

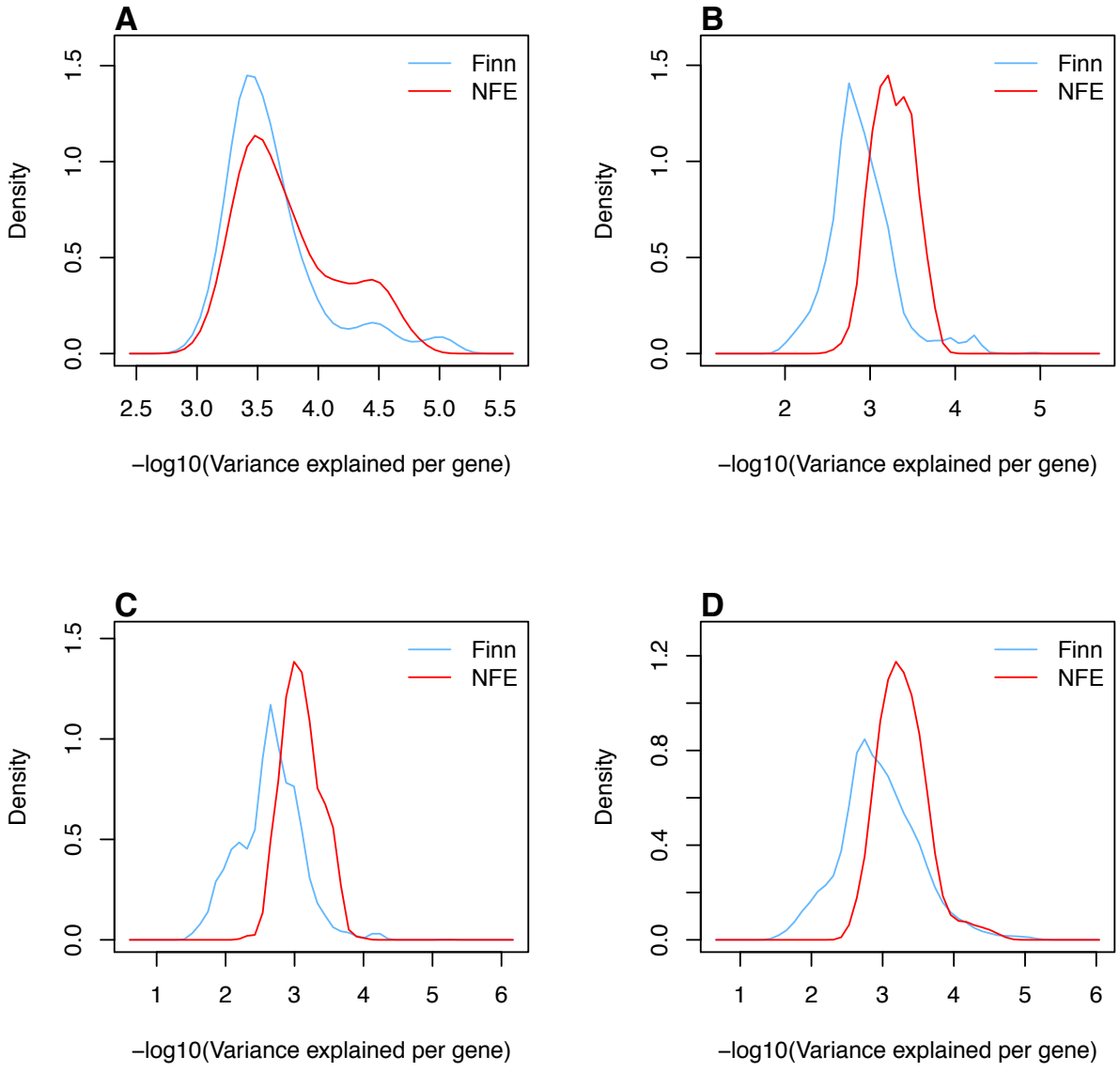

Figure S9. Distribution of variance explained per gene by variants with MAF below 5% under four different disease models in either 30,000 Finns or 30,000 NFEs, for genes detected in the Finns only. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

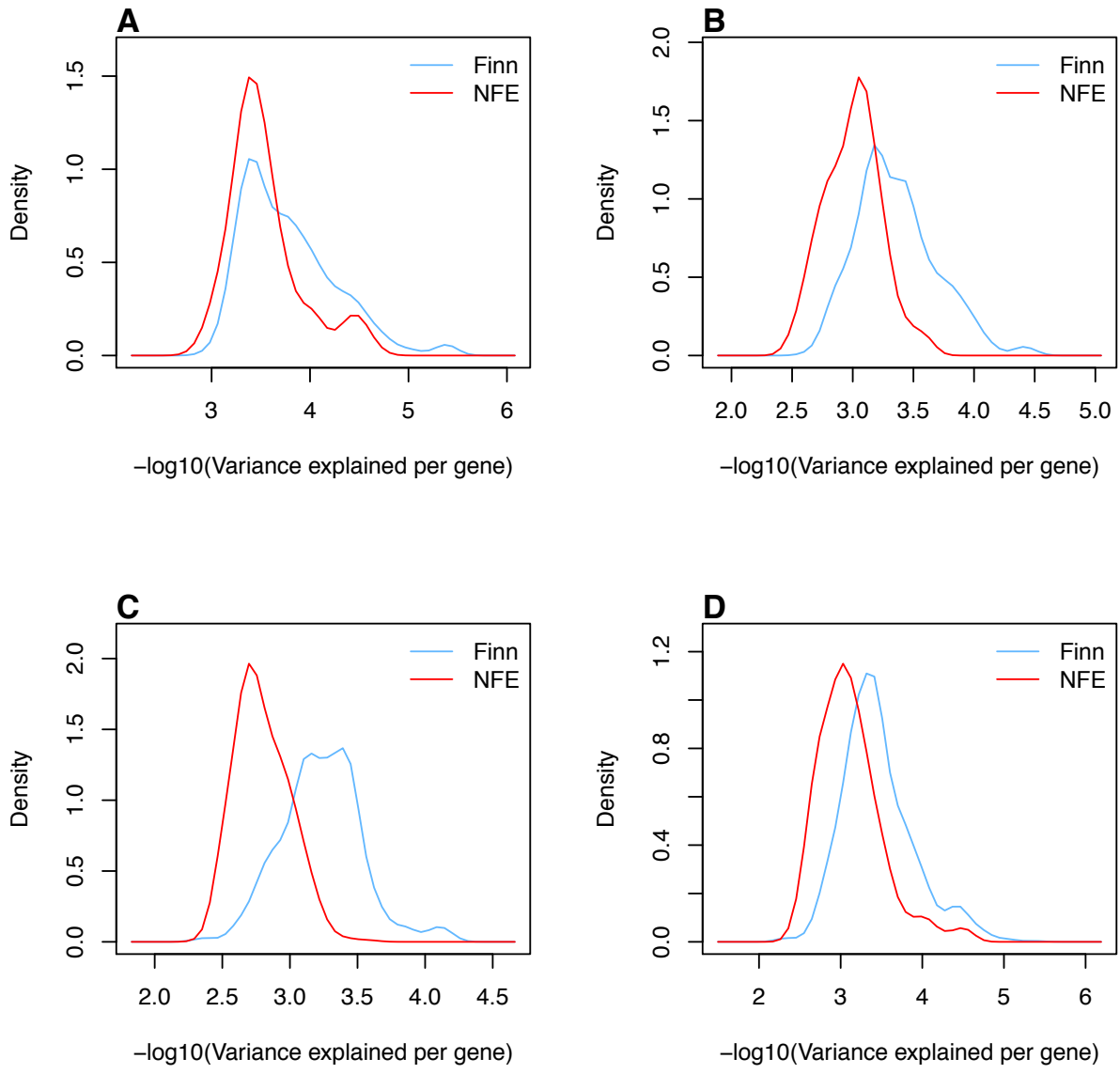

Figure S10. Distribution of variance explained per gene by variants with MAF below 5% under four different disease models, in either 30,000 Finns or 30,000 NFEs, for genes detected in the NFEs only. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

