*Supplementary Materials to*

**Power Analysis Provides Bounds for Genetic Architecture and Insights to Challenges Underlying Rare Variant Association Studies**

Table of Contents

* **Table S1.** First- and second-order approximations to the power of variance component tests.
* **Table S2.** Simulation scenarios and parameter values used for assessing accuracy of the first- and second - order approximations.
* **Table S3.** Summary of recently published association studies of rare variants.
* **Figure S1.** Evaluation of accuracy of the first-order approximation at various values of key parameters J=50, 100 and JC=30, 50 under simulation scenario S1 (MAF-independent EV)
* **Figure S2.** Evaluation of accuracy of the first-order approximation at various values of key parameters J=200, 400 and JC=10, 20, 30, 50 under simulation scenario S1 (MAF-independent EV).
* **Figure S3.** Evaluation of accuracy of the first-order approximation at various values of key parameters J=50, 100 and JC=10, 20, 30, 50 under simulation scenario S2 (MAF-independent ).
* **Figure S4.** Evaluation of accuracy of the first-order approximation at various values of key parameters J=50, 100 and JC=10, 20, 30, 50 under simulation scenario S3 (MAF-log-dependent ).
* **Figure S5.** Evaluation of accuracy of the second-order approximation at various values of key parameters J=200, 400 and JC=10, 20, 30, 50 under simulation scenario S1 (MAF-independent EV).
* **Figure S6.** Evaluation of accuracy of the second-order approximation at various values of key parameters J=50, 100 and JC=10, 20, 30, 50 under simulation scenario S2 (MAF-independent ).
* **Figure S7.** Evaluation of accuracy of the second-order approximation at various values of key parameters J=50, 100 and JC=10, 20, 30, 50 under simulation scenario S3 (MAF-log-dependent ).
* **Figure S8**. Evaluation of accuracy of the first-order approximation for sum-based tests at various values of key parameters J=50, 100 and JC=10, 20 under simulation scenario S1 (MAF-independent EV).
* **Figure S9**. Empirical distribution of MAF and the number of rare variants per locus estimated from Exome Aggregation Consortium (ExAC) that model one observed in the study of educational attainment that used exome sequencing.
* **Figure S10**. Empirical distribution of MAF and the number of rare variants per locus estimated from ExAC that model one observed in the study of blood pressure that used Exome Chip platform.
* **Figure S11.** Bounds for genetic architecture based on results reported in studies of education attainment (EA) and blood pressure (BP) under assumption of independence between MAF and genetic effects
* **Figure S12**. Effects of sensitivity and specificity for apriori variant screening on the power of variance component and burden tests under J=100, JC=10 and simulation scenario S1 (MAF-independent EV).
* **Figure S13.** Effects of sensitivity and specificity for apriori variant screening on the power of variance component and burden tests under J=50, 100, JC=10 and simulation scenario S2 (MAF-independent ).
* **Figure S14**. Bounds for genetic architecture based on results from single SNP analysis on rare variants (MAF<5%) reported in studies of education attainment (EA) and blood pressure (BP)

**Table S1.** **The first- and second-order approximations to the power of variance component tests.** Power of the variance component test depends on the first four cumulants (see formula 1). The first-order approximation rewrites cumulants ck as function of two key parameters: *J* -number of rare variants in a locus, *EV –* proportion of phenotypic variation explained by a locus and – weighting based on sample size *N* and MAF . The second-order approximation additionally incorporates - number of causal variants in a locus. For more details see Appendix A, B and C.

|  |  |  |  |
| --- | --- | --- | --- |
| Underlying Genetic Architecture[[1]](#footnote-1) | Mathematical Representation[[2]](#footnote-2) | First-order Approximation[[3]](#footnote-3) | Second-order Approximation[[4]](#footnote-4) |
| Proportion of variation explained by a variant is independent of its MAF |  |  |  |
| Genetic effect is independent of MAF |  |  |  |
| Genetic effect is function of MAF |  | where is an average change in due to one-unit change in . | where is an average change in due to one-unit change in . |

**Table S2.** **Simulation scenarios and parameter values for assessing accuracy of the first-order approximations**. Scenario S1 ('MAF-independent EV') assumes MAFs and EVs are mutually independent. Scenario S2 ("MAF-independent ") assumes MAFs and effect sizes defined in the unit of per copy of an allele () are mutually independent. Scenario S3 ("MAF-log-dependent ") assumes MAFs and effect sizes are dependent through log10 function. For the exact calculations for the Scenario S1, we directly generate EVjs from specific value EV and for Scenarios S2 and S3, we first generate then calculate corresponding EVjs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Parameters | Parameter Values | Parameters used in | | |
| First-order Approximation | Second-order Approximation | Exact-Calculations |
| *N* | Effective sample size | 10,000 | *Yes* | *Yes* | *Yes* |
| *J* | Total number of SNPs | 50, 100, 200, 400 | *Yes* | *Yes* | *Yes* |
| *EV* | Coefficient of explained phenotypic variation by a locus | Ranges between 0.001 and 0.01 | *Yes* | *Yes* | *Yes* |
| *pj* | MAF of SNP *j* | Gamma (1,300) with minimum and maximum values at 0.0002 and 0.01 | *Yes* | *Yes* | *Yes* |
| *JC* | Number of causal SNPs | 10, 20, 30, 50 | *No* | *Yes* | *Yes* |
| *Scenario S1 ("MAF-independent EV")* | | | | | |
| *EV****j*** | Coefficient of explained variation by SNP *j* | Randomly selected for each causal SNP under the constrain: | *No* | *No* | *Yes* |
| *Scenario S2 ("MAF-independent ")* | | | | | |
|  | MAF adjusted average effect of SNP j | MAF adjusted average effect of rare variant | *No* | *No* | *Yes* |
|  | Genetic effect of *jth* variant | , then coefficients of explained variations are scaled by the constant so that | *No* | *No* | *Yes* |
| *Scenario S3 ("MAF-log-dependent ")* | | | | | |
| *C* | Adjustment |  | *No* | *No* | *Yes* |
|  | Genetic effect of *jth* variant | then coefficients of explained variations are scaled by the constant so that | *No* | *No* | *Yes* |

