The American Journal of Human Genetics, Volume *99*

Supplemental Data

Determinants of Power in Gene-Based Burden Testing for Monogenic Disorders

Michael H. Guo, Andrew Dauber, Margaret F. Lippincott, Yee-Ming Chan, Rany M. Salem, and Joel N. Hirschhorn

Supplementary Notes

Additional Explanations of Parameters

Locus Heterogeneity

Locus heterogeneity refers to the number of genes in which pathogenic mutations can cause a given disease. We assign a parameter f_{case} to denote the proportion of cases caused by mutations in a given disease-associated gene, which is inversely related to locus heterogeneity. Throughout, we use a bolded $\boldsymbol{f_{case}}$ to represent a summary parameter for the average contributions of disease-associated genes to a given disorder, and unbolded f_{case} to represent the contribution of a given gene.

In our framework, we assume that there are no subgroups of the disorder that can be recognized clinically; therefore, different genes associated with a given disorder are indistinguishable phenotypically. Throughout the majority of the manuscript, we also assume that there are no phenocopies for the disorder and that all cases have disease as result of a pathogenic mutation in a monogenic disease-associated gene (i.e., the cases do not have disease from a non-genetic cause and or a polygenic burden of disease risk-increasing variants). Thus, the fractional contributions of the disease-associated genes sum to 100%. In Figure S10, we also consider the impact of potential phenocopies. Also, we assume that there are no known genes for a given disorder. If disease-associated genes are known and are screened against in the case cohort, the fractional contributions of the remaining unknown genes are increased.

Penetrance

Penetrance, π, is the proportion of individuals carrying pathogenic variants who develop disease. Incomplete penetrance influences the background rate of variation as pathogenic variants that do not manifest disease. When observing the number of qualifying variants empirically from the data, we cannot readily determine the proportion of qualifying variants present in controls that represent incompletely penetrant variants versus the proportion that represent benign variants misclassified as qualifying variants. It is also important to note that prevalence of disease (P) is an important consideration in the presence of incomplete penetrance. This is because the proportion of individuals in a disease-free control cohort who carry pathogenic mutations in a given disease-associated gene is proportional to $(P/\pi)\times(1-\pi)$. In the presence of full penetrance, this term goes to 0.

When penetrance approaches the disease prevalence, the likelihood of affected relatives sharing the same pathogenic variant(s) decreases; as such, we define a monogenic disorder as being caused by variants whose penetrance is sufficiently greater than disease prevalence to make it likely that affected relatives share the same pathogenic variants. The definition of what "sufficiently greater" is depends on the number of meiosis separating relatives and can be estimated using Bayes' Theorem. For example, when penetrance is ten-fold greater than the prevalence of a dominant disorder, then an affected individual will have at least a ~90% probability of sharing a pathogenic variant present in a sibling; for first cousins, the penetrance needs to be 70-fold greater than prevalence to achieve the same probability of sharing.

Sensitivity and Specificity to Distinguish Variants

The typical gene-based burden test applies filters (such as MAF and predicted effect on protein function) to try to enrich for variants that are more likely to be pathogenic. However, these filters are imperfect and thus have an associated specificity and sensitivity for each disease-associated gene. Increasing the stringency of each threshold can result in increased specificity, with fewer benign variants classified as qualifying variants, but can also likely decrease sensitivity, with fewer pathogenic variants classified as qualifying variants. The precise nature of trade-off between the stringency of the thresholds and the specificity and sensitivity is a complex relationship that is not readily assessable empirically and is beyond the scope of this work. This is because most currently assignments of the pathogenicity of variants are dependent on MAF and protein prediction annotations, introducing an inherent circularity in assessments of sensitivity and specificity.

A consideration that is intricately linked with sensitivity is the technical ability to detect pathogenic variants. Some parts of the exome and some forms of genetic variation are poorly sequenced and/or are difficult to variant call [1]. These poorly genotyped variants may be ascertained by lowering sequencing/variant calling quality thresholds, which would improve sensitivity, but at the cost of introducing noise in the form of artifactual variants [2]. In addition, current exome sequencing technologies fail to capture regions of the exome, often up to 20% [3]. Also, some forms of genetic variation are largely missed by exome sequencing, such as longer indels and CNVs.

Statistical Significance

Throughout our studies, we have used a p-value threshold of $2.5\text{x}10^{-6}$ for declaring association of a gene with disease. This represents $\alpha=0.05$ corrected for testing of approximately 20,000 genes. However, genes that meet this p-value threshold may not be truly disease-causing even in the absence of any artifacts (such as batch effects or mismatching of ancestry between cases and controls). Approximately 5% of disease-associated genes that meet this threshold will be false positives given the α. Additionally, a significant p-value in isolation is unlikely to be sufficient evidence to claim causality [4]. Almost certainly, additional functional or genetic replication will need to be performed.

Reaching a significant p-value actually does not require many cases. For example, for a gene with no controls (out of 2597 total controls) who carry a qualifying variant in a given gene, only 3 cases out of 10 total cases are needed to reach statistical significance under a Fisher's exact test ($p=4.07x10^{-8}$). If ***f*case** is 0.1, then in a given experiment, it is quite likely that any one of ten disease-associated genes will have 3 or more cases carrying variants in that gene when sequencing 10 disease cases.

dbGaP Sample Information/ Acknowledgements

The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers: phs000007 (FHS), phs000179 (COPDGene), phs000200 (WHI), phs000209 (MESA), phs000280 (ARIC), phs000286 (JHS), phs000287 (CHS), and phs000285 (CARDIA).

ARIC

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). The authors thank the staff and participants of the ARIC study for their important contributions. This study is part of the NHLBI Grand Opportunity Exome Sequencing Project (GO-ESP). Funding for GO-ESP was provided by NHLBI grants RC2 HL103010 (HeartGO), RC2 HL102923 (LungGO) and RC2 HL102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL102925 (BroadGO) and RC2 HL102926 (SeattleGO). HeartGO gratefully acknowledges the following groups and individuals who provided biological samples or data for this study. DNA samples and phenotypic data were obtained from the following studies supported by the NHLBI: the Atherosclerosis Risk in Communities (ARIC) study, the Coronary Artery Risk Development in Young Adults (CARDIA) study, Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), the Jackson Heart Study (JHS) and the Multi-Ethnic Study of Atherosclerosis (MESA). This manuscript was not prepared in collaboration with investigators of the Atherosclerosis Risk in

Communities Study and does not necessarily reflect the opinions or views of the ARIC, or NHLBI.

CARDIA

The Coronary Artery Risk Development in Young Adults Study (CARDIA) is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the University of Alabama at Birmingham (N01-HC95095 & N01-HC48047), University of Minnesota (N01-HC48048), Northwestern University (N01-HC48049), and Kaiser Foundation Research Institute (N01-HC48050). This study is part of the NHLBI Grand Opportunity Exome Sequencing Project (GO-ESP). Funding for GO-ESP was provided by NHLBI grants RC2 HL103010 (HeartGO), RC2 HL102923 (LungGO) and RC2 HL102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL102925 (BroadGO) and RC2 HL102926 (SeattleGO). HeartGO gratefully acknowledges the following groups and individuals who provided biological samples or data for this study. DNA samples and phenotypic data were obtained from the following studies supported by the NHLBI: the Atherosclerosis Risk in Communities (ARIC) study, the Coronary Artery Risk Development in Young Adults (CARDIA) study, Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), the Jackson Heart Study (JHS) and the Multi-Ethnic Study of Atherosclerosis (MESA). This manuscript was not approved by CARDIA. The opinions and conclusions contained in this publication are solely those of the authors, and are not endorsed by CARDIA or the NHLBI and should not be assumed to reflect the opinions or conclusions of either.

CHS

The research reported in this article was supported by contract numbers N01-HC-85079, N01-HC-85080, N01-HC-85081, N01-HC-85082, N01-HC-85083, N01-HC- 85084, N01-HC-85085,

N01-HC-85086, N01-HC-35129, N01 HC-15103, N01 HC- 55222, N01-HC-75150, N01-HC-45133, N01-HC-85239 and HHSN268201200036C; grant numbers U01 HL080295 from the National Heart, Lung, and Blood Institute and R01 AG-023629 from the National Institute on Aging, with additional contribution from the National Institute of Neurological Disorders and Stroke. A full list of principal CHS investigators and institutions can be found at http://www.chs-nhlbi.org/pi.htm. This study is part of the NHLBI Grand Opportunity Exome Sequencing Project (GO-ESP). Funding for GO-ESP was provided by NHLBI grants RC2 HL103010 (HeartGO), RC2 HL102923 (LungGO) and RC2 HL102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL102925 (BroadGO) and RC2 HL102926 (SeattleGO). HeartGO gratefully acknowledges the following groups and individuals who provided biological samples or data for this study. DNA samples and phenotypic data were obtained from the following studies supported by the NHLBI: the Atherosclerosis Risk in Communities (ARIC) study, the Coronary Artery Risk Development in Young Adults (CARDIA) study, Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), the Jackson Heart Study (JHS) and the Multi-Ethnic Study of Atherosclerosis (MESA). This manuscript was not prepared in collaboration with CHS investigators and does not necessarily reflect the opinions or views of CHS, or the NHLBI.

COPDGene

This research used data generated by the COPDGene study, which was supported by NIH grants U01HL089856 and U01HL089897. The COPDGene project is also supported by the COPD Foundation through contributions made by an Industry AdvisoryBoard comprised of Pfizer, AstraZeneca, Boehringer Ingelheim, Novartis, and Sunovion. This study is part of the NHLBI Grand Opportunity Exome Sequencing Project (GO-ESP). Funding for GO-ESP was provided by NHLBI grants RC2 HL103010 (HeartGO), RC2 HL102923 (LungGO) and RC2 HL102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL102925 (BroadGO) and RC2 HL102926 (SeattleGO). This manuscript was not prepared in collaboration with investigators of the COPDGene and does not necessarily reflect the opinions or views of the COPDGene, or NIH.