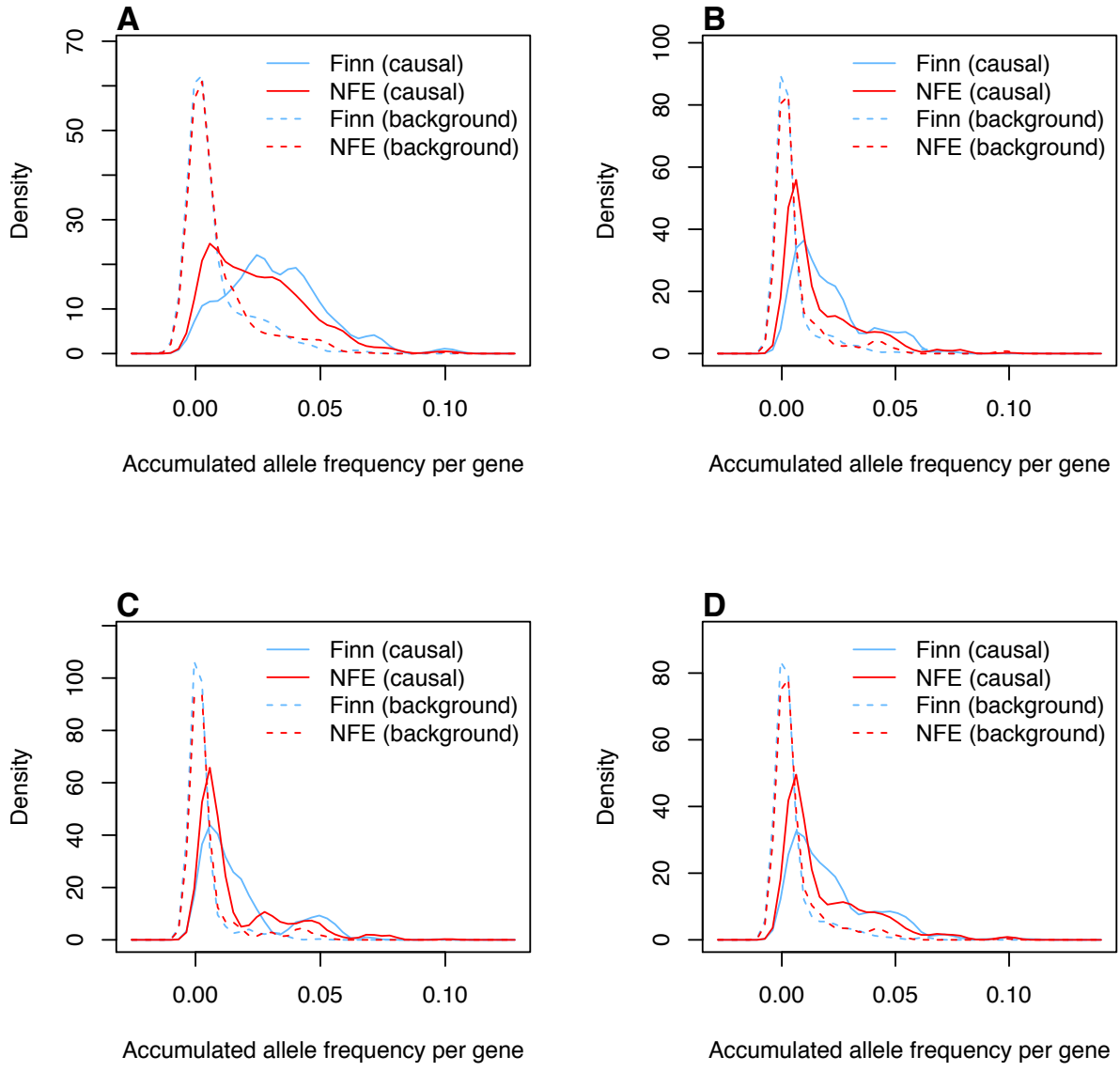

Figure S11. Accumulated allele frequency of causal variants (solid lines) or background variants (dashed lines) with MAF below 5% per gene under four different disease models, for genes detected in the Finns only. The distributions for causal variants in the Finns shift upwards compared to the NFEs. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