**Table S3.** **Summary of recently published association studies of rare variants.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Study** | **Genetic Platform,**  **Sample Size** | **Trait** | **Association analysis with rare variants[[5]](#footnote-5)** | | |
| **Gene based**  **tests** | **# of significant findings with gene based test** | **# of significant rare variant findings with single variant test** |
| A polygenic burden of rare disruptive mutations in schizophrenia1 | WES: 5000 | Case/Control ~ 1/1 | SKAT and Burden Tests with a gene as a unit | No findings | NA |
| Whole-genome sequencing identifies EN1 as determinant of bone density and fracture2 | WGS: ~2,900  WES: ~3,500  Imputation[[6]](#footnote-6): ~26,500 | Multiple QTs | SKAT with sliding window with 30 SNP | 0-1 gene[[7]](#footnote-7) | 0-1 rare variant |
| The UK10K project identifies rare variants in health and disease3 | WGS: ~3,500  Imputation: ~9,200 | Multiple QTs | SKAT with sliding window with 50 SNP | 0-1 genes | 0-1 rare variants |
| The genetic architecture of type 2 diabetes4 | WGS: ~2,600  WES: ~13,000  Exome Array: 80,000 | Case/Control ~ 1/1  Case/Control ~ 1/1  Case/Control ~ 1/2 | SKAT with a gene as a unit | No findings | 6 rare variants[[8]](#footnote-8) |
| Ultra-rare disruptive and damaging mutations influence educational attainment in general population5 | WGS: ~2,700  WES: ~11,300 | QT | Burden Tests with a gene as a unit | No findings | NA |
| Inherited coding variants at the CDKN2A locus influence susceptibility to acute lymphoblastic leukemia in children6 | Exome Chip: ~12,000 | Case/Control ~ 1/5 | SKAT with a gene as a unit | No findings | 1 rare variant |
| Meta-analysis of rare and common exome chip variants identifies *S1PR4* and other loci influencing blood cell traits7 | Exome Chip: ~52,000 | Multiple QTs | SKAT and Burden Tests with a gene as a unit | 1-2 genes | 1-3 rare variants |
| Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility8 | Exome Chip: ~61,000 | QT | SKAT with a gene as a unit | 1 gene | No findings |
| Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci9 | Exome Chip: ~145,000 | Multiple QT | SKAT and Burden Tests with a gene as a unit | 1-2 genes | 1-2 rare variants |

**Figure S1.** **Evaluation of the accuracy of the first order approximations at various values of key parameters J=50, 100 and JC=30, 50 under simulation scenario S1 (MAF-independent EV).**  J-number of rare variants in a locus and JC -number of causal variants in a locus. Exact Formula represents estimated average power using exact theoretical formulas for the SKAT test statistic. The First Order Approximation represents estimated average power using the first order approximation for the SKAT test statistic.



**Figure S2.** **Evaluation of the accuracy of the first order approximations at various values of key parameters J=200, 400 and JC=10, 20, 30, 50 under simulation scenario S1 (MAF-independent EV).**  J-number of rare variants in a locus and JC -number of causal variants in a locus. Exact Formula represents estimated average power using exact theoretical formulas for the SKAT test statistic. The First Order Approximation represents estimated average power using the first order approximation for the SKAT test statistic.



