Reducing the Impact of Minor Allele Frequency and Linkage Disequilibrium on Variable Importance Measures

Raymond Walters[1], Charles Laurin[1], and Gitta Lubke[1,2]
[1] University of Notre Dame [2] VU University Amsterdam

Note on Data Generation

The reported simulations involving only SNPs with no true association with the phenotype rely on generating a simulated SNP with a specified MAF m that is in LD with a neighboring empirical SNP with a specified correlation ρ. In order to achieve the desired correlation with the empirical SNP, the MAF of the simulated SNP must be sufficiently close to the MAF of the empirical SNP to allow the necessary degree of similarity between the two SNPs. As a result, it is not possible to generate data for this set of simulations under all combinations of the specified MAF and LD conditions.

Biswas and Hwang (2002) provide equations that can be used to determine which combinations of MAF and LD can be achieved at the population level. For an empirical SNP with MAF $m = .286$, and a simulated SNP with some MAF, it is possible to determine the upper bound of the correlation between this SNP and a simulated SNP. The results of this computation are shown in Supplementary Figure 10. This figure indicates, for instance, that we will only be able to simulate a SNP with $\rho = .9$ for MAF $m = .3$; this LD condition will need to be omitted for the other MAF conditions.

In a finite sample, however, it is possible to slightly exceed these boundaries. With sufficient iterations of the data generation procedure, these larger sample values can be attained consistently. For the current simulations, this allows us to consider MAF $m = .05$ and $m = .10$ with $\rho = 0.6$. We include these conditions to provide the fullest possible coverage of MAF and LD values for the simulations.

Supplementary Results

Impact of Number of Subsets Including a SNP

The proposed LD subsetting algorithm creates a group of subsets that are guaranteed to include each SNP one or more times. Notably, not all SNPs are guaranteed to be included in the same number of subsets. Indeed, we anticipate that SNPs not in LD will be included in more subsets since they are more likely to be added to subsets after the initial selection.

There is room for concern, then, that the final importance of a SNP could be affected by the number of subsets it is included in as a part of LD subsetting. Analyses using the LD subsetting intentionally average each SNPs importance only across subsets containing that SNP to avoid bias from the number of subsets, an effect of the number of subsets containing a given SNP is still theoretically possible. To alleviate this concern we provide evidence that there is no relationship between a SNP's importance and the number of LD subsets it is placed in by the LD subsetting algorithm.

Using data from the simulation reported in Section 3.1, Supplementary Table 5 presents the observed correlation between the number of subsets containing a SNP and the final importance of the SNP with LD subsetting. Results are reported using Kendall's τ to accommodate the skewed distribution of importances and allow for possible non-linear relationships between importances and subsets. In addition, the analysis of the impact of number of LD subsets is stratified by the LD and MAF conditions of the simulation to avoid confounding the results with the simulation settings. The results in Supplementary Table 5 show no significant association between the number of subsets selected to include a given SNP and that SNP's final importance for any of the SL methods considered in this study. Therefore the simulation results suggest that the LD subsetting algorithm is not adversely affected by averaging the importance of each SNP across a differing numbers of subsets.

Additional Results for Random Forests

For the majority of conditions the results from random forests (RF) are highly consistent with the results of gradient boosting (GBM), but for completeness we present the full results of RF here.

Impact of LD and MAF with null SNPs

Without subsetting for linkage disequilibrium (LD), RF shows strong effects on LD and minor allele frequency (MAF) on variable importance for non-functional SNPs. For the Gini importance there is a very clear trend towards lower importance for SNPs with low MAF and SNPs in high LD (Supplementary Fig. S3a). The results for the MDA importance also show a significant effect of MAF at $m = .05$ and $m = .30$. Significant results may reflect an effect of MAF on the variability of importance scores rather than the median, which would be consistent with the findings of Nicodemus and Malley (2009). On the other hand, the observed medians suggest a trend towards higher significance for SNPs in strong LD (Supplementary Fig. S4a), with the reduced variability at $m = .05$ and the additional LD condition $\rho = .9$ at $m = .30$ providing improved power to detect an effect of MAF on the median importance. The reason for the deviation of these results from those observed by Nicodemus and Malley (2009) remains unclear, and should be the subject of future research.

Impact of LD subsetting with null SNPs

The impact of the LD subsetting procedure on the Gini importance is dramatic; the influence of LD appears to be completely eliminated (Supplementary Fig. S3b). The reduced influence of LD is also visible for the MDA importance (Supplementary Fig. S4b). The Kruskal-Wallis test confirms that there is no longer a significant effect of LD on the Gini or MDA importance (Supplementary Table S1). Strong influences of MAF remain, however, with highly significant effects observed for the Gini importance, and apparent differences in the variability of the MDA importance linked to MAF.