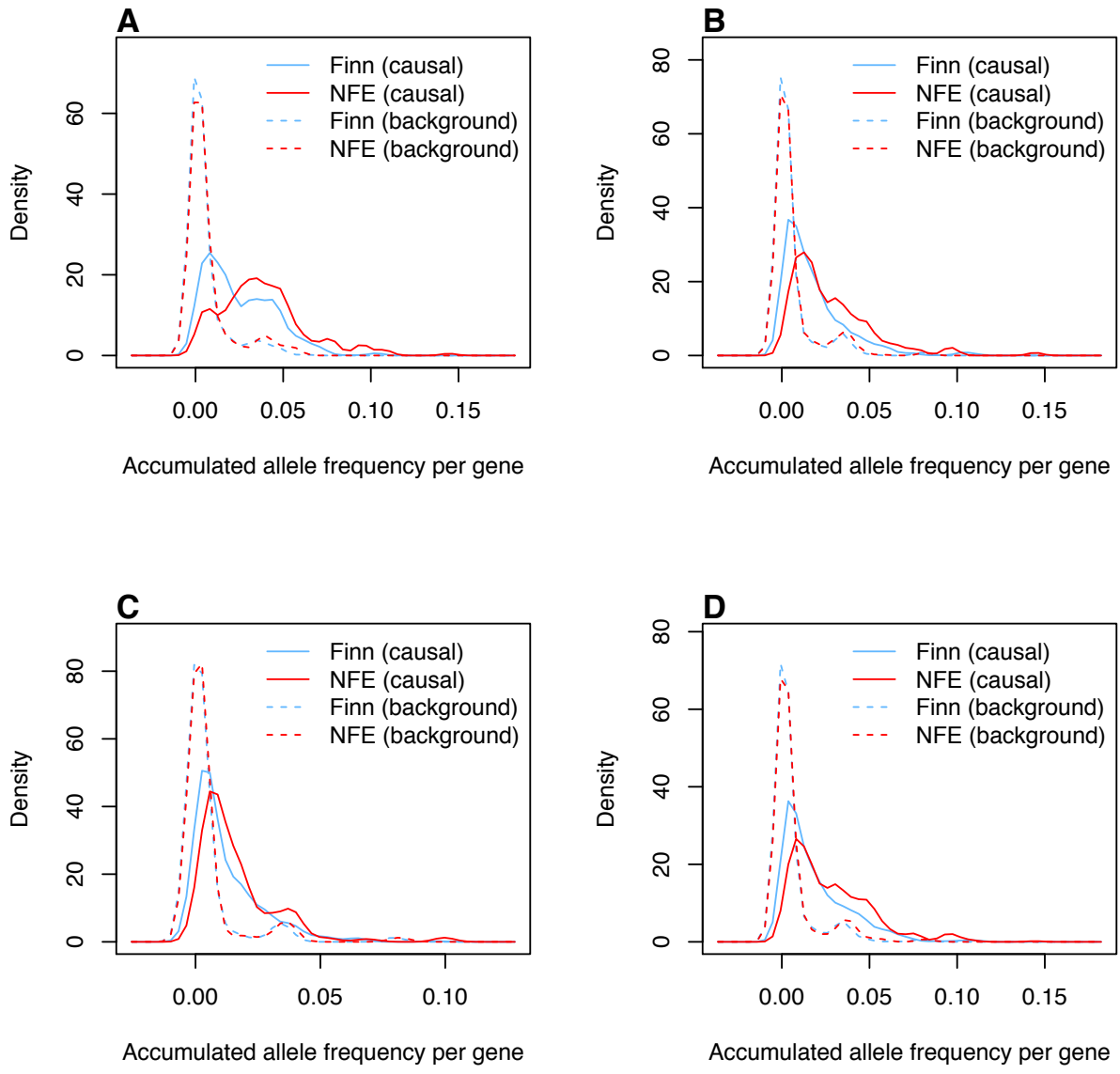

Figure S12. Accumulated allele frequency of causal variants (solid lines) or background variants (dashed lines) with MAF below 5% per gene under four different disease models, for genes detected in the NFEs only. The distributions for causal variants in the Finns shift downwards compared to the NFEs. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

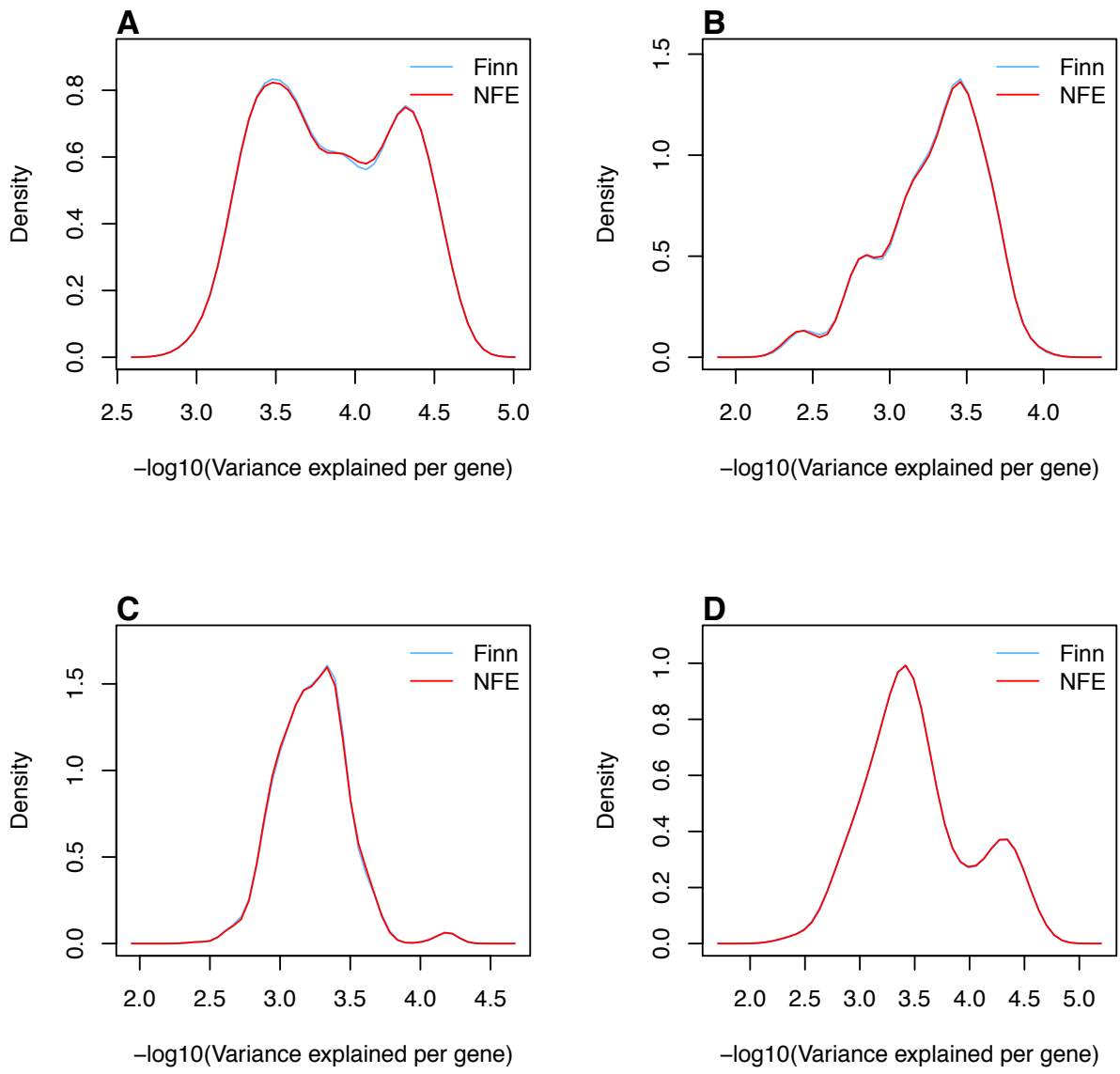

Figure S13. Distribution of variance explained per gene by variants with MAF below 5% under four different disease models, in either 30,000 Finns or 30,000 NFEs. The genes were sampled so as to match the variance explained in the Finns and the NFEs. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