FHS

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195). The project utilized data from the Heart Cohorts Exome Sequencing Project (FHS), part of the NHLBI Grand Opportunity Exome Sequencing Project (GO-ESP). Funding for GO-ESP was provided by NHLBI grants RC2 HL103010 (HeartGO), RC2 HL102923 (LungGO) and RC2 HL102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL102925 (BroadGO) and RC2 HL102926 (SeattleGO). HeartGO gratefully acknowledges the following groups and individuals who provided biological samples or data for this study. DNA samples and phenotypic data were obtained from the following studies supported by the NHLBI: the Atherosclerosis Risk in Communities (ARIC) study, the Coronary Artery Risk Development in Young Adults (CARDIA) study, Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), the Jackson Heart Study (JHS) and the Multi-Ethnic Study of Atherosclerosis (MESA). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI.

JHS

The Jackson Heart Study is supported and conducted in collaboration with Jackson State University (N01-HC-95170), University of Mississippi Medical Center (N01-HC-95171), and Touglaoo College (N01-HC-95172) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD), with additional support from the National Institute on Biomedical Imaging and Bioengineering (NIBIB). This study is part of the NHLBI Grand Opportunity Exome Sequencing Project (GO-ESP). Funding for GO-ESP was provided by NHLBI grants RC2 HL103010 (HeartGO), RC2 HL102923 (LungGO) and RC2 HL102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL102925 (BroadGO) and RC2 HL102926 (SeattleGO). HeartGO gratefully acknowledges the following groups and individuals who provided biological samples or data for this study. DNA samples and phenotypic data were obtained from the following studies supported by the NHLBI: the Atherosclerosis Risk in Communities (ARIC) study, the Coronary Artery Risk Development in Young Adults (CARDIA) study, Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), the Jackson Heart Study (JHS) and the Multi-Ethnic Study of Atherosclerosis (MESA). This manuscript was not prepared in collaboration with JHS investigators and does not necessarily reflect the opinions or views of JHS, or the NHLBI.

MESA

MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169 and CTSA UL1-RR-024156. This study is part of the NHLBI Grand

Opportunity Exome Sequencing Project (GO-ESP). Funding for GO-ESP was provided by NHLBI grants RC2 HL103010 (HeartGO), RC2 HL102923 (LungGO) and RC2 HL102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL102925 (BroadGO) and RC2 HL102926 (SeattleGO). HeartGO gratefully acknowledges the following groups and individuals who provided biological samples or data for this study. DNA samples and phenotypic data were obtained from the following studies supported by the NHLBI: the Atherosclerosis Risk in Communities (ARIC) study, the Coronary Artery Risk Development in Young Adults (CARDIA) study, Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), the Jackson Heart Study (JHS) and the Multi-Ethnic Study of Atherosclerosis (MESA). This manuscript was not prepared in collaboration with MESA investigators and does not necessarily reflect the opinions or views of MESA, or the NHLBI.

WHI

The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118-32119, 32122, 42107-26, 42129-32, and 44221. The Women's Health Initiative Sequencing Project (WHISP) is funded by Grant Number RC2 HL102924. This study is part of the NHLBI Grand Opportunity Exome Sequencing Project (GO-ESP). Funding for GO-ESP was provided by NHLBI grants RC2 HL103010 (HeartGO), RC2 HL102923 (LungGO) and RC2 HL102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL102925 (BroadGO) and RC2 HL102926 (SeattleGO). This manuscript was not prepared in collaboration with investigators of the WHI, has not been reviewed and/or approved by the Women's Health Initiative (WHI), and does not necessarily reflect the opinions of the WHI investigators or the NHLBI.

Figure S1

For each gene (n=1,2…20000):

	CASES	CONTROLS
CARRY QUALIFYING VARIANT	**CASE$_{QV}$** Estimate based on simulated total case size, background variation, f_{case}, and sensitivity	**CONTROL$_{QV}$** Observe from sequencing data based on thresholds
NO QUALIFYING VARIANT	**CASE$_{NQV}$** Simulate	**CONTROL$_{NQV}$** Observe from sequencing data based on thresholds

Figure S1: Simulation framework
For each gene, we construct a 2x2 contingency table with cases and controls and presence/absence of qualifying variant. We estimate the number of controls carrying (background variation, CONTROL$_{QV}$) and not carrying (CONTROL$_{NQV}$) qualifying variants at a set of thresholds from the control exome sequencing data (n=2597). We simulate the total case size (CASE$_{QV}$+CASE$_{NQV}$), and based on the total case size, background variation, and genetic architecture parameters (f_{case} and sensitivity), estimate the number of cases carrying a qualifying variant (CASE$_{QV}$). A p-value can then be calculated based on this 2x2 contingency table.

Legends for Figure S2-S6

Figure S2: Background rates at different MAF thresholds
Background rate of variation (proportion of controls carrying qualifying variants) in each gene considering all nonsynonymous variants for private (S2A), MAF≤0.01% (S2B), MAF≤0.1% (S2C), and MAF≤1% (S2D). Genes are ranked on the horizontal axis from the least variable to the most variable. Each point on the plot represents a single gene. Plot truncated at background rate of 0.2 (20%).

Figure S3: Sample size needs at different MAF thresholds
Sample size needs to have 80% power to detect each gene for private (S3A), MAF≤0.01% (S3B), MAF≤0.1% (S3C), and MAF≤1% (S3D) under the base model. Analyses performed using all nonsynonymous variants under a dominant model. Plot truncated at 300 samples.

Figure S4: Background rates at different protein-deleteriousness thresholds
Background rate of variation (proportion of controls carrying qualifying variants) in each gene at MAF≤0.1%. Analyses performed under a dominant model using all nonsynonymous variants (S4A), LOF plus damaging missense variants (S4B) or LOF variants only (S4C). Genes are ranked on the horizontal axis from the least variable to the most variable. Each point on the plot represents a single gene. Plot truncated at background rate of 0.2 (20%).

Figure S5: Sample size needs at different protein-deleteriousness thresholds
Sample size needs to have 80% power to detect each gene at MAF threshold≤0.1% under the base model. Analyses performed under a dominant model using all nonsynonymous variants (S5A), LOF plus damaging missense variants (S5B) or LOF variants only (S5C). Plot truncated at 300 samples.

Figure S6: Background rates at different MAF thresholds under recessive model
Background rate of variation (proportion of controls carrying qualifying variants) in each gene for private (S6A), MAF≤0.01% (S6B), MAF≤0.1% (S6C), and MAF≤1% (S6D). Genes are ranked on the horizontal axis from the least variable to the most variable. Each point on the plot represents a single gene. Plot truncated at background rate of 0.2 (20%).

Figure S7: Sample size needs at different MAF thresholds under recessive model
Sample size needs to have 80% power to detect each gene for private (S7A), MAF≤0.01% (S7B), MAF≤0.1% (S7C), and MAF≤1% (S7D) under the base model, except considering a recessive model. Analyses performed using all nonsynonymous variants. Plot truncated at 300 samples.

Figure S2

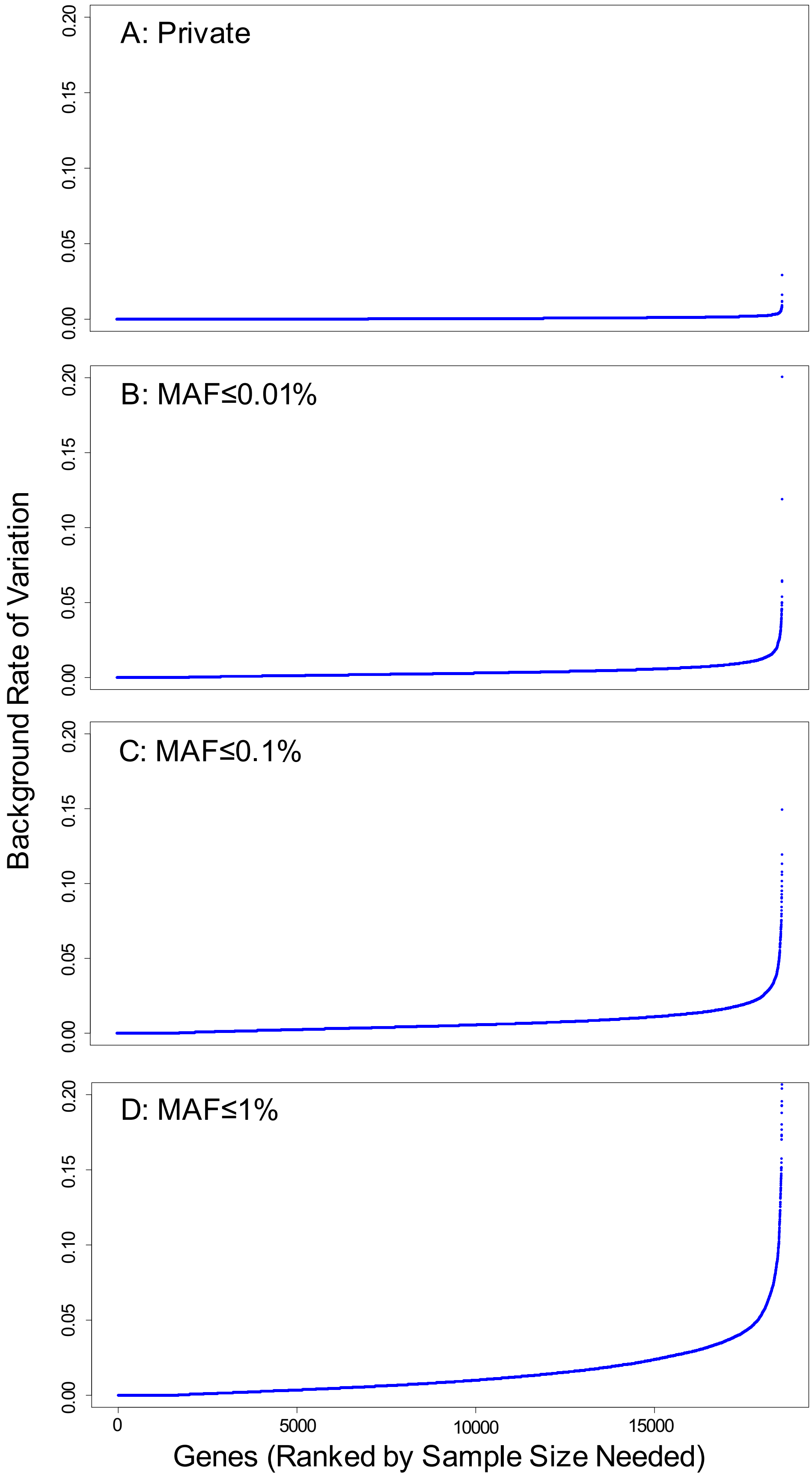

A: Private
B: MAF≤0.01%
C: MAF≤0.1%
D: MAF≤1%
Background Rate of Variation
0.00
0.05
0.10
0.15
0.20
0
5000
10000
15000
Genes (Ranked by Sample Size Needed)