Sensitivity of RF with Functional SNPs

In the simulation with functional SNPs the Gini importance is influenced by both LD and MAF, while the MDA importance shows signs of only being weakly influenced by MAF. Specifically, for the Gini importance lower importances are observed for functional SNPs in LD and for SNPs with extreme MAF (eg. near 0 or .5; Supplementary Fig. S6a). For the MDA importance, higher MAF is associated with lower observed importance values (Supplementary Fig. S7a). The resulting detection rate for identifying functional SNPs is reported in Supplementary Table S3 and acts as the baseline for assessing the potential benefit of LD subsetting.

After LD subsetting, the effect of LD on the Gini importance for functional SNPs appears to be reduced, though the effect remains significant for SNPs with moderate MAF ($m = .30$; Supplementary Table S4). On the other hand, the effect of MAF on the Gini importance is noticeably stronger after LD subsetting, with higher importances corresponding to high MAF (Supplementary Fig. S6b). This strong effect of MAF drastically reduces the ability of the Gini importance to detect functional SNPs with low or moderate MAF (Supplementary Table S3).

Meanwhile the MDA importance for functional SNPs remains unaffected by LD after LD subsetting. While plotting the importances from each condition suggest a continued trend toward lower importance values for SNPs with high MAF (Supplementary Fig. S7b), the effect is non-significant (Supplementary Table S4). LD subsetting also appears to marginally improve the detection rate of the MDA importance in all conditions (Supplementary Table S3).

Unfortunately, even after LD subsetting RF frequently under-performs the Armitage trend test (ATT). The difference is generally modest, but in the majority of cases confidence intervals for the detection rate of RF do not overlap with confidence intervals for the ATT (Supplementary Table S3). The primary exception is the MDA importance after LD subsetting, which achieves a detection rate within sampling variation of the trend in approximately half of the LD and MAF conditions considered here. This is comparable to the performance of the uncorrected GBM importance, but still markedly less favorable than the strong performance of GBM with LD subsetting.

The result is not entirely disheartening, however, since the simulation design specifically considers an ideal scenario for the ATT. Specifically, we assume an linear additive genetic model, which is consistent with the ATT and does not emphasize the potential benefit of RF's flexibility to simultaneously consider epistatic and dominance effects. In addition, we do not consider the potential benefit of using pseudocovariates (PCVs) with RF. Given the strong negative impact of MAF on the Gini importance after LD subsetting, including a pseudocovariate correction for MAF could dramatically benefit the sensitivity of RF with the Gini importance.

Role of Regression Coefficients and Effect Size

In studying the impact of MAF on variable importance for functional SNPs, the simulation results suggest that for the GBM and RF MDA importances high MAF is associated with lower observed variable importance for a constant effect size (Fig. 3a, Supplementary Fig. S7a). This effect is in the opposite direction of the observed trend for null SNPs, where high MAF is associated with higher importance for the GBM importance (Fig. 2a), and has no significant impact on the MDA importance (Supplementary Fig. S4a). This apparent reversal of the effect of MAF is due to our research design rather than a changing effect of MAF.

In the simulations with functional SNPs we generate functional SNPs that explain 1% and 2% of the variance in a continuous outcome. To generate these effect sizes at a range of MAF values, it becomes necessary to account for the variance of each SNP. Using the standard linear model, we use the 6 simulated effect SNPs to generate

$$y_i = \sum_j \beta_j x_{ij} + \epsilon_i \tag{1}$$

where the ϵ_i are distributed i.i.d as $N(0, \sigma^2)$, and we sum over the effect SNPs $j \in \{1, 5, 13, 17, 25, 26\}$ (see Supplementary Figure S5). Since the functional SNPs are independent, it follows that

$$var(y) = \sum_j var(\beta_j x_j) + var(\epsilon_{ij}) \tag{2}$$

$$= \sum_j \beta_j^2 var(x_j) + \sigma^2 \tag{3}$$

It is evident, then, that the proportion of variance explained for a given SNP is proportional to $\beta_j^2 var(x_j)$. Note that the SNPs are each binomial variables with MAF m_j, so $var(x_j) = 2m_j(1 - m_j)$. Thus to keep the effect size of the functional SNPs constant, in terms of variance explained, SNPs with higher MAF (larger variance) must be given smaller unstandardized regression coefficients β_j (Supplementary Table S6).