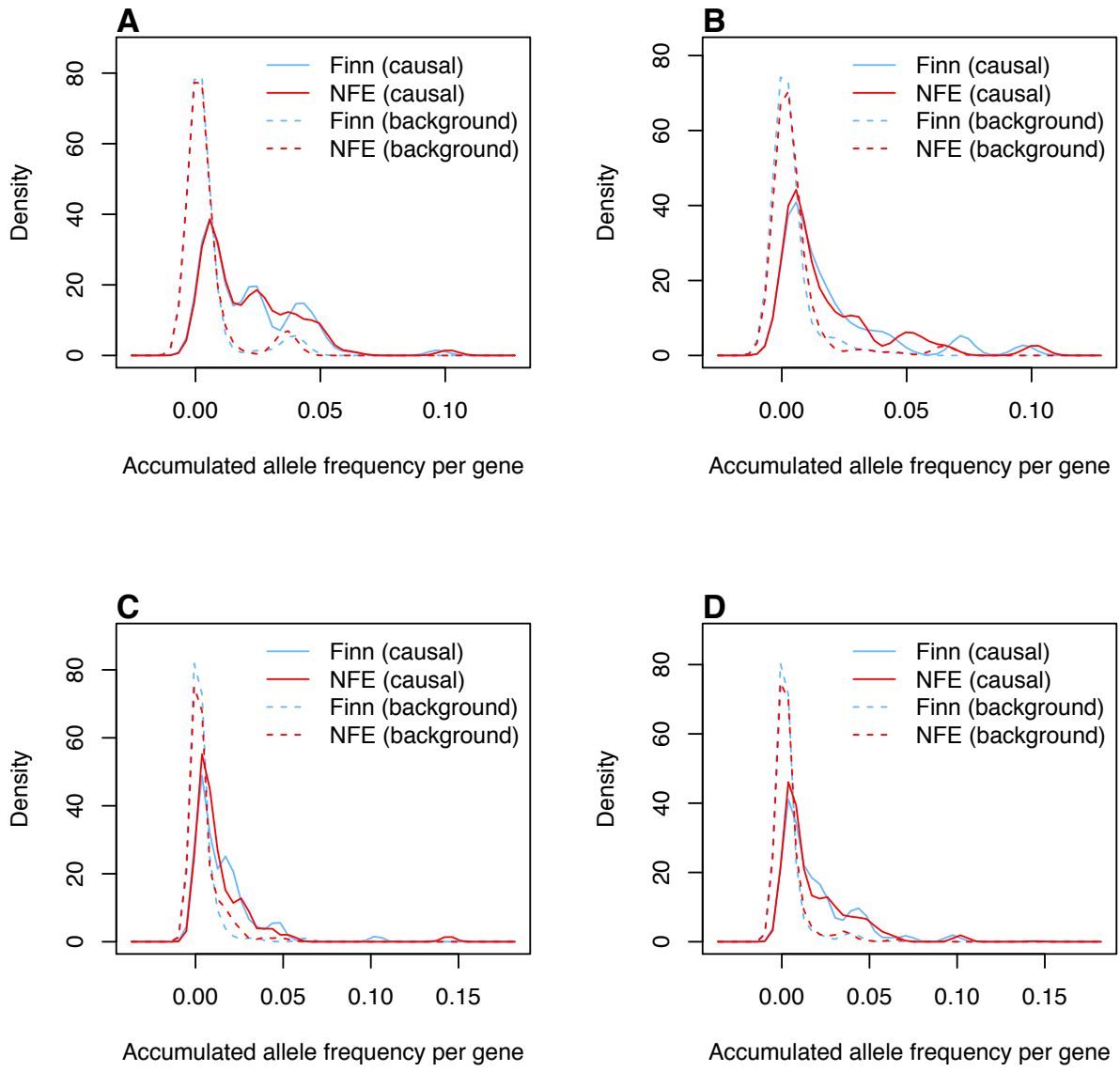

Figure S14. Accumulated allele frequency of causal variants (solid lines) or background variants (dashed lines) with MAF below 5% per gene under four different disease models. The genes were sampled so as to match the variance explained in the Finns and the NFEs. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

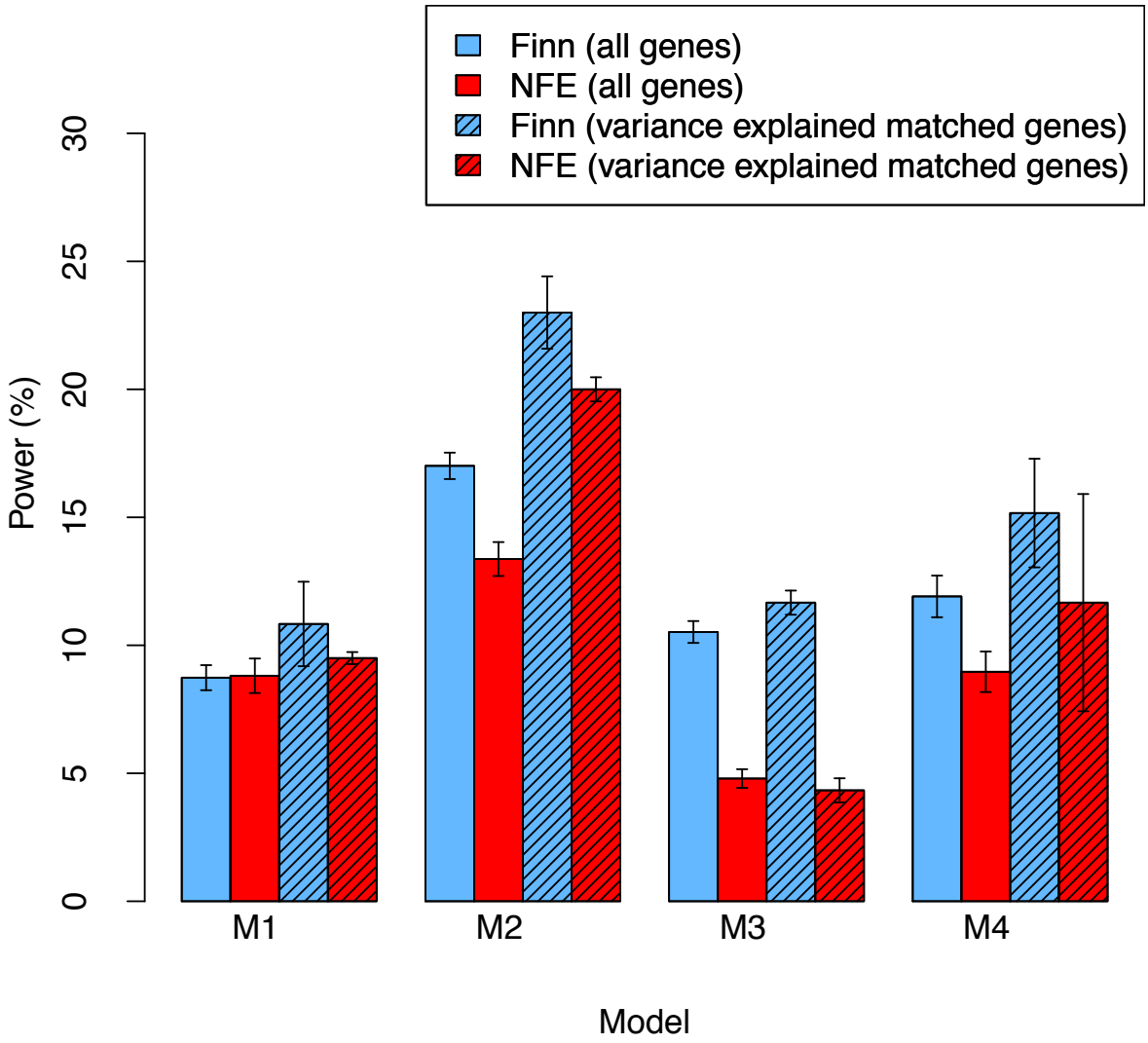

Figure S15. Power of SKAT-O test in 30,000 Finns or 30,000 NFEs under four different disease models, either for all genes, or for a set of genes sampled by matching the variance explained in the Finns and the NFEs.