Figure S3

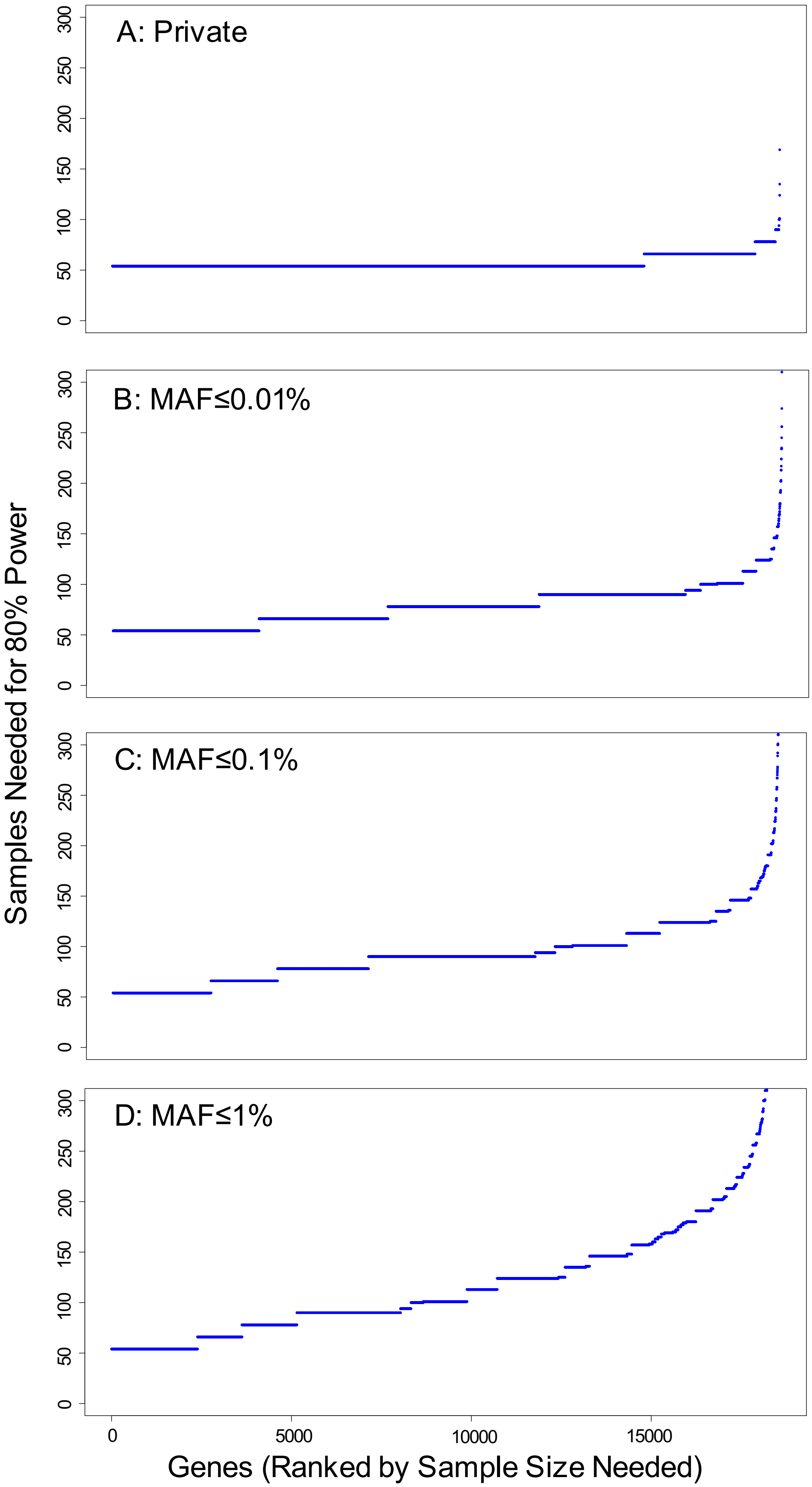
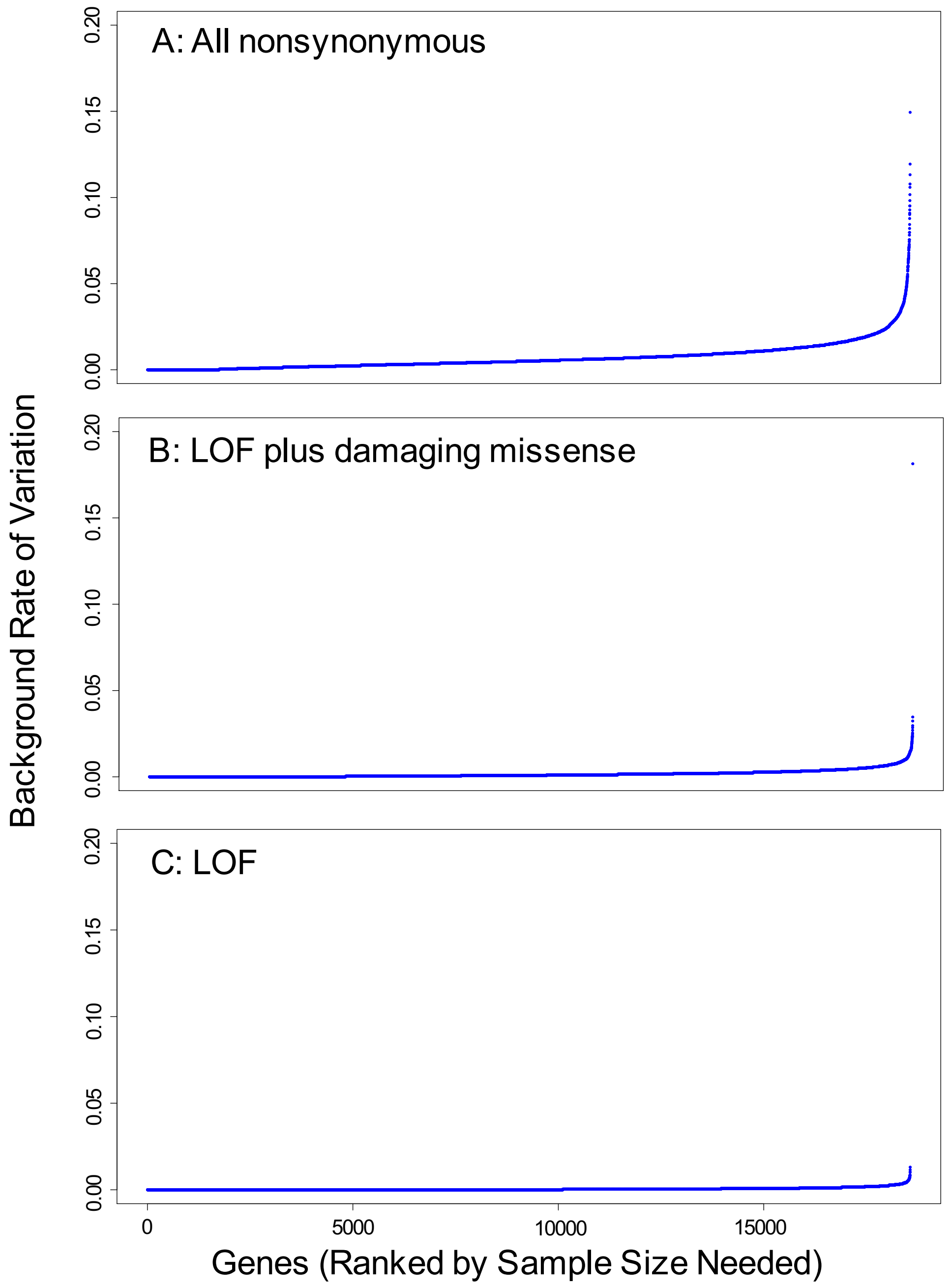

A: Private
B: MAF≤0.01%
C: MAF≤0.1%
D: MAF≤1%
Samples Needed for 80% Power
0
50
100
150
200
250
300
Genes (Ranked by Sample Size Needed)
0
5000
10000
15000

Figure S4



A: All nonsynonymous
B: LOF plus damaging missense
C: LOF
Background Rate of Variation
0.00
0.05
0.10
0.15
0.20
0
5000
10000
15000
Genes (Ranked by Sample Size Needed)