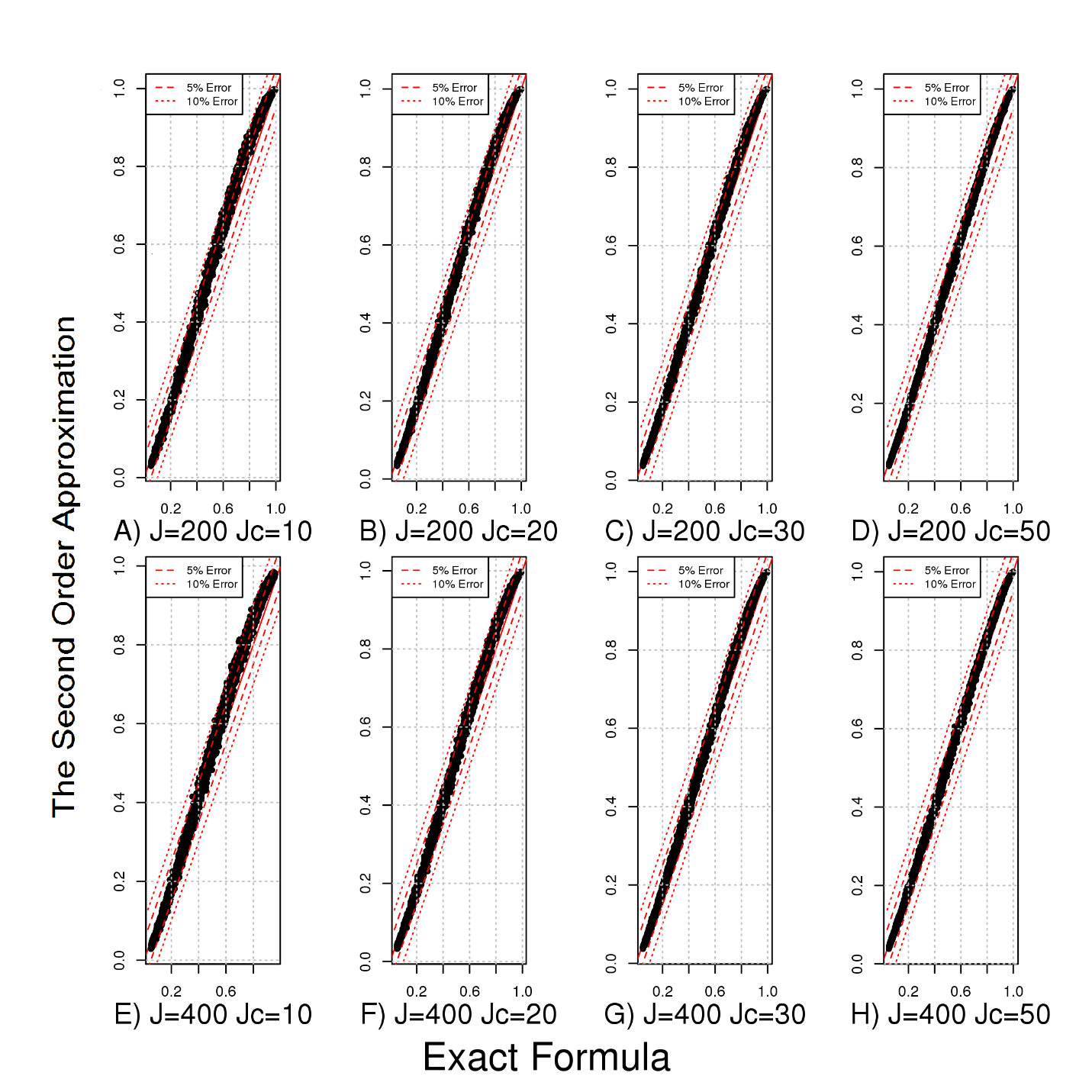
**Figure S3. Evaluation of the accuracy of the first order approximations at various values of key parameters J=50, 100 and JC=10, 20, 30, 50 under simulation scenario S2 (MAF-independent ).**  J-number of rare variants in a locus and JC -number of causal variants in a locus. Exact Formula represents estimated average power using exact theoretical formulas for the SKAT test statistic. The First Order Approximation represents estimated average power using the first order approximation for the SKAT test statistic.

**

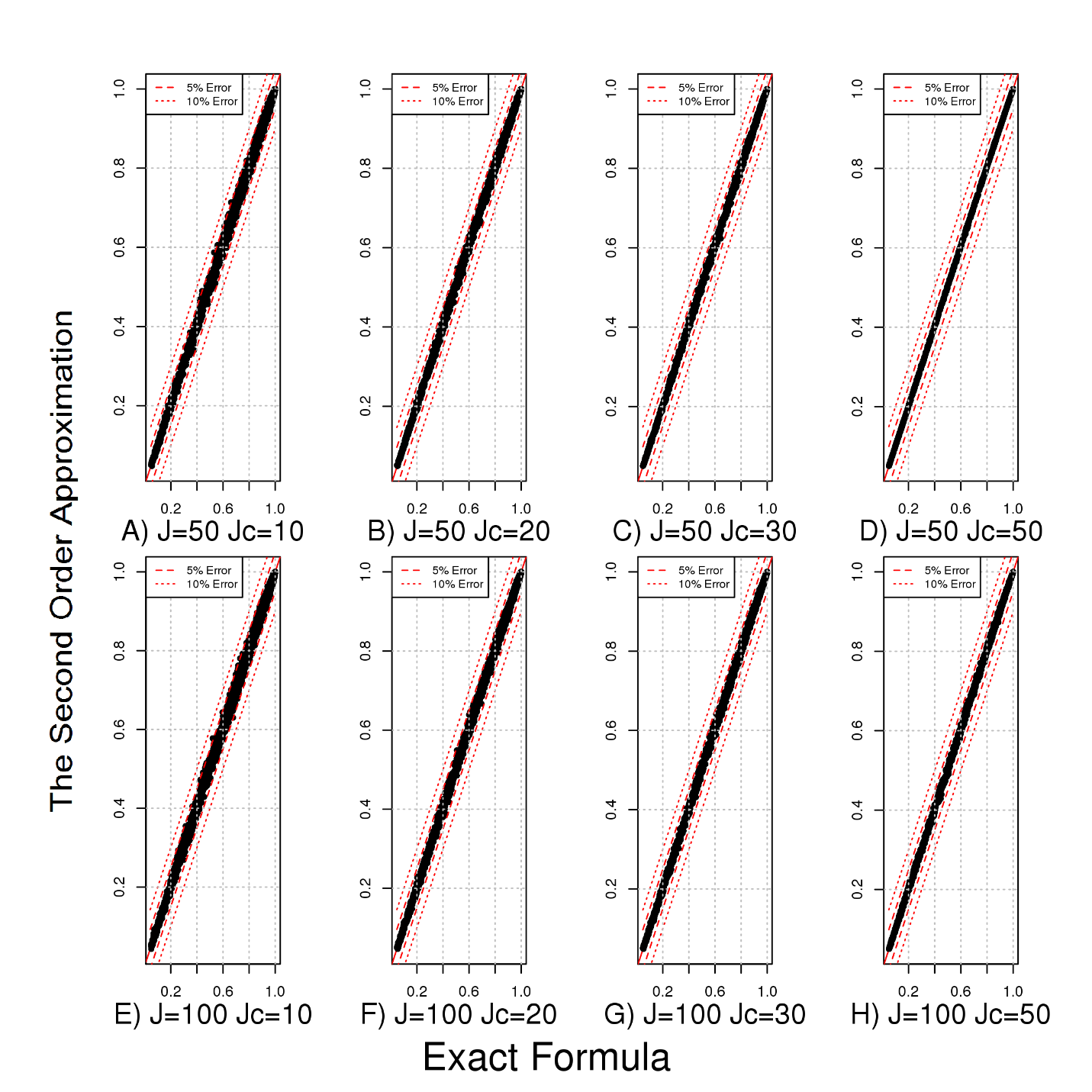
**Figure S4.** **Evaluation of the accuracy of the first order approximations at various values of key parameters J=50, 100 and JC=10, 20, 30, 50 under simulation scenario S3 (MAF-log-dependent ).** J-number of rare variants in a locus and JC -number of causal variants in a locus. Exact Formula represents estimated average power using exact theoretical formulas for the SKAT test statistic. The First Order Approximation represents estimated average power using the first order approximation for the SKAT test statistic.