While this ensures the effect size is kept equal across MAF conditions, it results in a smaller regression coefficient for the SNPs with larger MAF. If, however, the variable importances reflect the magnitude of the unstandardized regression coefficient rather than the standardized effect size, then this design could be expected to result in lower observed importances for the SNPs with high MAF as a result of the chosen regression coefficients. In that situation the lower observed importance for SNPs with high MAF could be considered an expected result, rather than an unexpected "effect" of MAF.

If the simulation is repeated with equal regression coefficients rather than equal effect sizes, then the expected effect of MAF returns, with higher MAF associated with higher variable importance (Supplementary Fig. S11). This is consistent with previous work by Boulesteix *et al.* (2011), who also defined effect size in terms of regression coefficients rather than variance explained. In this framework, the inclusion of pseudocovariates reduces the impact of MAF on the GBM importance, though a strong effect of MAF remains (Supplementary Table S7). The impact of LD on the results is unchanged by using constant regression coefficients rather than variance explained. Including pseudocovariates provides a slight increase in sensitivity for SNPs with low MAF, but that improvement is still offset by a slight decrease in sensitivity for more common SNPs and detection rates for LD subsetting without pseudocovariates remain closer to the ATT (Supplementary Table S8).

In short, the relationship between MAF and variable importance for functional SNPs is fully intertwined with how the effect size for the functional SNPs is defined. The decision to use pseudocovariates with the GBM or RF MDA importance will likely be guided by the researchers desire to balance sensitivity to SNPs with high and low MAF.

Results for Tag SNPs and Larger SNP Effects

In addition to the functional SNPs each explaining 1% of the variance in a continuous outcome variable, our simulations to assess the sensitivity of each method also included functional SNPs explaining 2% of the variance in the outcome, and tag SNPs within the LD block for each functional SNP (Supplementary Fig. S5). We briefly investigate the results for those SNPs here.

Tag SNPs

In practice, there is no assurance that the functional SNP associated with a given phenotype will be included for genotyping. Instead, the available genotype data may only include tag SNPs in imperfect LD with the true functional SNP. This is a challenge for RF and GBM since competition between correlated tag SNPs will prevent any single SNP from achieving high variable importance.

The design of the simulation with functional SNPs includes a large number of tag SNPs (Supplementary Fig. S5). Specifically, in the set for each MAF $x_2, \ldots, x_4$ act as tag SNPs for x_1, the SNPs $x_6, \ldots, x_8$ tag x_5, SNPs $x_{14}, \ldots, x_{16}$ tag x_{13}, and $x_{18}, \ldots, x_{20}$ tag x_{17}. The remaining SNPs with no association with the phenotype, $x_9, \ldots, x_{12}; x_{21}, \ldots, x_{24}; x_{27}$, and x_{28} provide a reference for the observed variable importance of null SNPs.

Supplementary Figure S8a plots the median observed GBM importance for all 28 simulated SNPs with each MAF. Comparing the observed importance for the tag SNPs to their respective functional SNPs clearly shows that high importance is given to the functional SNP, with little importance for the tag SNPs. Completing the analysis with LD subsetting, however, allows each of the tag SNPs to be considered independently. As a result, the GBM importance with LD subsetting leads to a much smaller difference in the variable importance between functional SNPs and corresponding tag SNPs in strong LD (ex. $x_1, \ldots, x_8$; Supplementary Fig. S8b). A similar effect of LD subsetting is observed for the MDA importance (not shown).

Importantly, this latter pattern closely resembles the results provided by the ATT (Supplementary Fig. S9). Since we apply the ATT separately to each SNP, tag SNPs in strong LD may be recognized as nearly as significant as the corresponding functional SNP.

This similarity with the results of the ATT is not as clear for all of the tested SL methods. Specifically, the Gini importance provides much poorer separation between the importances of functional SNPs, tag SNPs, and null SNPs with no association with the phenotype (Supplementary Fig. S12a). While the high importance of tag SNPs can be beneficial, the reduced separation from the null SNPs increases the difficulty of identifying SNPs that are truly associated with the phenotype. The use of LD subsetting further reduces this separation (Supplementary Fig. S12b). Combined with the substantial effect of MAF on the importances, this accounts for the poor detection rate for the Gini importance (Supplementary Table S3). The visible size of the effect of MAF compared to the limited difference between effect SNPs and null SNPs highlights the potential value of a correction for the effect of MAF.

SNPs with Large Effect Sizes

The functional SNPs explaining 2% of the variance in the outcome variable correspond to an effect size much larger than is generally found in genome-wide studies. Still, comparing the influence of LD and MAF at this larger effect size to the results observed for the SNPs explaining 1% of the variance provides information about the role of effect size on the influence of LD and MAF. Thus the impact of effect size in these conditions may shape our future expectations as we push towards finding SNPs with smaller effects sizes.