Figure S5

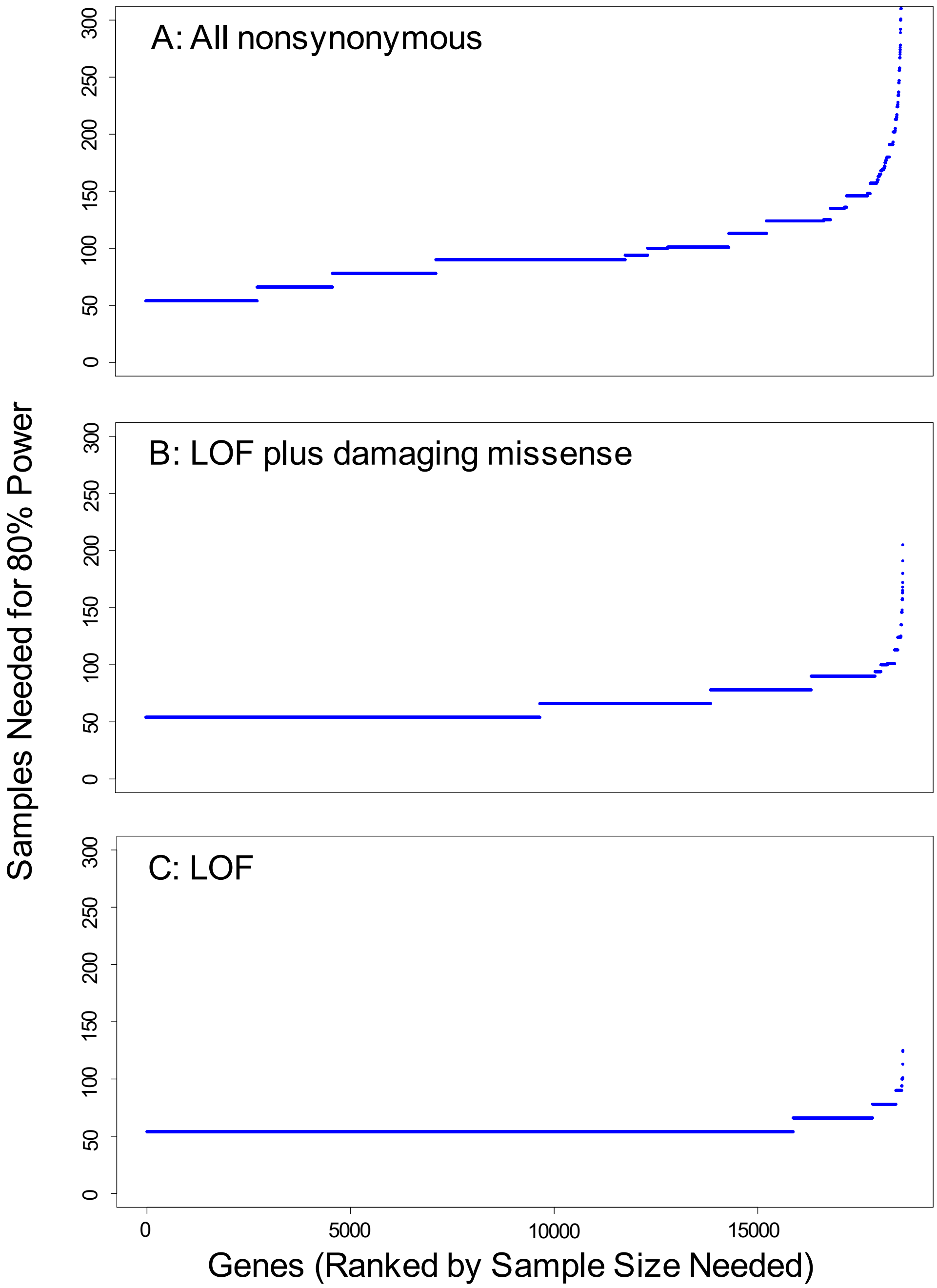

A: All nonsynonymous
B: LOF plus damaging missense
C: LOF
Samples Needed for 80% Power
0
50
100
150
200
250
300
0
5000
10000
15000
Genes (Ranked by Sample Size Needed)

Figure S6

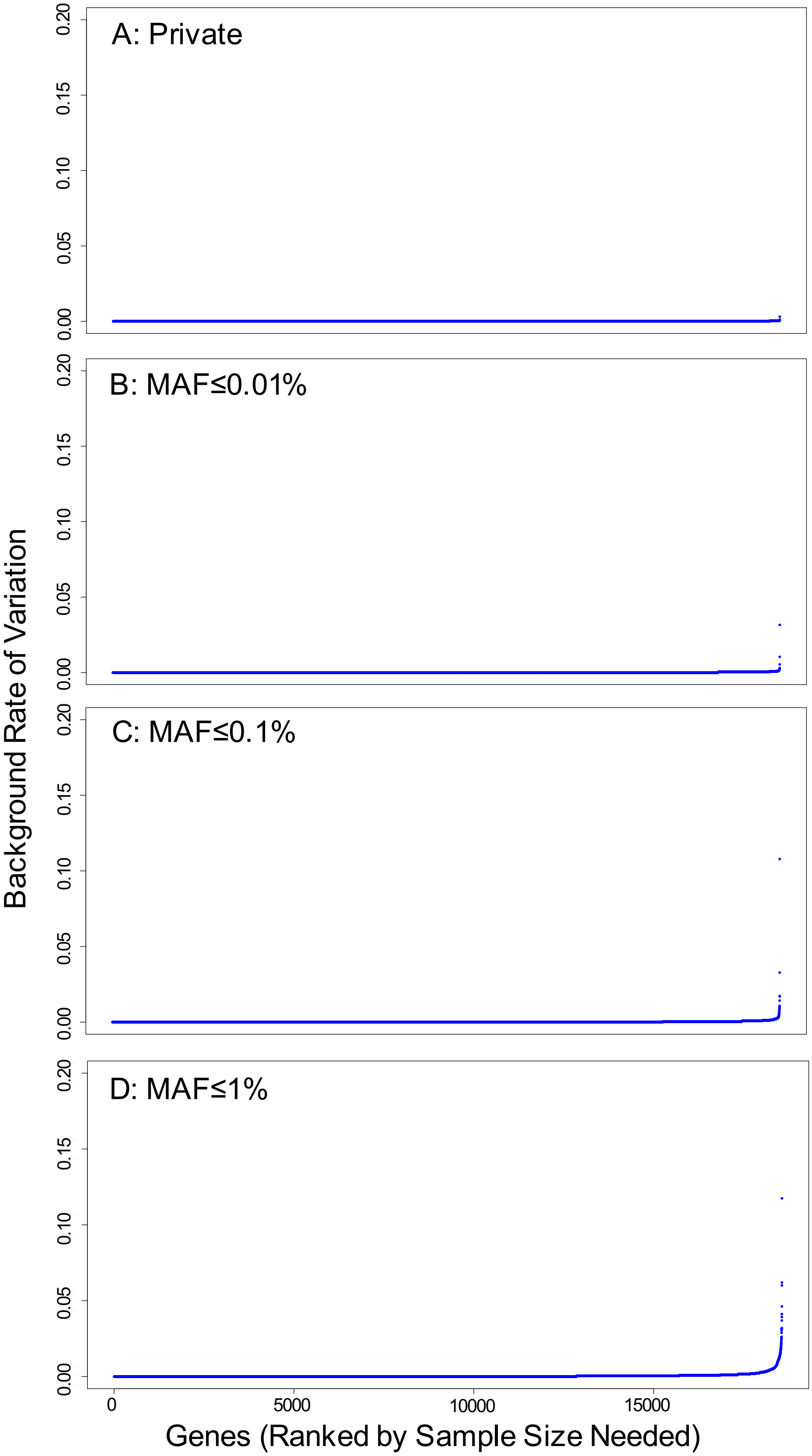

A: Private
B: MAF≤0.01%
C: MAF≤0.1%
D: MAF≤1%
Background Rate of Variation
0.00
0.05
0.10
0.15
0.20
0
5000
10000
15000
Genes (Ranked by Sample Size Needed)