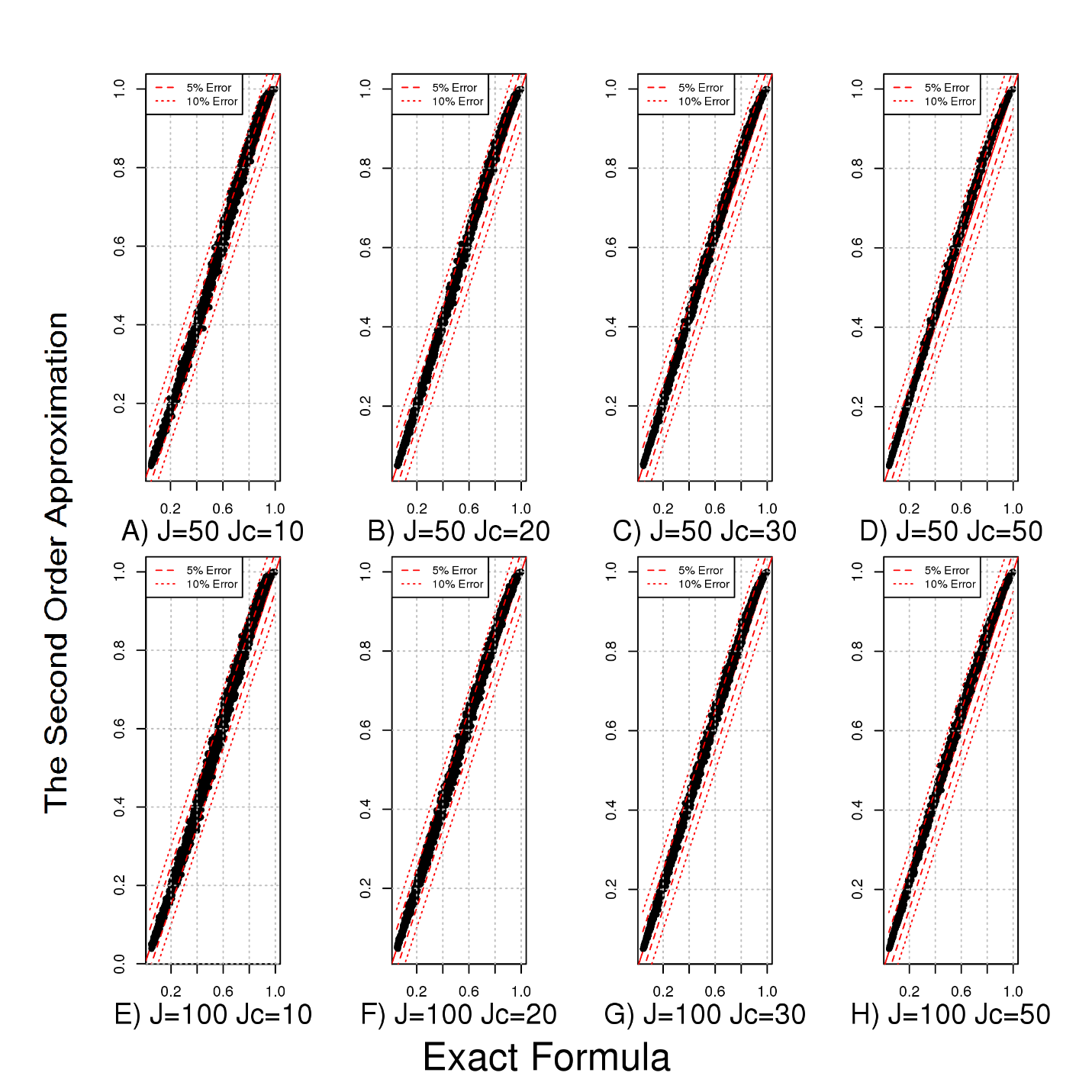
**Figure S5.** **Evaluation of the accuracy of the second order approximation at various values of key parameters J=200, 400 and JC=10, 20, 30, 50 under simulation scenario S1 (MAF-independent EV).**  J-number of rare variants in a locus and JC -number of causal variants in a locus. Exact Formula represents estimated average power using exact theoretical formulas for the SKAT test statistic. The Second Order Approximation represents estimated average power using the first order approximation for the SKAT test statistic.