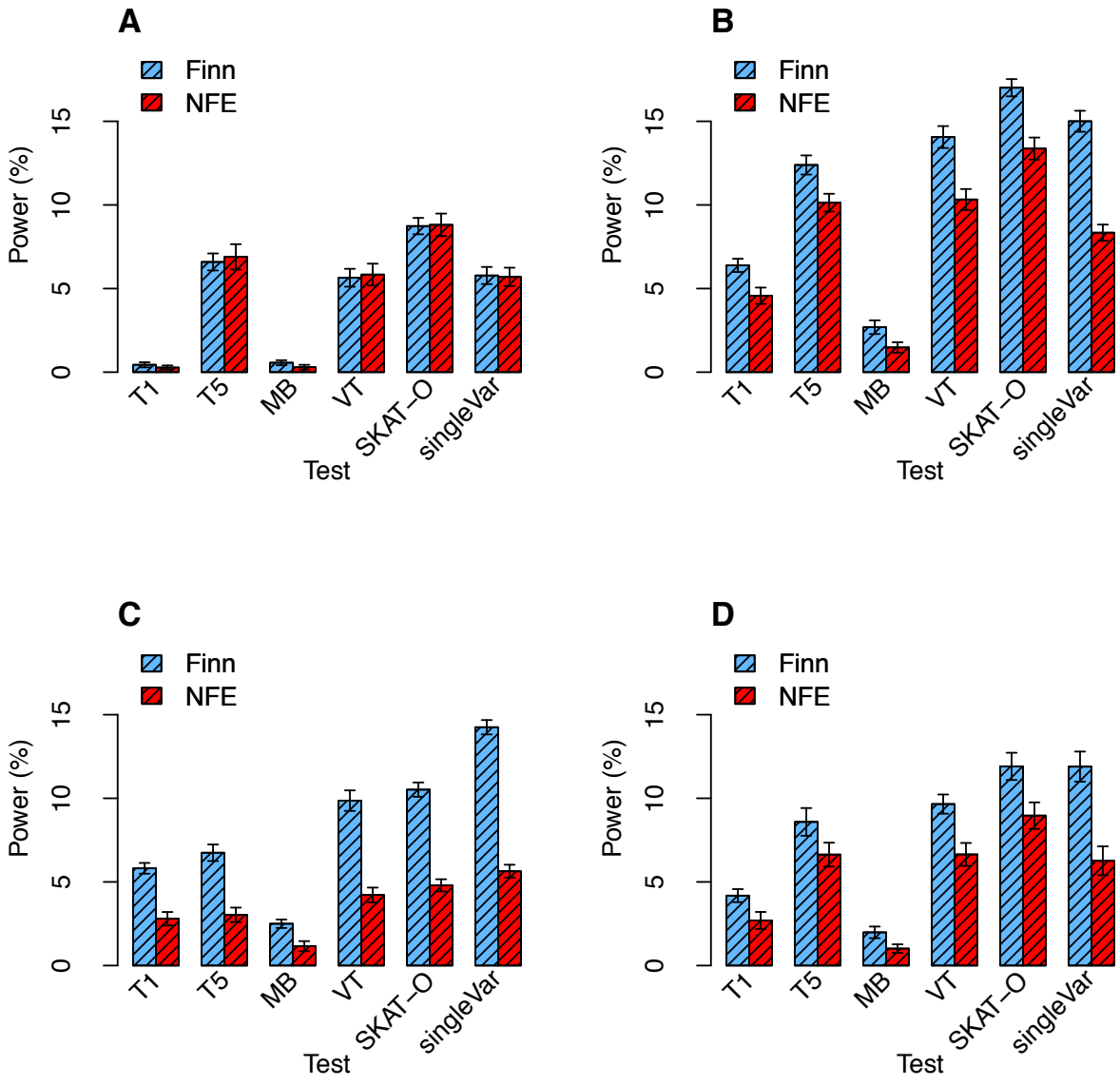

Figure S16. Power of exome sequencing studies in 30,000 Finns vs. 30,000 NFEs. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene). We simulated a quantitative trait (h2 = 80%) for which aggregated coding variation in 1,000 genes explains the total heritability. Models M1-4 were generated by varying the degree of coupling (τ) between a causal variant's phenotypic effect and the strength of purifying selection against that variant. We implemented five gene-based tests (T1, T5, MB, VT, SKAT-O) in addition to the single variant tests (singleVar) (see Methods).

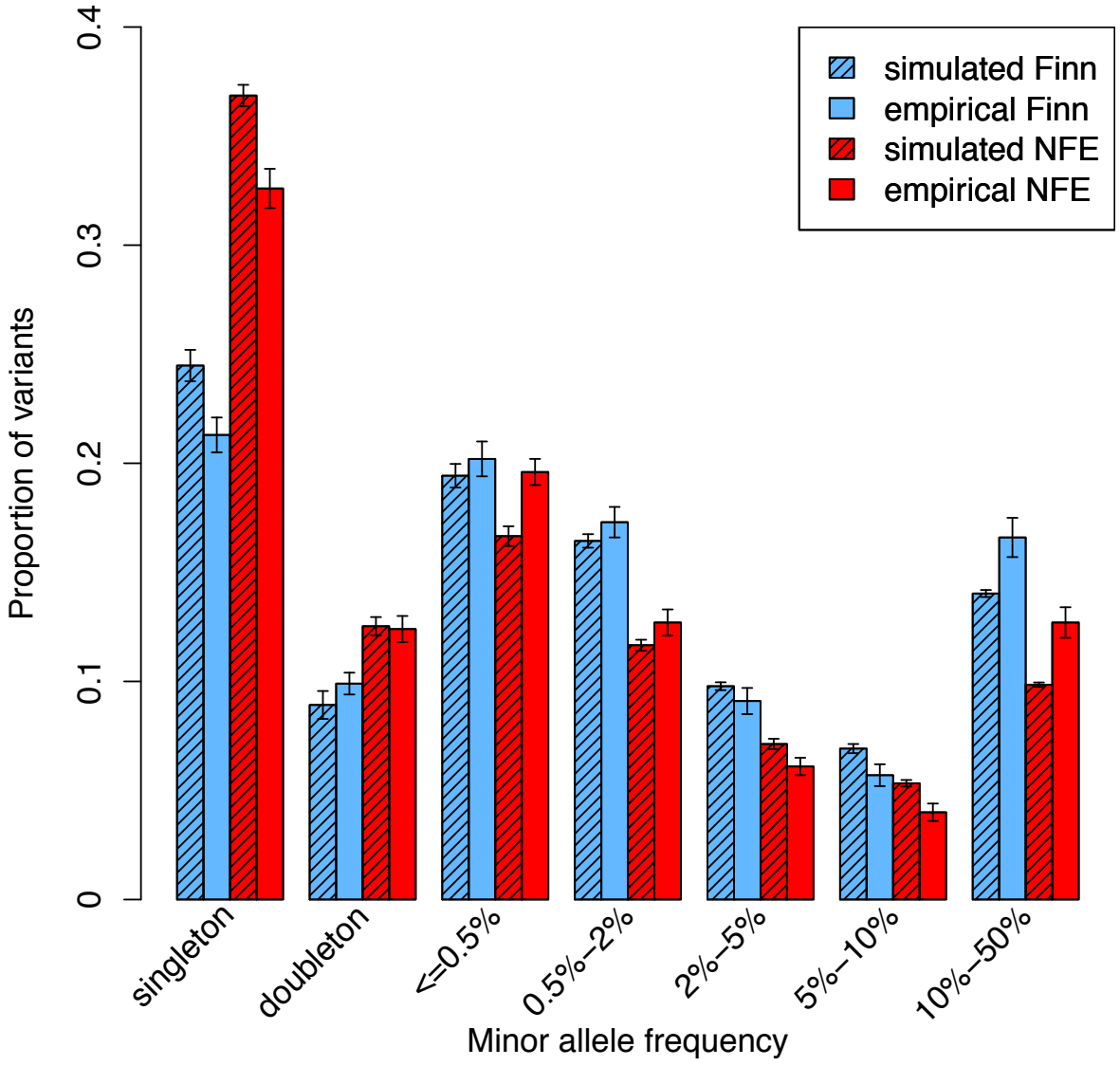

Figure S17. Agreement of empirical allele frequency spectra with the modeled spectra of exome chip data.