Figure S7

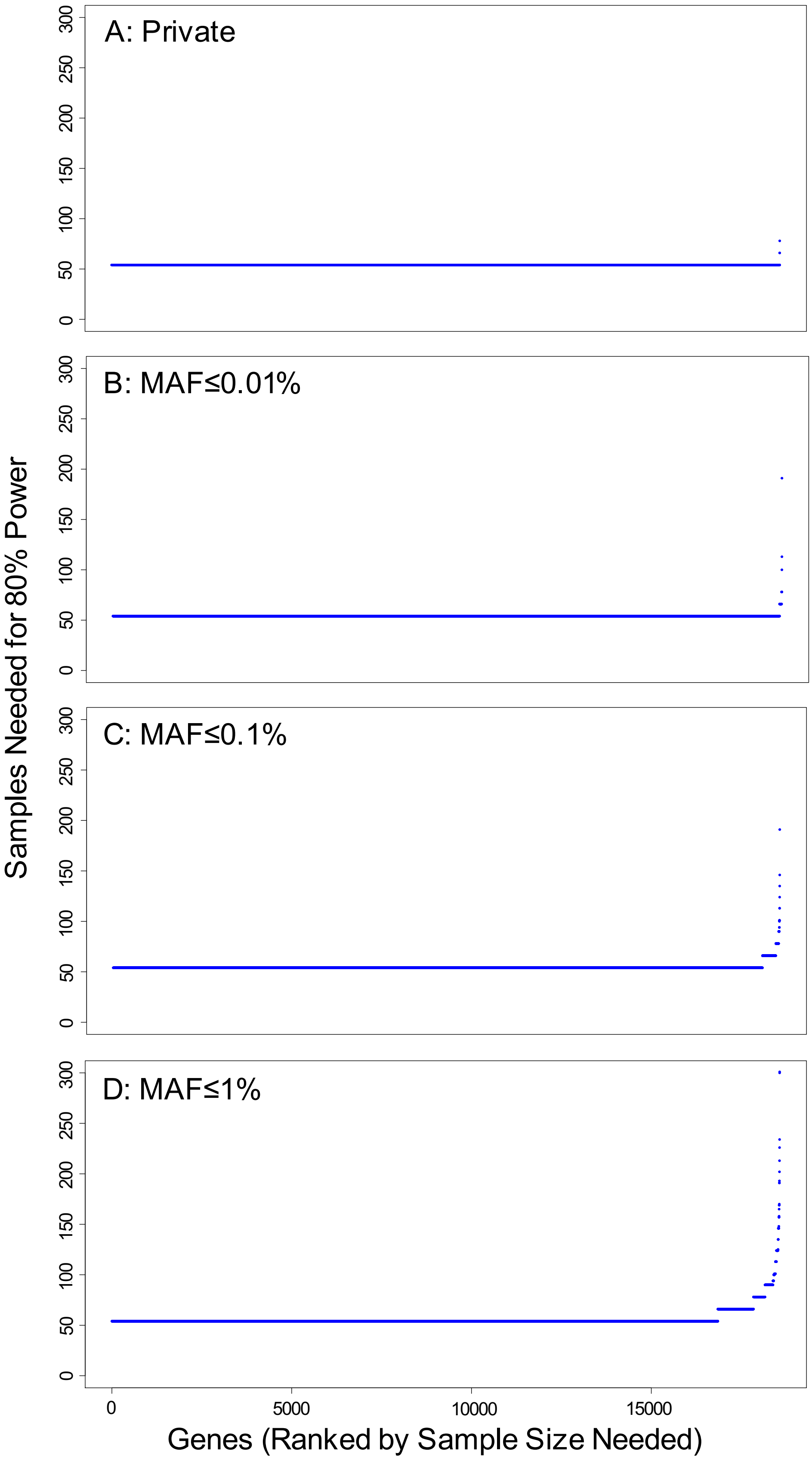

A: Private
B: MAF≤0.01%
C: MAF≤0.1%
D: MAF≤1%
Samples Needed for 80% Power
0
50
100
150
200
250
300
Genes (Ranked by Sample Size Needed)
0
5000
10000
15000

Figure S8

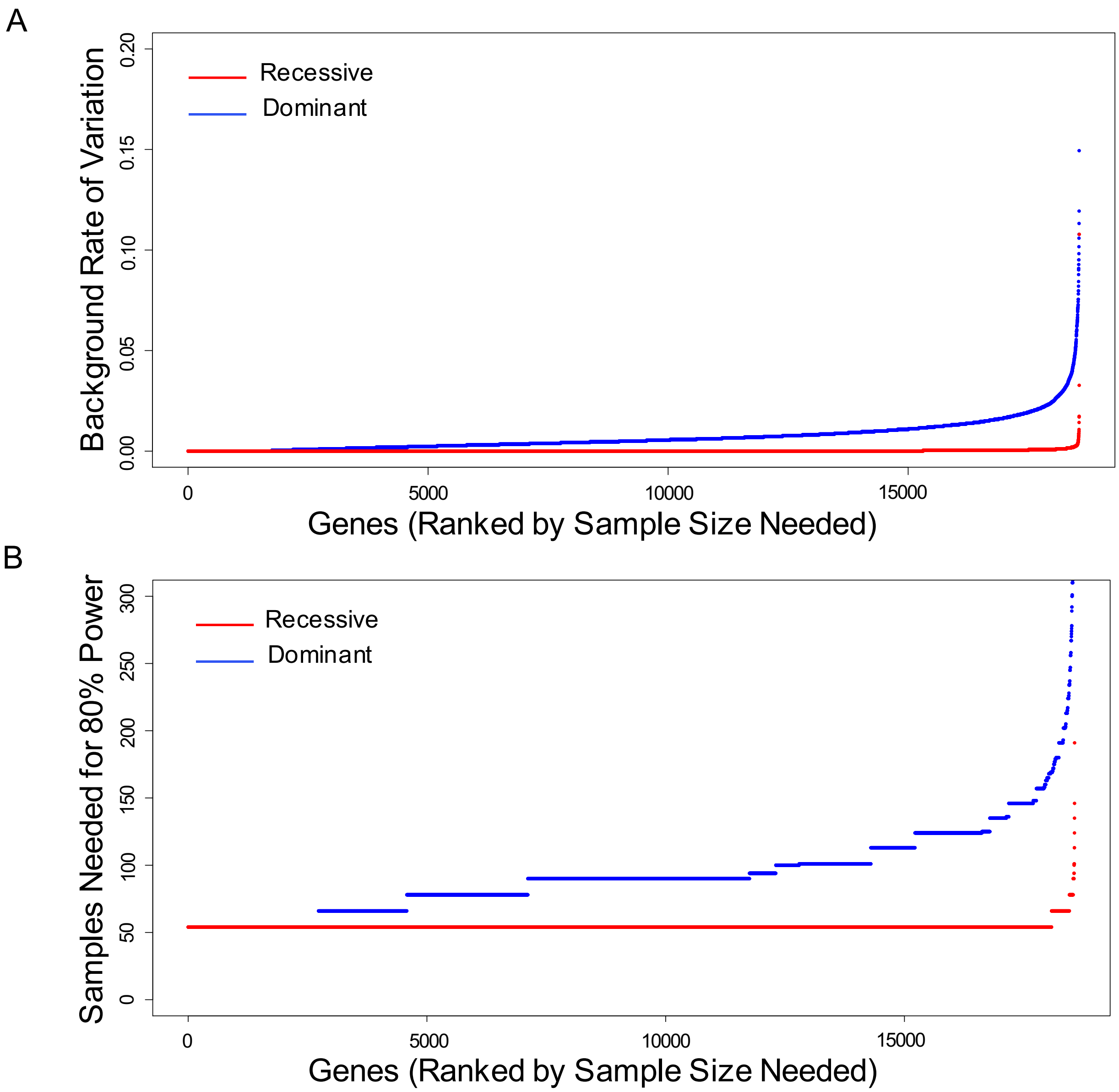


Figure S8: Comparison of dominant and recessive models

A) Background rate of variation (proportion of controls carrying qualifying variants) in each gene considering all nonsynonymous variants at MAF≤ 0.1% under a base model for a recessive (red) or dominant (blue, same as Figure 2A) disorder. Genes are ranked on the horizontal axis from the least variable to the most variable. Each point on the plot represents a single gene. Plot truncated at background rate of 0.2 (20%).

B) Sample size needs to have 80% power to detect each gene in the genome under a base model for a recessive (red), or dominant (blue, same as Figure 2B) disorder. Simulations were performed using with all nonsynonymous variants at MAF ≤ 0.1%. Plot truncated at 300 samples.

Figure S9

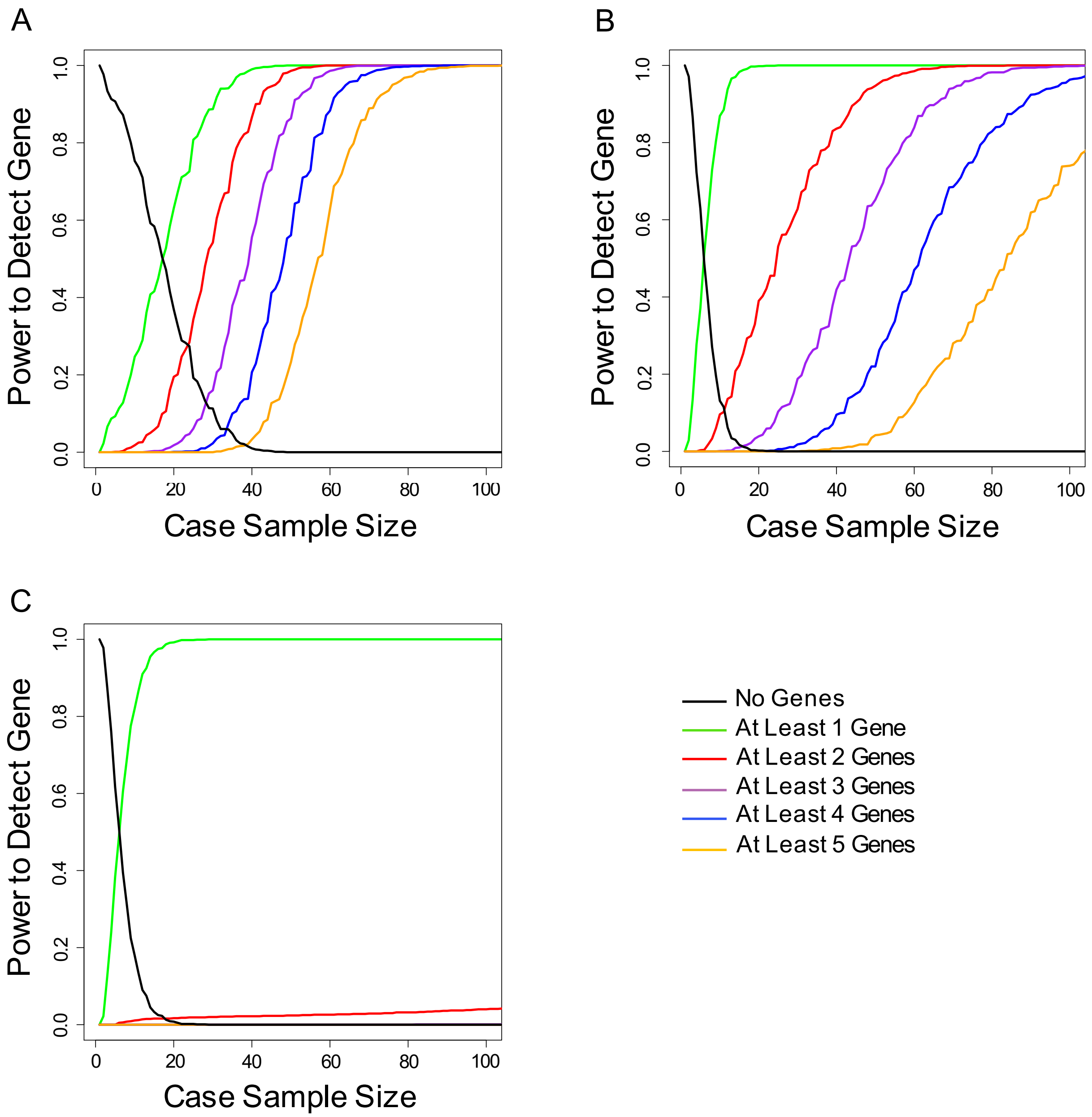


Figure S9: Power for unequal contributions to disease cases
Power to detect at least one disease-associated gene (green), at least two (red), at least three (purple), at least four (blue), at least five genes (orange), and no genes (black) at increasing case sample sizes. Analyses were performed for three separate sets of disease-associated gene contributions (Set 1,2,3 in A-C respectively) and considering all nonsynonymous variants at MAF≤0.1% under a dominant model. In set 1, each of ten genes contributes to 10% of cases (base model); in set 2, one gene contributes to 50% of cases and five genes each contribute to 10% of cases; in set 3, one gene contributes to 50% of cases, while 50 additional genes each contribute to 1% of cases.