The impact of LD and MAF on the variance importances for the SNPs with large effects is consistent with the results at the smaller effect size (Supplementary Tables S9 and S10). The only noticeable difference in the results is in the magnitude of the effects of MAF and LD. For example, we observe a much stronger effect of LD on the RF Gini and GBM importances prior to LD subsetting. There is also a stronger effect of MAF on the MDA importance and on the GBM Importance after the inclusion of pseudocovariates. The latter result may be a byproduct of the sensitivity of these methods to the size of the unstandardized regression coefficient for functional SNPs rather than the variance explained.

Unsurprisingly, higher detection rates were observed for all methods with at the larger effect size for the functional SNPs (Supplementary Table S11). As with the smaller effect SNPs, the SL methods frequently maintain detection rates within sampling variation of the observed rates for the ATT. Given that

the detection rate for most of these methods approach 1.0 due to the large effect size, there is less variability in the results to notice any major trends. Still, the results the SL methods are consistent with the results from the smaller effect size; GBM with LD subsetting continues to provide strong results, and the Gini importance, especially after LD subsetting, often performs poorly.

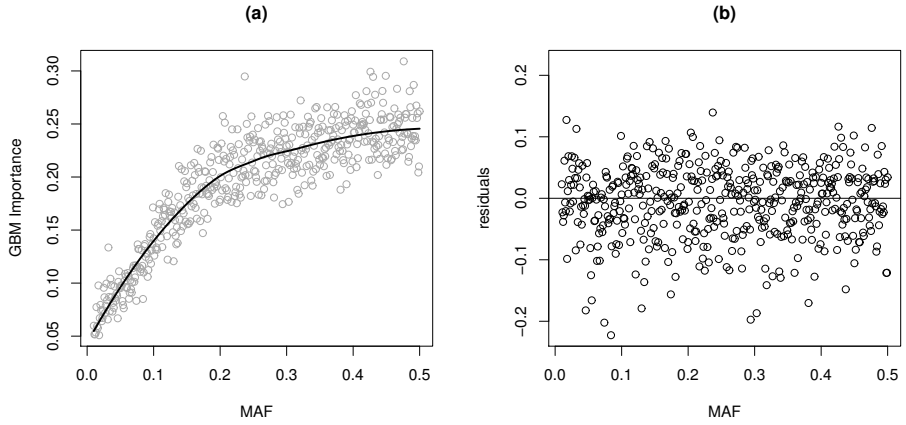
References

Biswas, A. and Hwang, J. (2002). A new bivariate binomial distribution. *Statistics and Probability Letters*, **60**, 231–240.

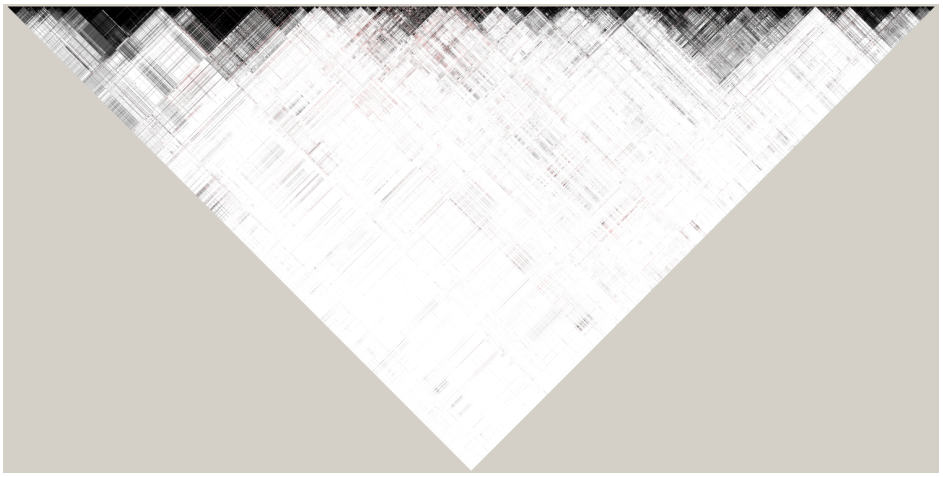
Boulesteix, A. *et al.* (2011). Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. *Brief. Bioinformatics*.

Clopper, C. and Pearson, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**(4), 404–413.