**Figure S6.** **Evaluation of the accuracy of the second order approximations at various values of key parameters J=50, 100 and JC=10, 20, 30, 50 under simulation scenario S2 (MAF-independent ).** J-number of rare variants in a locus and JC -number of causal variants in a locus. Exact Formula represents estimated average power using exact theoretical formulas for the SKAT test statistic. The Second Order Approximation represents estimated average power using the second order approximation for the SKAT test statistic.

**

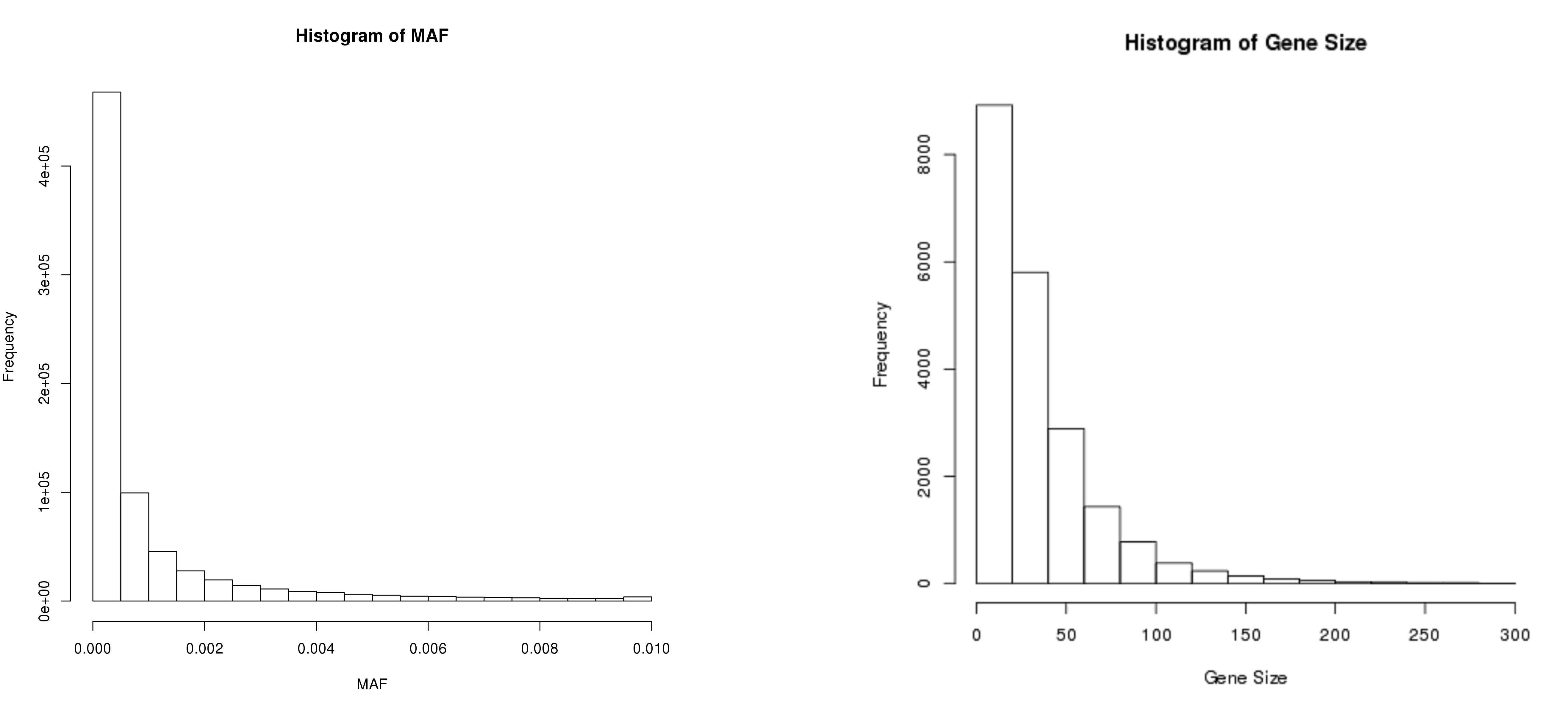
**Figure S7**. **Evaluation of the accuracy of the second order approximations under simulation at various values of key parameters J=50, 100 and JC=10, 20, 30, 50 under simulation scenario** **S3 (MAF-log-dependent ).** Exact Formula represents estimated average power using exact theoretical formulas for the SKAT test statistic. The Second Order Approximation represents estimated average power using the second order approximation for the SKAT test statistic.