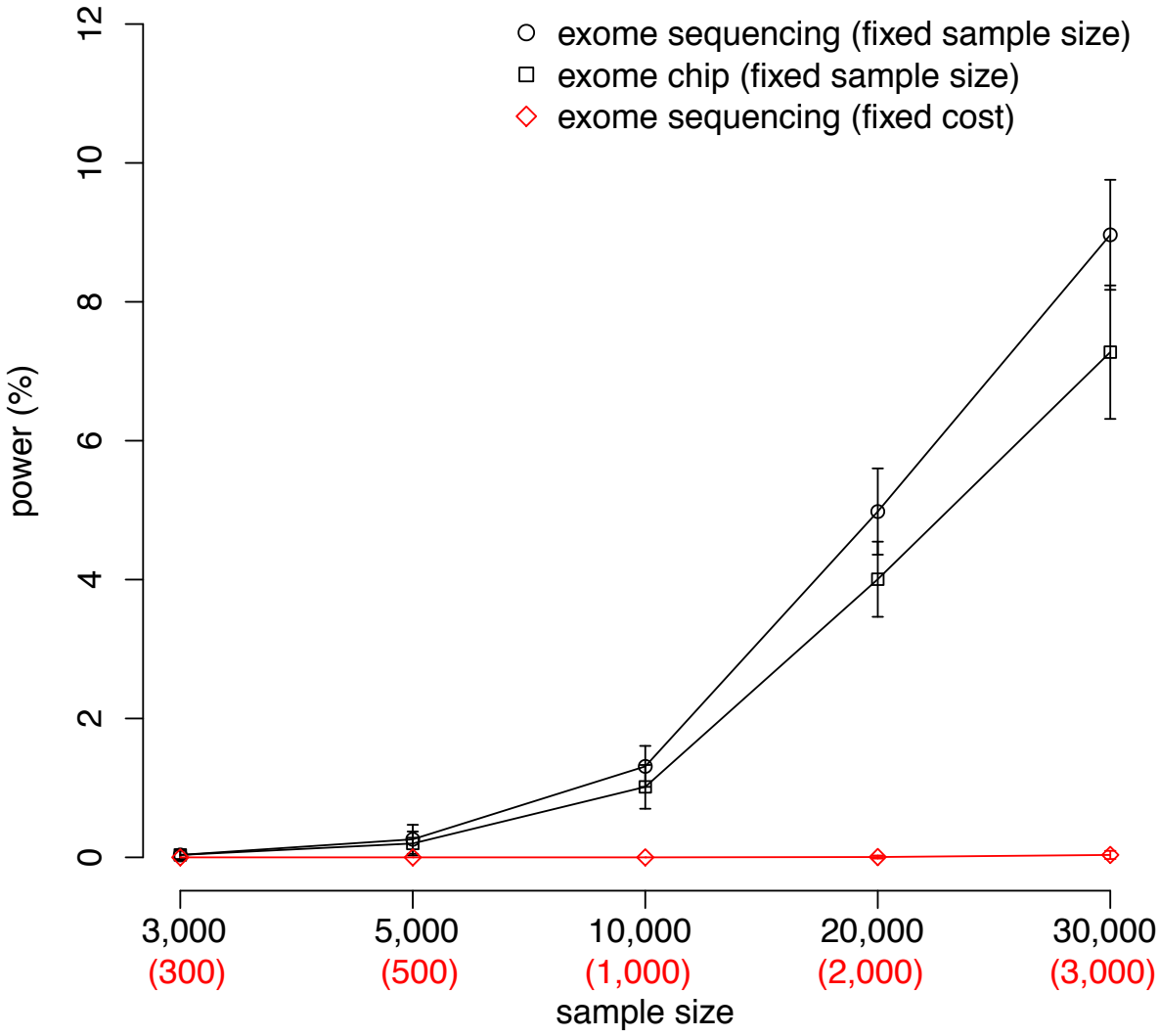

Figure S18. Power of exome chip study *vs* exome sequencing study in the NFEs under M4 using SKAT-O test. As different genes are likely to have different pleiotropic effects and are therefore exposed to different strengths of purifying selection, M4 is generated to represent a potentially more realistic scenario where τ (the degree of coupling between a causal variant's phenotypic effect and the strength of purifying selection against that variant) is randomly chosen among 0, 0.5 and 1 for each effect gene. The top two lines show power comparison at a fixed sample size; the bottom two lines show power comparison at a fixed cost (and thus only a tenth of the samples were sequenced).

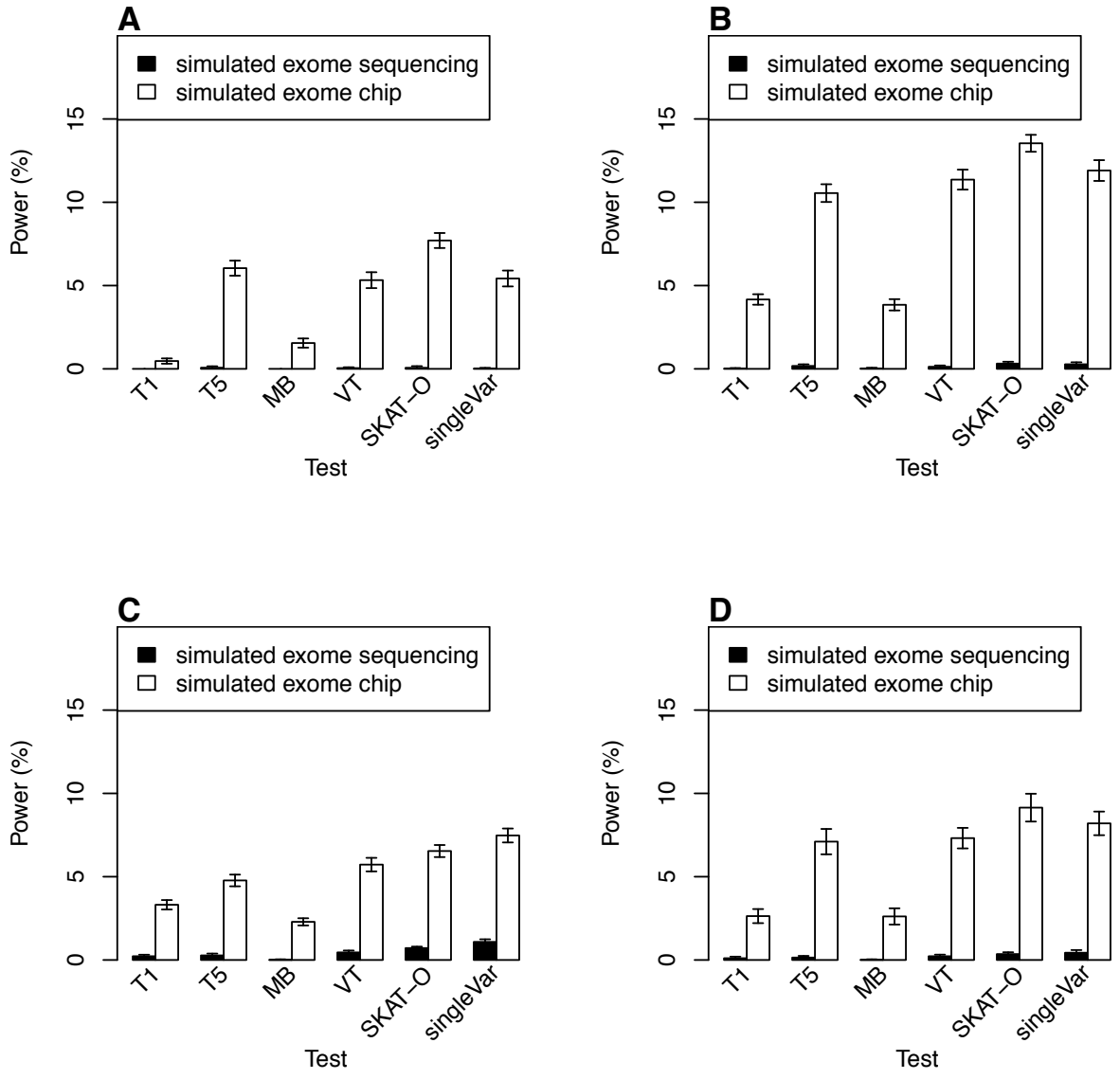

Figure S19. Power of exome chip study (N=30,000) *vs* exome sequencing study (N=3,000) in the Finns under four different disease models. (A) M1: τ=0; (B) M2: τ=0.5; (C) M3: τ=1; (D) M4: τ randomly sampled from 0, 0.5, and 1 for each effect gene.