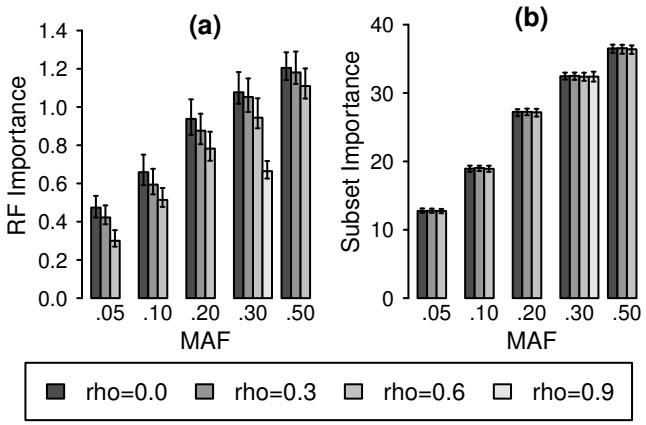
Nicodemus, K. K. and Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, **25**(15), 1884–1890.



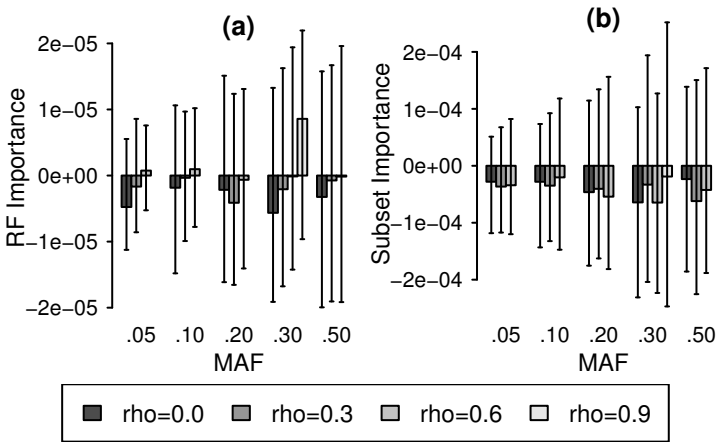
Supplementary Fig. S1: Pseudocovariate (PCV) correction to GBM importance. A loess curve is fit to the importances for the pseudocovariates aggregated across subsets (a). The estimated importance due to MAF is then subtracted, and the importance is divided by the standard deviation of the pseudocovariates, yielding the corrected importance (b).



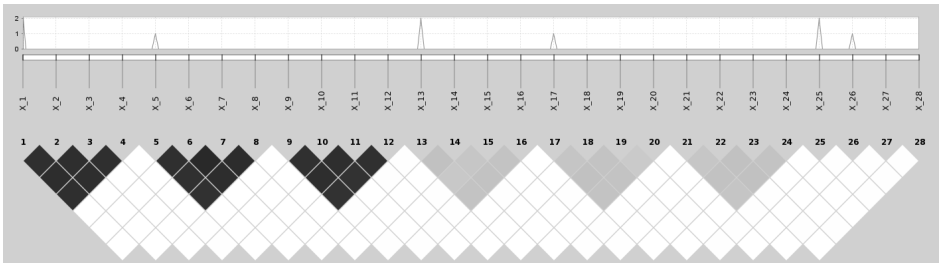
Supplementary Fig. S2: Map of LD blocks in 3000 SNP region on Chromosome 14 used as noise in the simulations. White indicates low D', black indicates high D' with high LOD, and pink indicates high D' with low LOD.



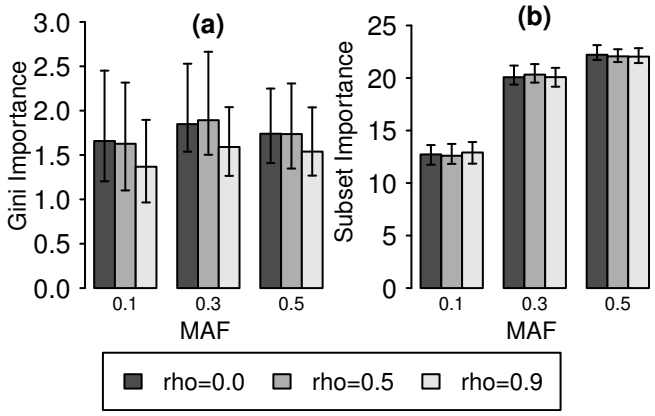
Supplementary Fig. S3: Median observed RF Gini variable importance by LD and MAF. Observed median importance with (a) no correction, and (b) LD subsetting. The skewed distribution of importances makes SEs uninformative, so error bars instead indicate observed upper and lower quartiles. Results show the reduced effect of LD and the remaining effect of MAF after LD subsetting.