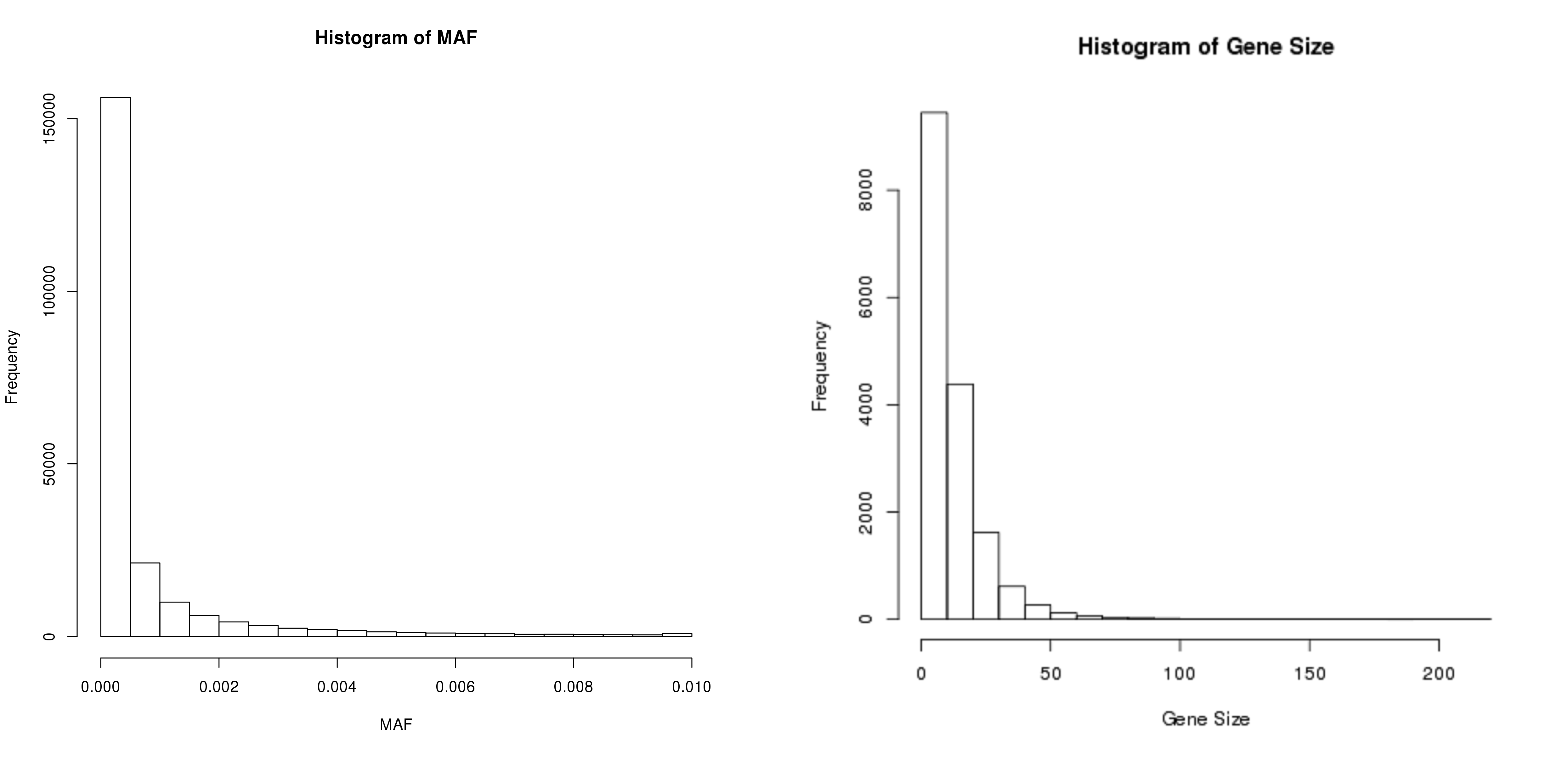
**Figure S8. Evaluation of the accuracy of the first order approximations for the burden test statistic at various values of key parameters J=50, 100 and JC=10, 20 under simulation scenario S1 (MAF-independent EV).**  Exact Formula represents estimated average power using exact theoretical formulas for the burden test statistic. The Approximation represents estimated average power using the first order approximation for the burden test statistic.