Figure S10

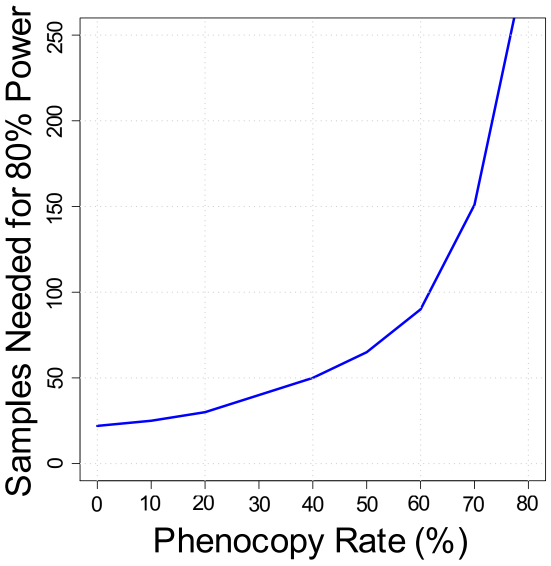


Figure S10: Effect of phenocopies
Samples needed for 80% power to detect at least one gene associated with disease as a function of phenocopy rate (expressed as a percentage). Phenocopy rate represents the percentage of disease cases who do not have disease due to pathogenic mutations in a monogenic disease-associated gene.

Figure S11

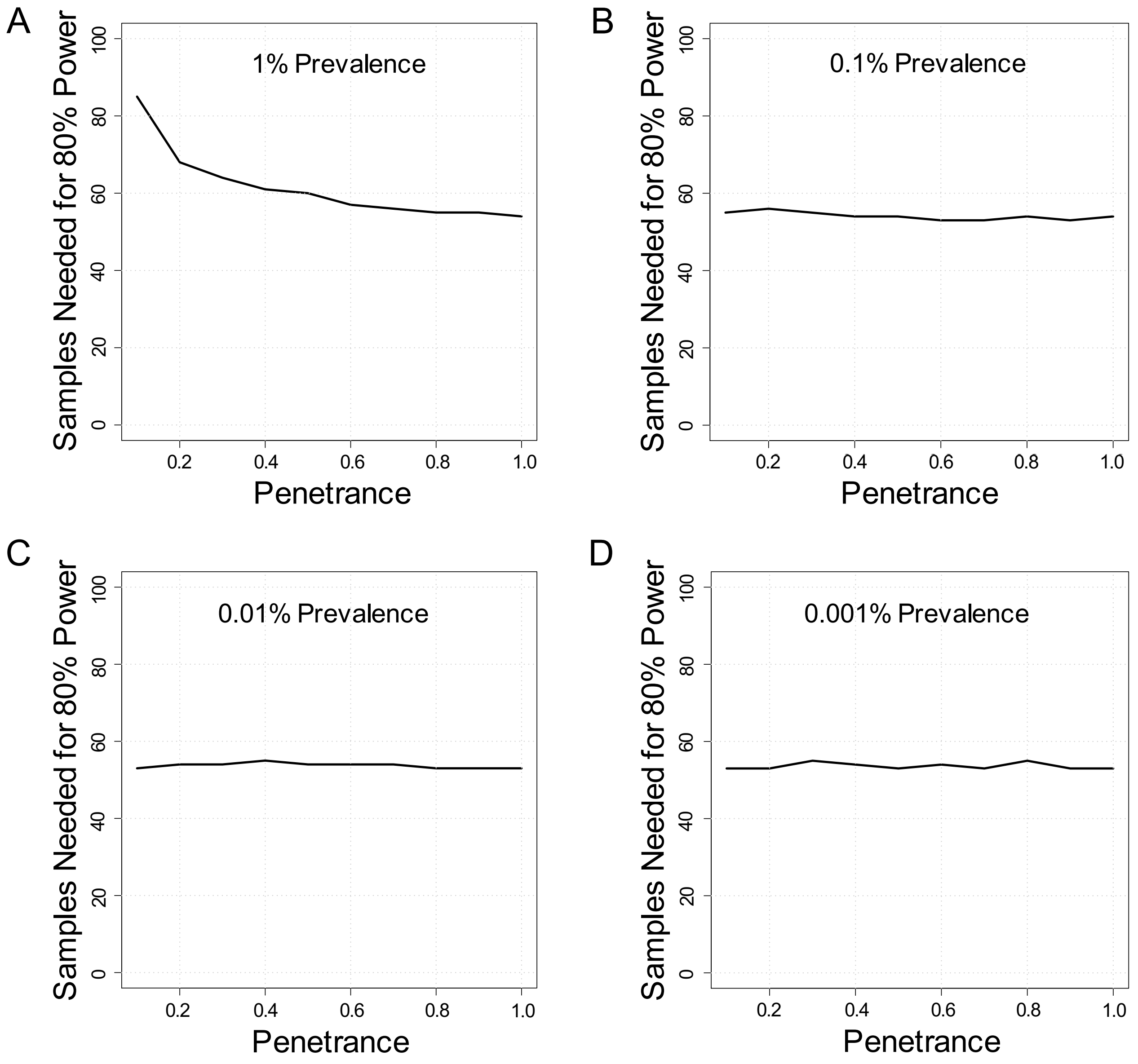


Figure S11: Effect of penetrance at *f_{case}* of 0.05

Effect of penetrance on sample sizes needed for 80% power to detect at least one disease-associated gene. Simulations were performed at varying disease prevalence of 1% (A), 0.1% (B), 0.01% (C) or 0.001% (D). Values of penetrance ranged from 0.1 to 1.0. Simulations were performed assuming a dominant disorder with 10 disease-associated genes, each of which contributes to 5% of cases (***f_{case}***=0.05).

Figure S12

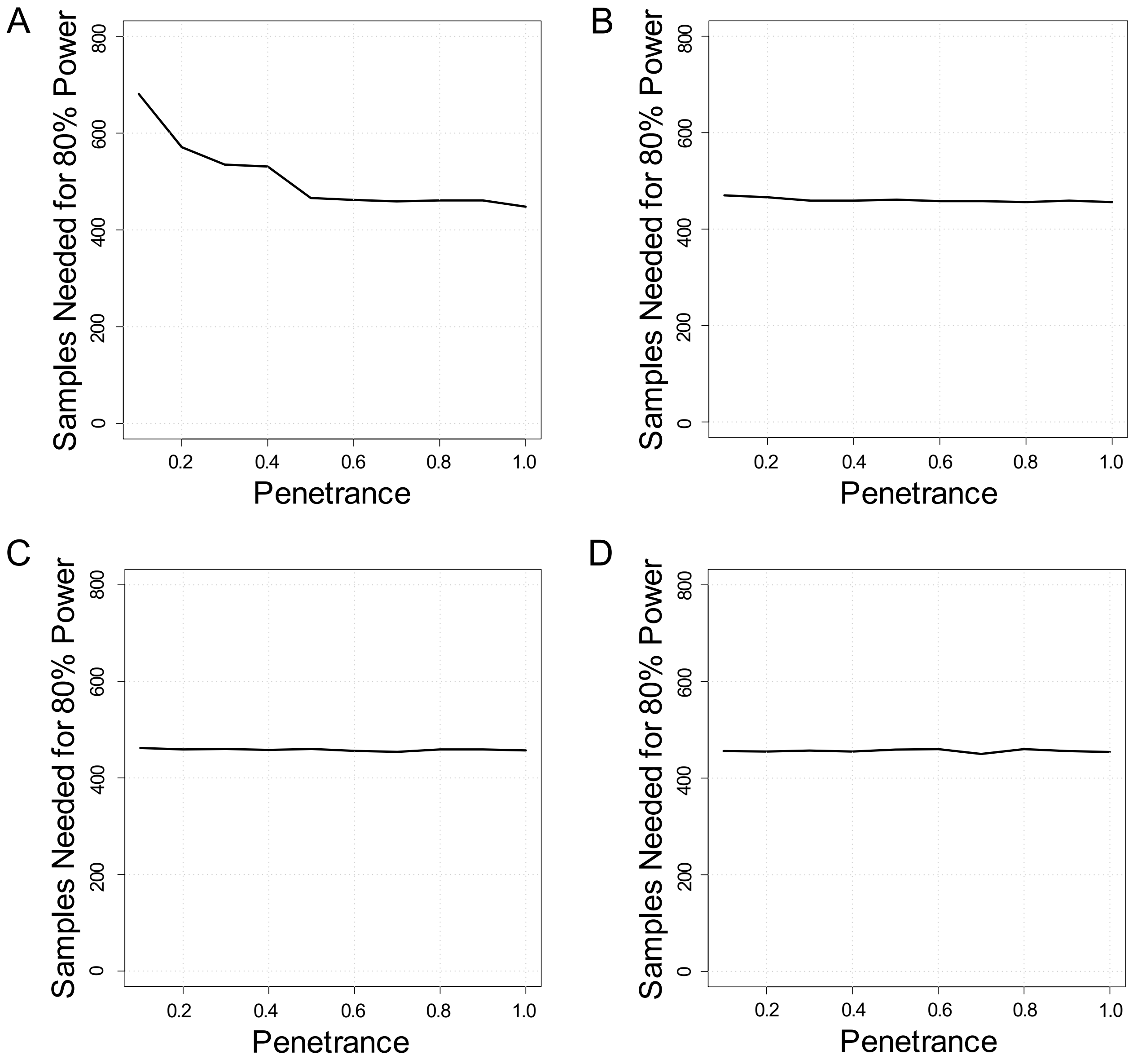


Figure S12: Effect of Penetrance at f_{case} of 0.01

Effect of penetrance on sample sizes needed for 80% power to detect at least one disease-associated gene. Simulations were performed at varying disease prevalence of 1% (A), 0.1% (B), 0.01% (C) or 0.001% (D). Values of penetrance ranged from 0.1 to 1.0. Simulations were performed assuming a dominant disorder with 100 disease-associated genes, each of which contributes to 1% of cases (f_{case}=0.01).