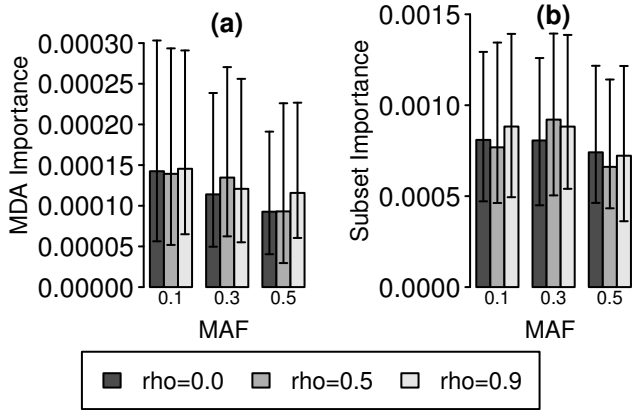
Supplementary Fig. S4: Median observed RF MDA variable importance by LD and MAF. Observed median importance with (a) no correction, and (b) LD subsetting. The skewed distribution of importances makes SEs uninformative, so error bars instead indicate observed upper and lower quartiles.



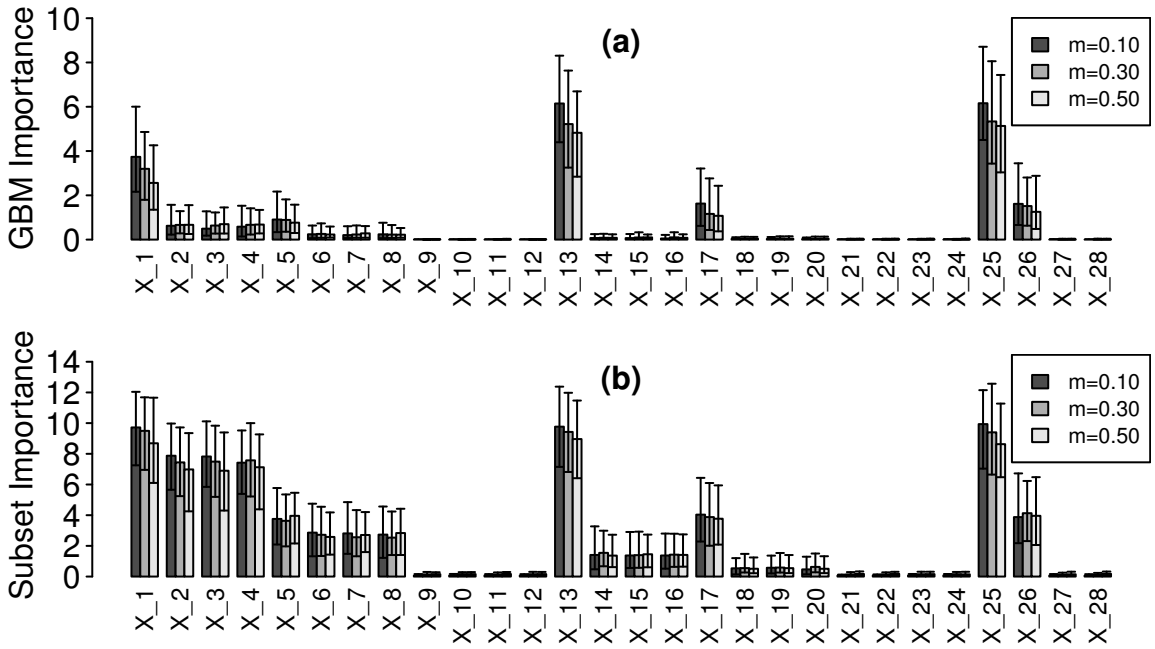
Supplementary Fig. S5: LD blocks and effect sizes for one set of 28 simulated SNPs. The effect size of each SNP, in terms of the percentage of variance explained for the continuous latent factor, is shown in the upper plot. The lower plot shows the pairwise population correlations between the SNPs. Black indicates $\rho = .9$, gray indicates $\rho = .5$, and white indicates $\rho = 0$. Sets with this design were simulated for each MAF $m = 0.1, 0.3, 0.5$.