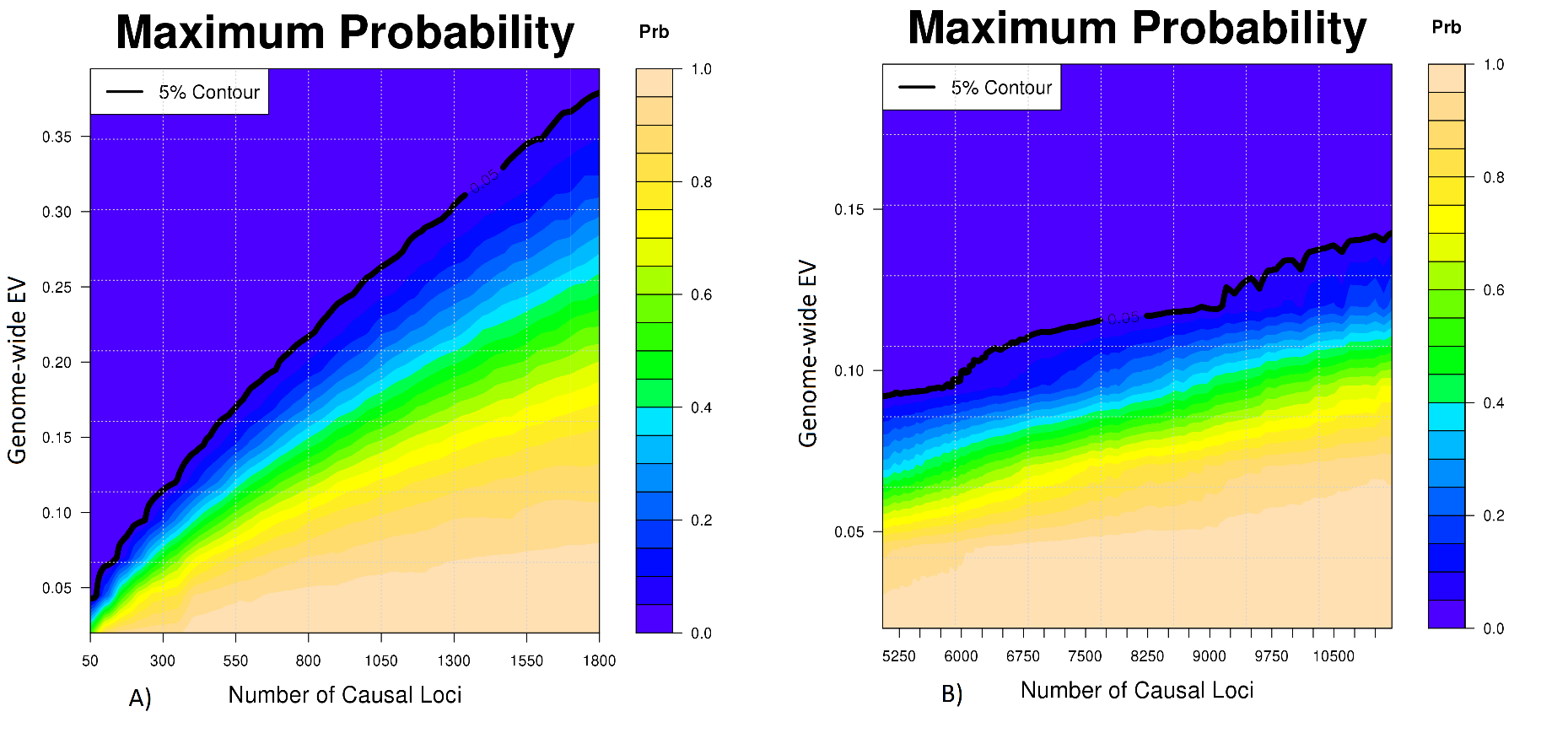
**Figure S9.** **Empirical distribution of MAF and the number of rare variants per locus estimated in Exome Aggregation Consortium (ExAC)10 that model one observed in the study of educational attainment that used exome sequencing5.** We observe **743,094** variants with MAFs ranging between 0.0001 and 0.01 and 20,895 genes with at least two rare variants. Average number of variants in a locus is **35.5**.



**Figure S10. Empirical distribution of MAF and the number of rare variants per locus estimated from ExAC that model one observed in the study of blood pressure that used Exome Chip platform9.** We observe **215,674** variants with MAFs<0.01 and **16,000** genes with at least two rare variants.; Average number of variants in a locus is **13***.*

****

**Figure S11.** **Bounds for genetic architecture based on results reported in studies of education attainment (EA) and blood pressure(BP) under assumption of independence between MAF and genetic effects .** Panel (A) shows maximum probability of observing no discoveries in the EA study, which used whole exome sequencing platform, as a function of the number of underlying causal loci K and the total variation explained by them with a sample size of 14,000. Panel (B) shows the maximum probability of observing three statistical significant discoveries in the BP study, which used exome chip, as a function of the number of underlying causal loci K and the total variation explained by them with a sample size of 140,000. In both cases, it’s assumed gene-based tests have been performed using the SKAT test statistics at the level of  2.5·10-6. Probabilities are estimated by (2) and assumption of independence between MAF and genetic effects . The black line shows approximate contours (bounds) corresponding to probability of 5%.



**Figure S12.** **Effects of sensitivity and specificity for apriori variant screening on the power of variance component and burden tests under simulation scenario S1 (MAF-independent EV).** Number of variants in a locus is set to **J=100** and number of causal variants to **JC=10.** The setting corresponds toa baseline power (i.e. if all variants were included in the study) of 40%, and 11% for variance component and sum-based tests, respectively**.**