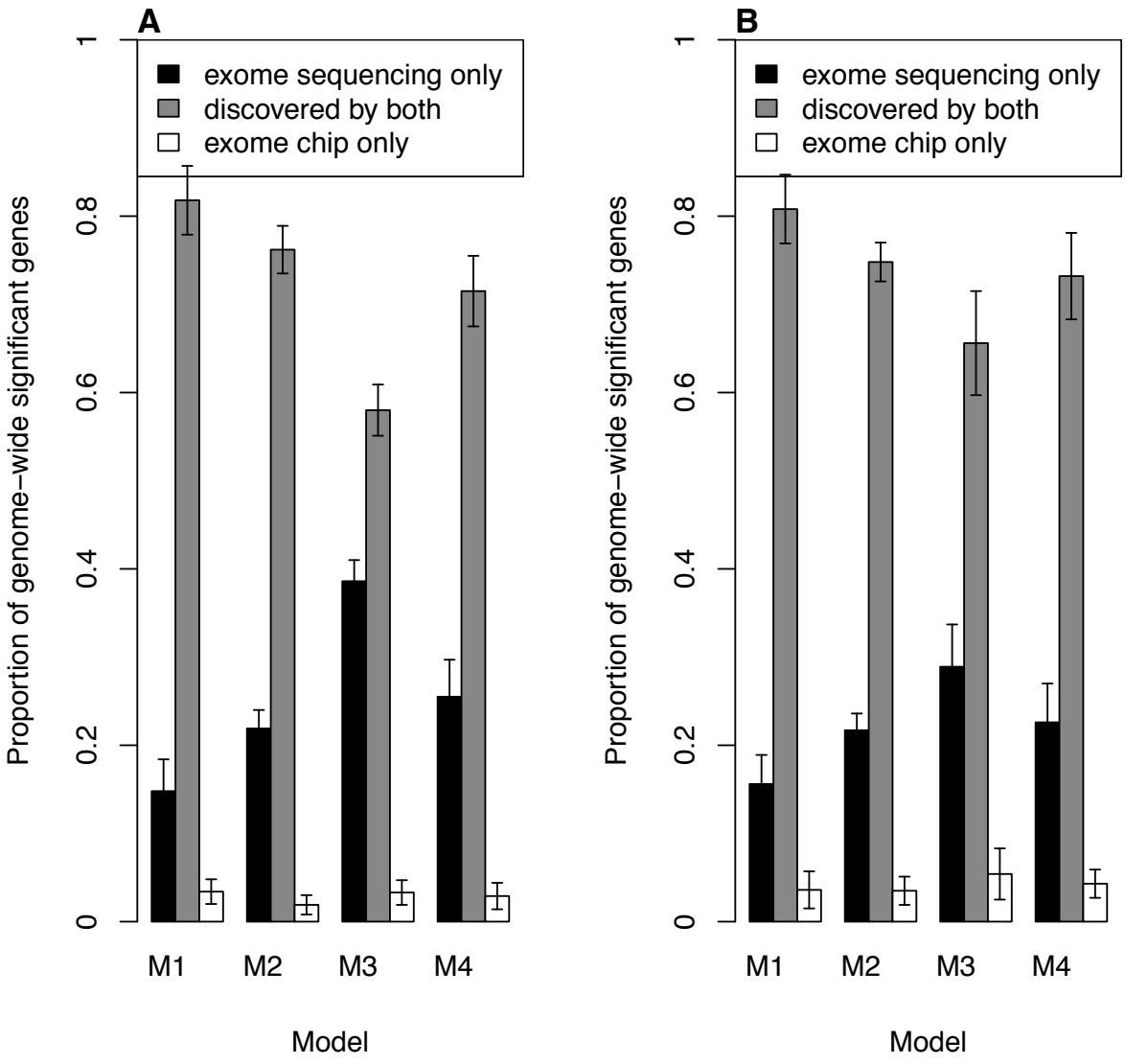

Figure S20. Proportion of genes detected by exome sequencing (N=30,000) only, or by exome chip (N=30,000) only, or by both (using SKAT-O test) under four different disease models (M1: τ=0; M2: τ=0.5; M3: τ=1; M4: τ randomly sampled from 0, 0.5, and 1 for each effect gene). As τ gets larger, the proportion of genes detected by exome sequencing only increases. (A) Results in Finns; (B) results in NFEs.

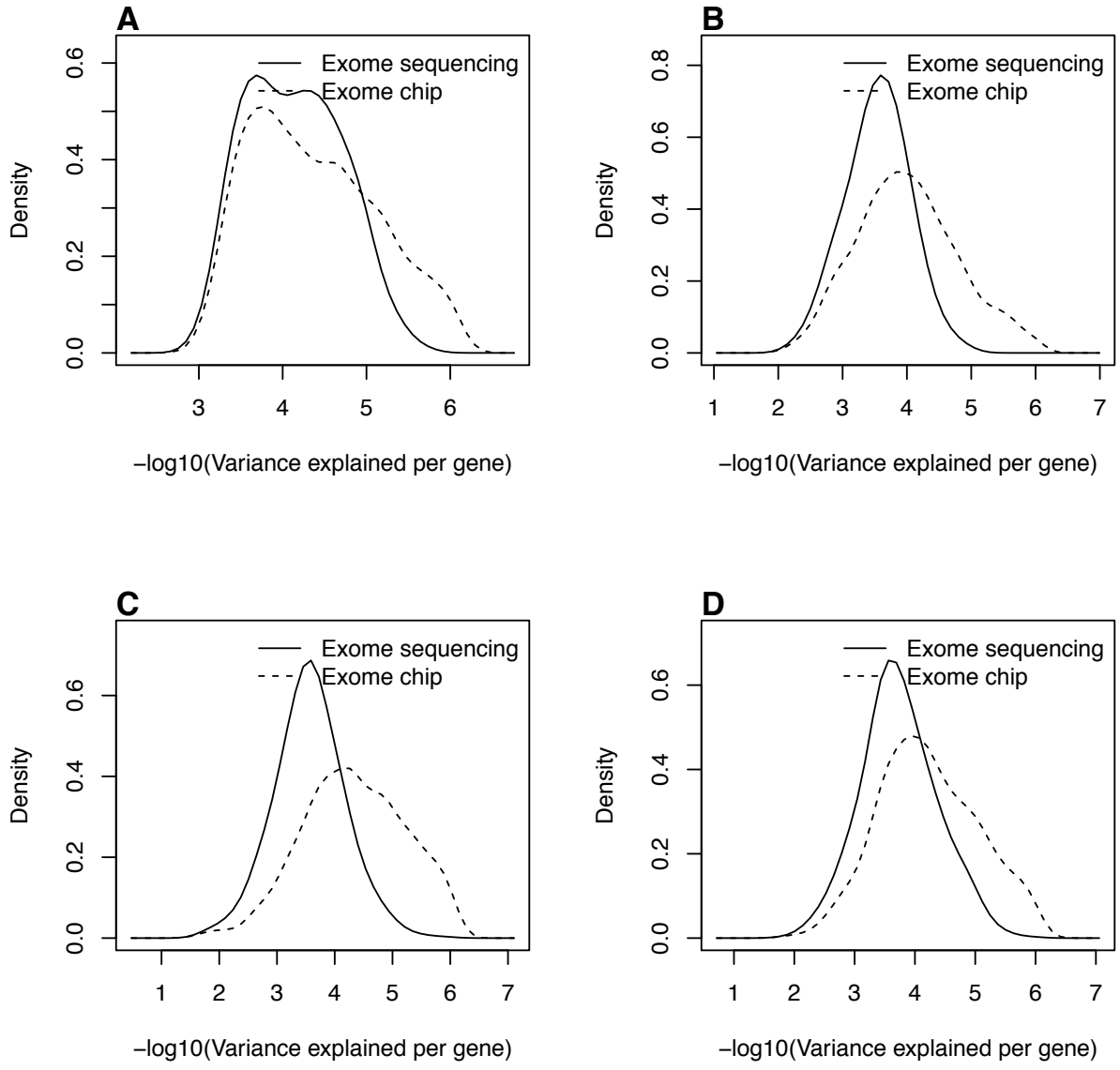

Figure S21. Distribution of variance explained per gene by variants with MAF below 5% in exome sequencing (solid line) or exome chip (dashed line) data of 30,000 Finns, under four different disease models. (A) M1 (τ=0); (B) M2 (τ=0.5); (C) M3 (τ=1); (D) M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene).