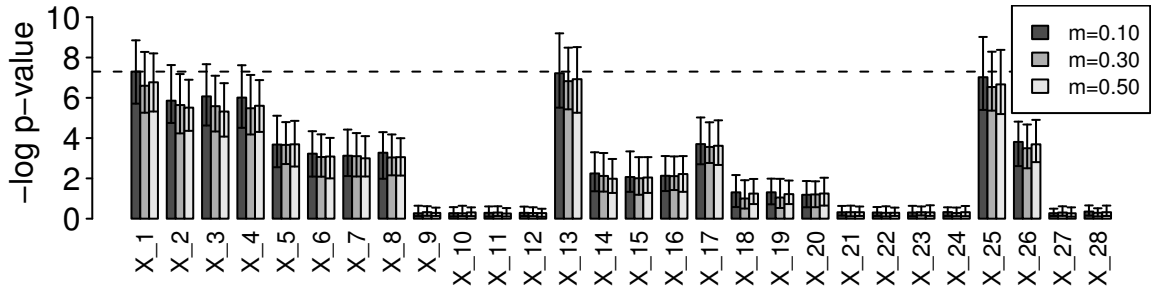
Supplementary Fig. S6: RF Gini importance for functional SNPs. Comparison of median observed RF Gini importance using (a) no correction, and (b) LD subsetting. The skewed distribution of importances makes SEs uninformative, so error bars instead indicate observed upper and lower quartiles.



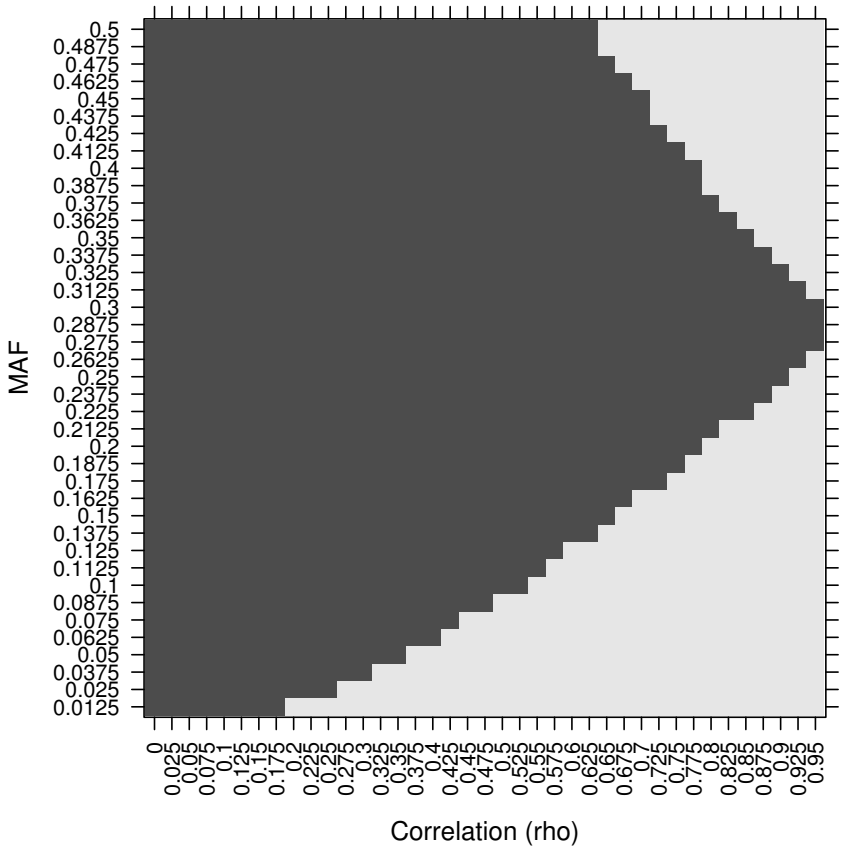
Supplementary Fig. S7: RF MDA importance for functional SNPs. Comparison of median observed RF MDA importance using (a) no correction, and (b) LD subsetting. The skewed distribution of importances makes SEs uninformative, so error bars instead indicate observed upper and lower quartiles.



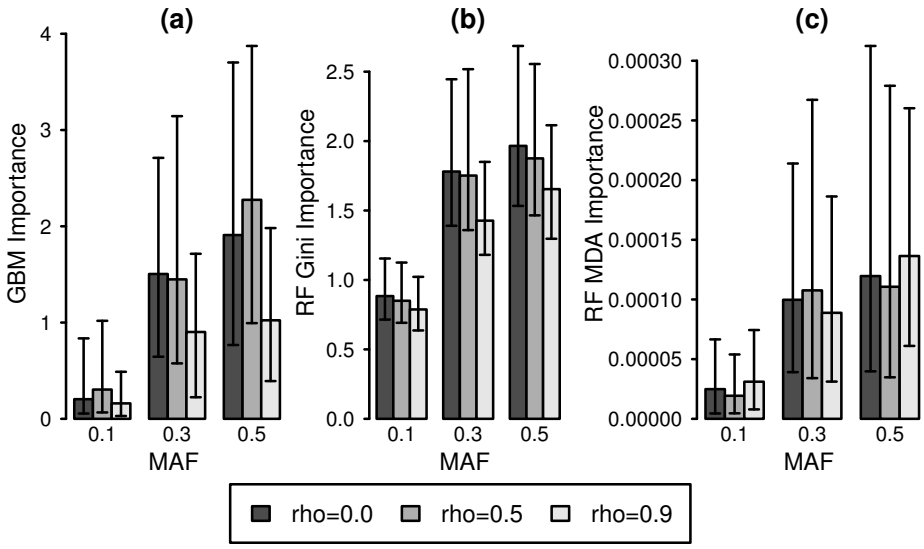
Supplementary Fig. S8: GBM variable importance for simulated SNPs including functional SNPs. Median GBM variable importance using (a) uncorrected importance and (b) LD subsetting. The skewed distribution of importances makes SEs uninformative, so error bars instead indicate observed upper and lower quartiles.