Figure S13

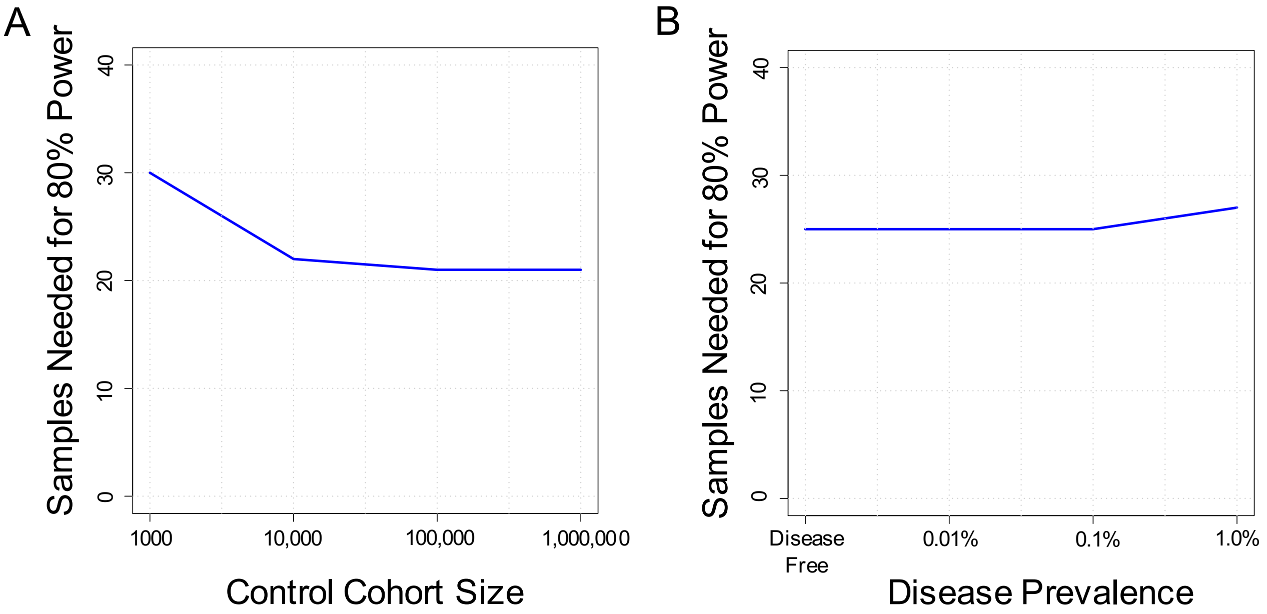


Figure S13: Effect of characteristics of control cohort

A) Effect of control cohort sizes on samples needed for 80% power to detect at least one disease-associated gene at different control cohort sizes (1000, 10000, 100000, 1000000). Simulations performed under base model.

B) Effect of using population-based cohort on samples needed for 80% power to detect at least one disease-associated gene. Simulations were performed assuming a disease-free control cohort, as well as disease prevalences of 0.01%, 0.1% and 1.0%. Simulations performed under base model.

Figure S14

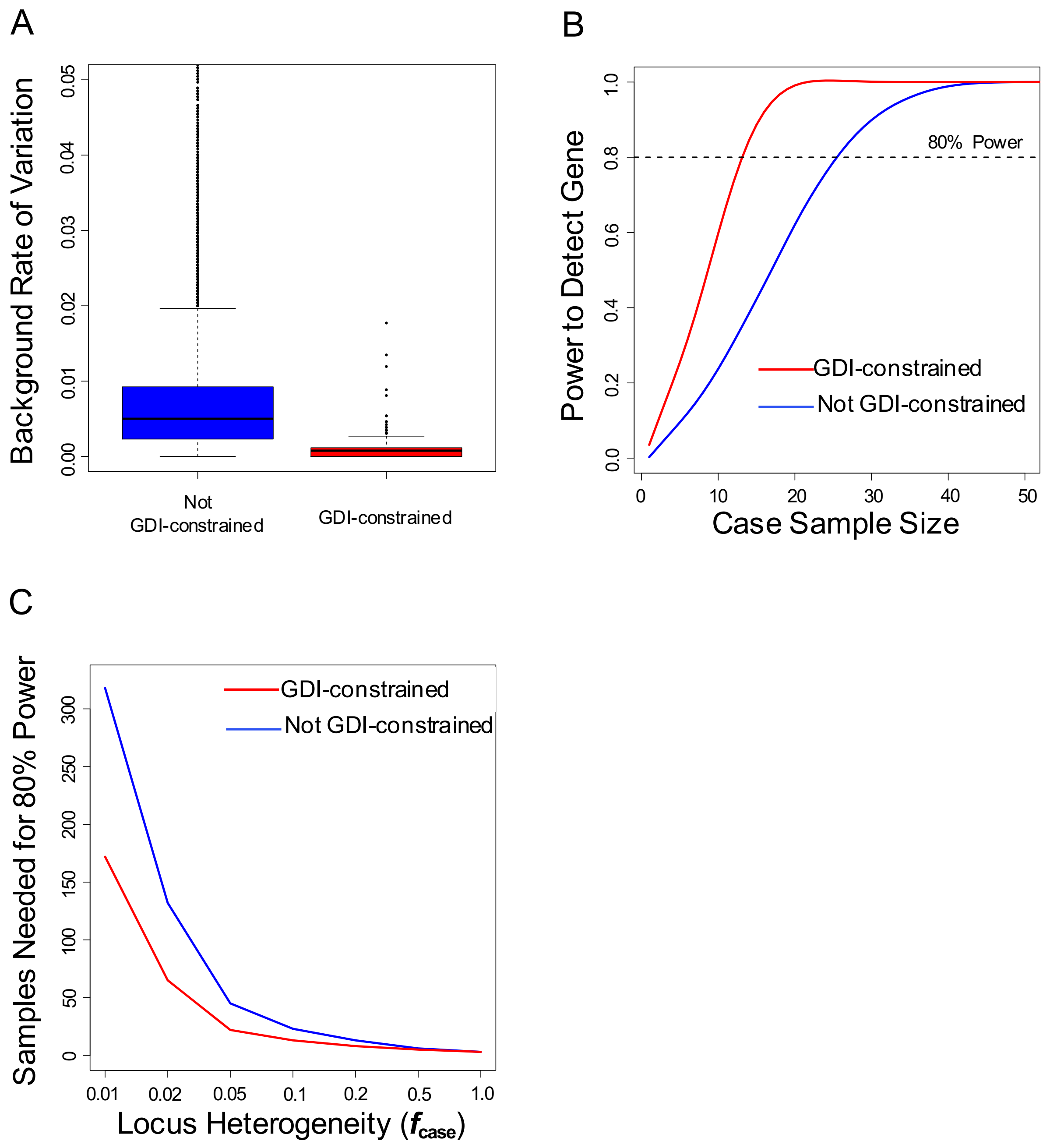


Figure S14: Power for GDI-constrained genes

A) Background rate of variation for GDI-constrained genes as compared to all other genes. Plot truncated at a background rate of 0.05.

B) Power to detect at least one gene for a disease with 10 associated genes, each of which contributes to 10% of cases (f_{case}=0.1). Analyses were performed for all genes in the genome (blue) as compared to GDI-constrained genes (red). All parameters are the same as Figure 3A.

C) Sample sizes needed to have 80% power to detect at least one gene associated with a disease using all genes in the genome (blue) as compared to GDI-constrained genes (red). All parameters are the same as Figure 3B.