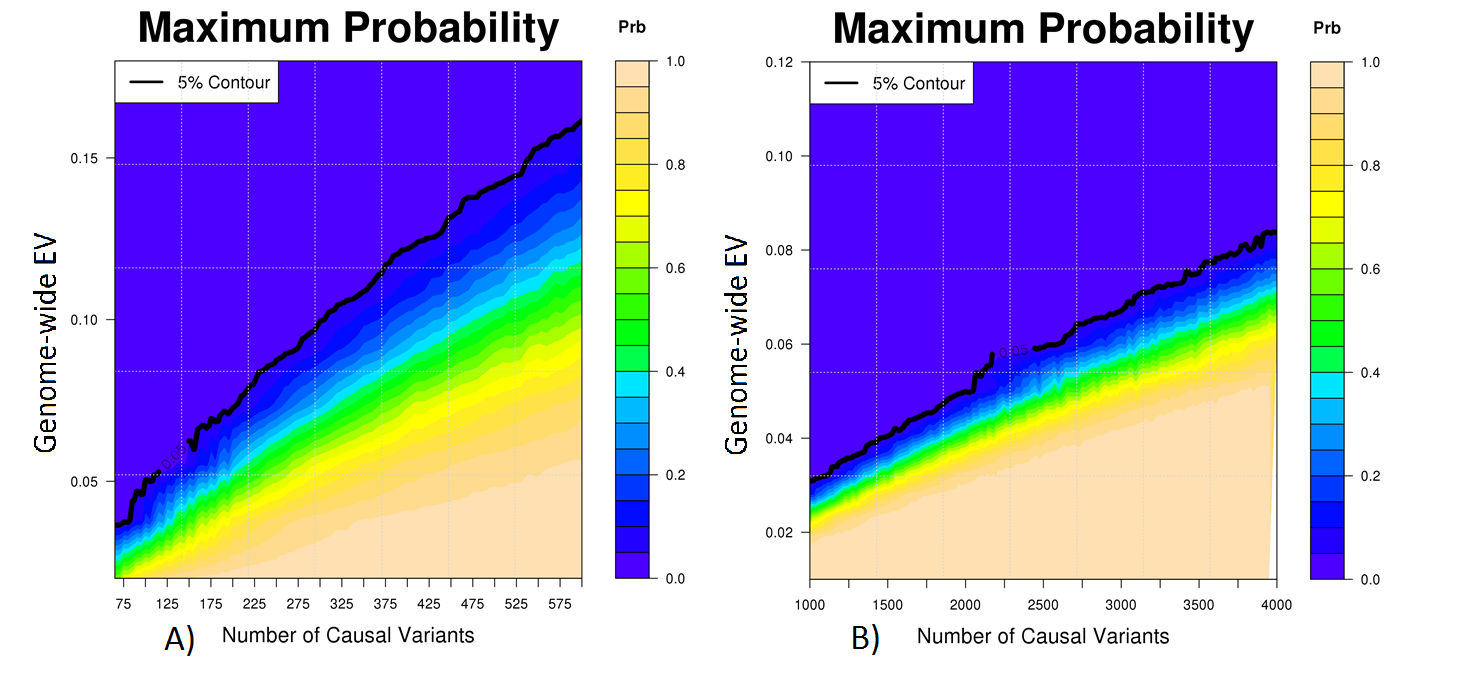
**

**Figure S13.** **Effects of sensitivity and specificity for apriori variant screening on the power of variance component and burden tests under simulation scenario S2 (MAF-independent ).** Number of variants in a locus is set to **J=50, 100** and number of causal variants to **JC=10.** The setting (A) **J=50** and **JC=10** corresponds toa baseline power (i.e. if all variants were included in the study) of 40%, and 5.6% for variance component and sum-based tests, respectively**.** The setting (B) **J=100** and **JC=10** corresponds toa baseline power (i.e. if all variants were included in the study) of 40%, and 5.6% for variance component and sum-based tests, respectively

**A)**

**B)**

**Figure S14**. **Bounds for genetic architecture based on results from single SNP analysis reported in studies of education attainment (EA) and blood pressure(BP).** Panel (A) shows maximum probability of observing no discoveries in the EA study, which used whole exome sequencing platform, as a function of the number of underlying causal SNPs K and the total variation explained by them with a sample size of 14,000. Panel (B) shows the maximum probability of observing three statistical significant discoveries in the BP study, which used exome chip, as a function of the number of underlying causal SNPs K and the total variation explained by them with a sample size of 140,000. In both cases, it’s assumed single SNP analysis was performed at the level of 5·10-8. Probabilities are estimated by (2). The black line shows approximate contours (bounds) corresponding to probability of 5%.



**References**

1. Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kahler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature 506, 185-+.

2. Zheng, H.F., Forgetta, V., Hsu, Y.H., Estrada, K., Rosello-Diez, A., Leo, P.J., Dahia, C.L., Park-Min, K.H., Tobias, J.H., Kooperberg, C., et al. (2015). Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 526, 112-117.

3. Consortium, U.K., Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature 526, 82-90.

4. Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D.J., et al. (2016). The genetic architecture of type 2 diabetes. Nature 536, 41-47.

5. Ganna, A., Genovese, G., Howrigan, D.P., Byrnes, A., Kurki, M.I., Zekavat, S.M., Whelan, C.W., Kals, M., Nivard, M.G., Bloemendal, A., et al. (2016). Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. Nat Neurosci 19, 1563-1565.

6. Xu, H., Zhang, H., Yang, W.J., Yadav, R., Morrison, A.C., Qian, M.X., Devidas, M., Liu, Y., Perez-Andreu, V., Zhao, X.J., et al. (2015). Inherited coding variants at the CDKN2A locus influence susceptibility to acute lymphoblastic leukaemia in children. Nature Communications 6.

7. Group, C.C.H.W. (2016). Meta-analysis of rare and common exome chip variants identifies S1PR4 and other loci influencing blood cell traits. Nat Genet 48, 867-876.

8. Wessel, J., Chu, A.Y., Willems, S.M., Wang, S., Yaghootkar, H., Brody, J.A., Dauriz, M., Hivert, M.F., Raghavan, S., Lipovich, L., et al. (2015). Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nat Commun 6, 5897.

9. Liu, C., Kraja, A.T., Smith, J.A., Brody, J.A., Franceschini, N., Bis, J.C., Rice, K., Morrison, A.C., Lu, Y., Weiss, S., et al. (2016). Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. Nat Genet 48, 1162-1170.

10. Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., Fennell, T., O'Donnell-Luria, A., Ware, J., Hill, A., Cummings, B., et al. (2015). Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv.

1. Three common assumptions about relationship between genetic effect of a SNP and MAF [↑](#footnote-ref-1)
2. Mathematical representation of the relationship between genetic effect of a SNP and MAF [↑](#footnote-ref-2)
3. First-order approximations of cumulants of a variance component test [↑](#footnote-ref-3)
4. Second-order approximations of cumulants of a variance component test [↑](#footnote-ref-4)
5. Variants with MAF<1%. [↑](#footnote-ref-5)
6. Variants with MAF>0.1% [↑](#footnote-ref-6)
7. Identified by single SNP analysis [↑](#footnote-ref-7)
8. Variants with MAF<5% [↑](#footnote-ref-8)