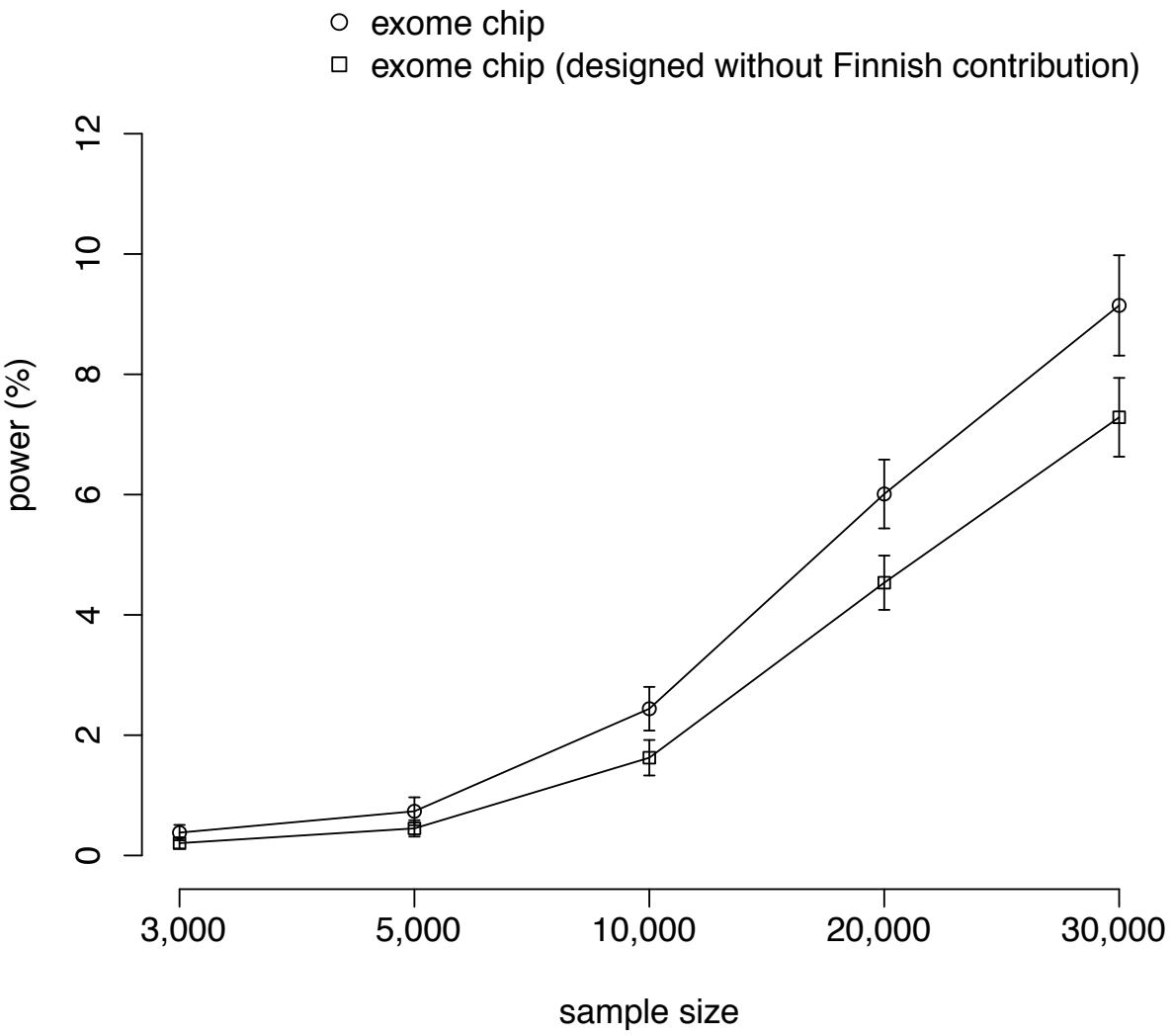

Figure S22. Power of two different exome chips in the Finns under M4 (τ randomly sampled from 0, 0.5, and 1 for each effect gene) using SKAT-O test. One chip design resembles that of the actual exome chip design (top line); the other chip design uses NFE samples only with no contribution from Finnish samples (bottom line).

Table S1. Birth place distribution of FUSION samples

Birth place	Number of samples
UUSIMAA, UUDENMAAN / NYLAND	27
TURKU-PORI, TURUN JA PORIN / ABO-BJORNEBORG, ABO-OCH-BJORNEBORG	92
HAME, HAMEEN / TAVASTEHUS	105
KYMI, KYMEN / KYMMMENE, VIBORG, VIIPURI	49
MIKKELI, MIKKELIN / SAINT MICHEL	61
POHJOIS-KARJALA, POHJOIS-KARJALAN / NORRA-KARALEN, NORRA KARENS	52
KUOPIO, KUOPION / KUOPIO	148
KESKI-SUOMI, KESKI-SUOMEN / MELLERSTA-FINLAND	76
VAASA, VASAAN / VASA, WASA	125
OULU, OULUN / ULEABORG	41
LAPPI, LAPIN / LAPPLAND	13
KARJALA, VIIPURI (area formerly part of Finland)	54
Total	**843**

Table S2. Three different models of Finnish population history

Parameters	Class 1 Model[a]	Class 2 Model[b]	Class 3 Model[c]
Bottleneck size	200-4000	1000	1000
Bottleneck time	1.5-3.5ky ago	2.5ky ago	2.5ky ago
Growth rate (per generation)	2.5-10%	5-10%	Slow phase:0.5-5% Fast phase: 8-30%
Gene flow into Finns	0	1-5%	0.5-7%
Minimal -log(P(data|model))	1419	426	267

[a]Founding bottleneck event followed by exponential growth of constant growth rate, with no gene flow between NFEs and Finns

[b]Founding bottleneck event followed by exponential growth of constant growth rate, with gene flow from NFEs into Finns

[c]Founding bottleneck event followed by a slow growth phase and a fast growth phase, with gene flow from NFEs into Finns

Table S3. Variants found in both samples tend to have higher allele counts in Finns

Variants	Finn-NFE		Finn-Finn		NFE-NFE	
	Difference	p value	Difference	p value	Difference	p value
Synonymous	0.415	0.000733	0.00451	0.911	0.0059	0.845
Missense	0.505	1.43e-06	0.00868	0.777	0.000157	0.994

For Variants shared between 250 Finns and 250 NFEs, their allele counts tend to be higher in Finns (paired t-test). As controls, we also checked allele counts for variants shared between 250 Finns and another 250 Finns, as well as between 250 NFEs and another 250 NFEs.