Figure S15

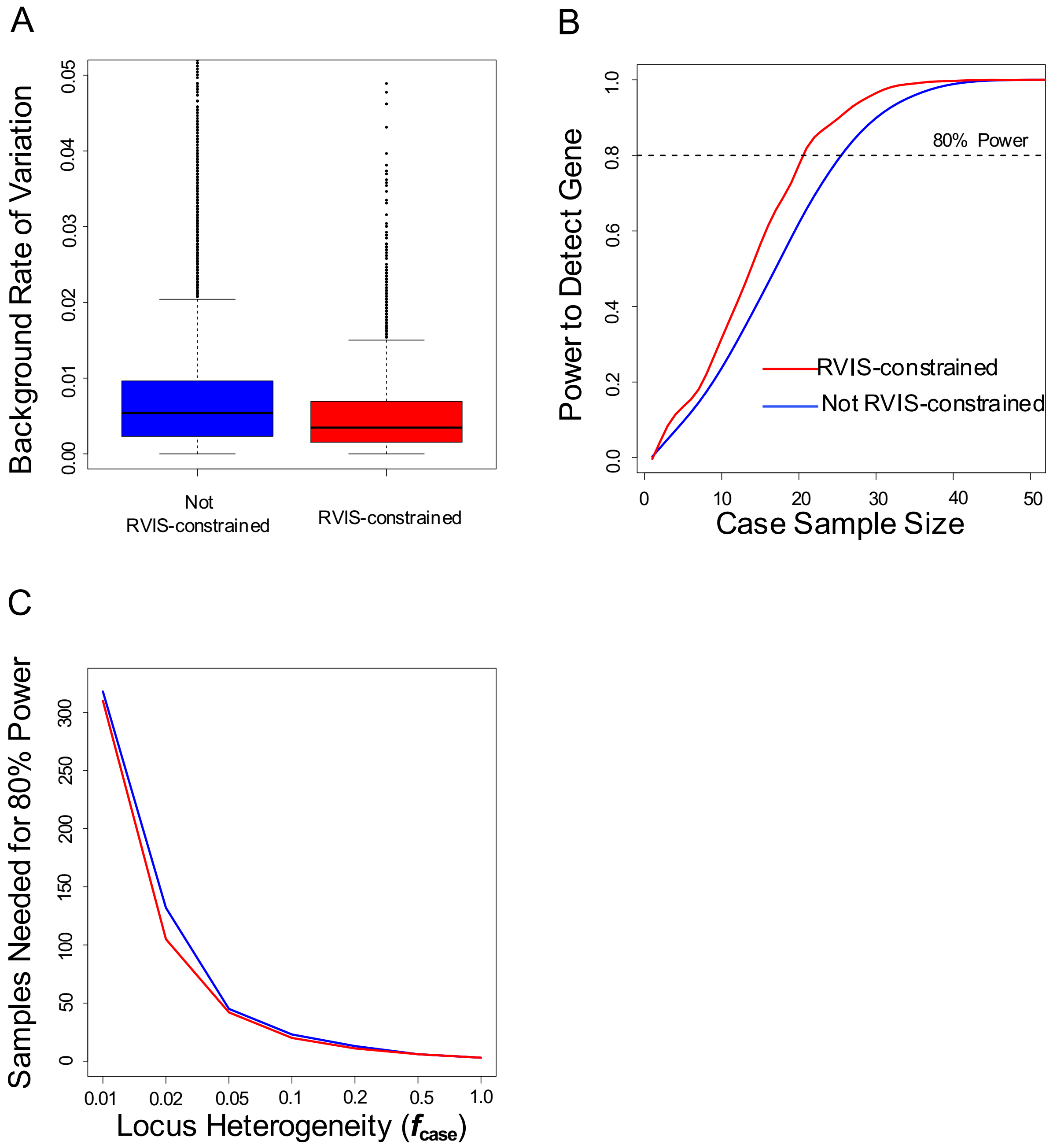


Figure S15: Power for RVIS-constrained genes

A) Background rate of variation for RVIS-constrained genes as compared to all other genes. Plot truncated at a background rate of 0.05.

B) Power to detect at least one gene for a disease with 10 associated genes, each of which contributes to 10% of cases (f_{case}=0.1). Analyses were performed for all genes in the genome (blue) as compared to RVIS-constrained genes (red). All parameters are the same as Figure 3A.

C) Sample sizes needed to have 80% power to detect at least one gene associated with a disease using all genes in the genome (blue) as compared to RVIS-constrained genes (red). All parameters are the same as Figure 3B.

Figure S16: Power for Known Disease Genes (Next Page)

A) Background rate of variation for dominant genes associated with disease according to OMIM compared with non-OMIM genes. Plot truncated at a background rate of 0.05.

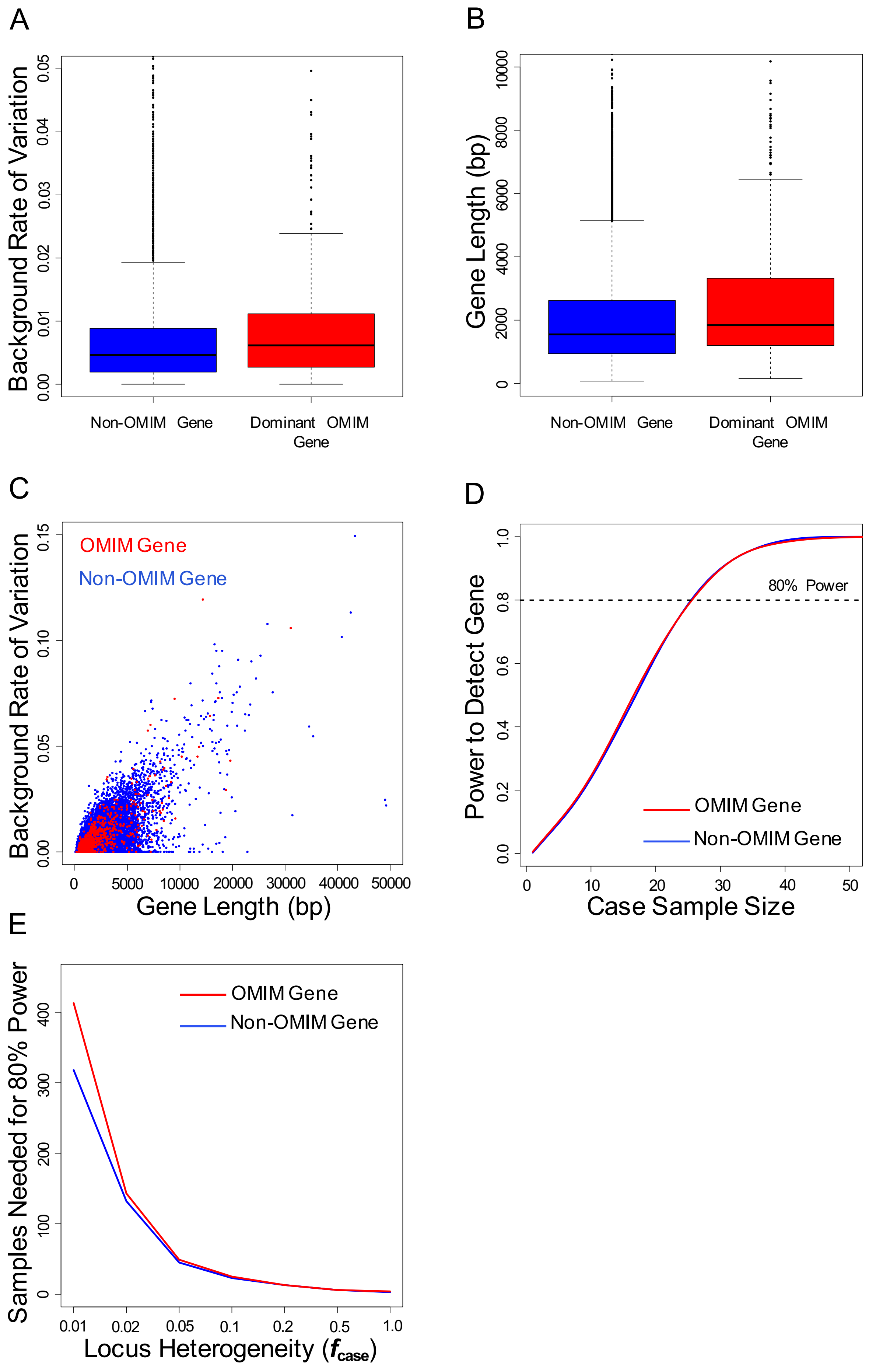
B) Length of coding region for dominant OMIM genes compared with non-OMIM genes.

C) Correlation of background rate of variation (y-axis) with coding gene length (x-axis). Red dots represent dominant OMIM genes, while all other genes in the genome are shown as blue dots.

D) Power to detect at least one gene for a disease with 10 associated genes, each of which contributes to 10% of cases ($\boldsymbol{f}_{\textbf{case}}$=0.1). Analyses were performed for all genes in the genome (blue) as compared to dominant OMIM genes (red). All parameters are the same as Figure 3A.

E) Sample sizes needed to have 80% power to detect at least one gene associated with a disease using all genes in the genome (blue) as compared to dominant OMIM genes (red). All parameters are the same as Figure 3B.

Figure S16


A
Background Rate of Variation
0.00
0.01
0.02
0.03
0.04
0.05
Non-OMIM Gene
Dominant OMIM Gene
B
Gene Length (bp)
0
2000
4000
6000
8000
10000
Non-OMIM Gene
Dominant OMIM Gene
C
OMIM Gene
Non-OMIM Gene
Background Rate of Variation
0.00
0.05
0.10
0.15
0
5000
10000
20000
30000
40000
50000
Gene Length (bp)
D
80% Power
Power to Detect Gene
0.0
0.2
0.4
0.6
0.8
1.0
OMIM Gene
Non-OMIM Gene
0
10
20
30
40
50
Case Sample Size
E
OMIM Gene
Non-OMIM Gene
Samples Needed for 80% Power
0
100
200
300
400
0.01
0.02
0.05
0.1
0.2
0.5
1.0
Locus Heterogeneity (f_{case})

References

1. Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.L., Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc. Natl. Acad. Sci. U. S. A. 112, 5473-5478.

2. Meynert, A.M., Ansari, M., FitzPatrick, D.R., Taylor, M.S. (2014). Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics 15, 247-2105-15-247.

3. Sulonen, A.M., Ellonen, P., Almusa, H., Lepisto, M., Eldfors, S., Hannula, S., Miettinen, T., Tyynismaa, H., Salo, P., Heckman, C. et al. (2011). Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol. 12, R94-2011-12-9-r94.

4. MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A. et al. (2014). Guidelines for investigating causality of sequence variants in human disease. Nature 508, 469-476.