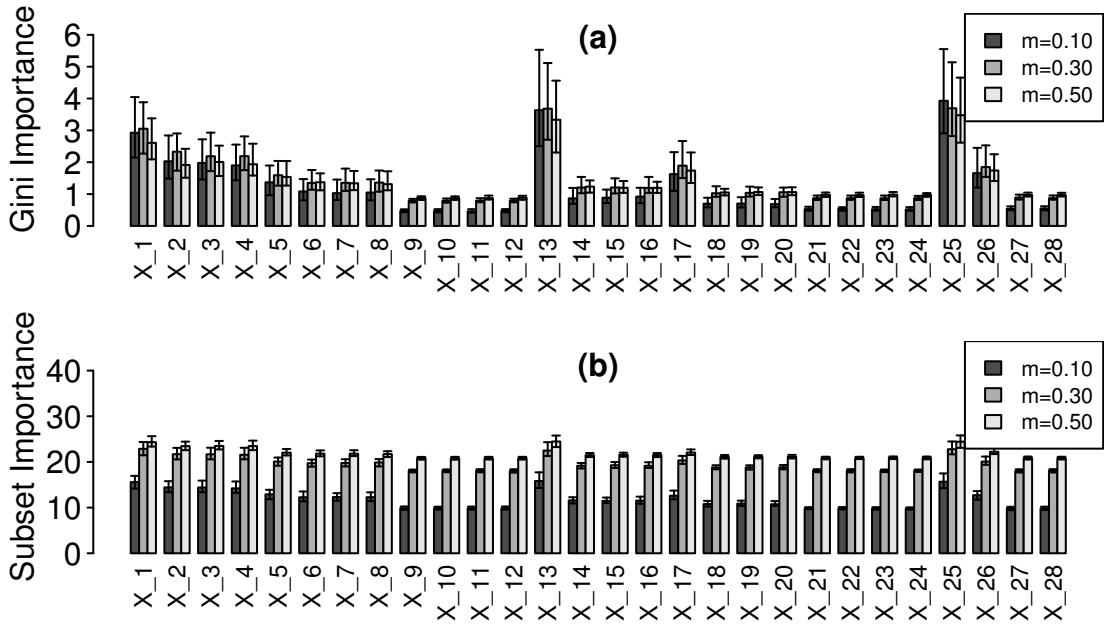
Supplementary Fig. S9: ATT results for simulated SNPs including functional SNPs. Median log p-values for the ATT for each SNP, with a horizontal reference line indicating the $p = 5 \times 10^{-8}$ threshold for genome-wide significance. To maintain consistency with figures for variable importance measures, error bars indicate observed upper and lower quartiles. Comparison to results for GBM with LD subsetting shows similar separation between functional SNPs, tag SNPs, and null SNPs.



Supplementary Fig. S10: Population limits on correlation with the empirical SNP. Plot depicts whether a given population correlation ρ can be achieved between a SNP with a given MAF and a SNP with MAF=.286, corresponding to the empirical SNP. Dark gray indicate valid combinations of MAF and ρ, light gray cells indicate invalid combinations.



Supplementary Fig. S11: Median variable importance for functional SNPs with unstandardized regression coefficients $\beta = .1414$. Results are shown by LD and MAF for (a) GBM importance, (b) RF Gini importance, and (c) RF MDA importance. The skewed distribution of importances makes SEs uninformative, so error bars instead indicate observed upper and lower quartiles.



Supplementary Fig. S12: RF Gini importance for functional SNPs. Comparison of median RF Gini importance using (a) no correction, and (b) LD subsetting. The skewed distribution of importances makes SEs uninformative, so error bars instead indicate observed upper and lower quartiles. Results show the dominant effect of MAF after LD subsetting.

Supplementary Table S1: Kruskal-Wallis test on RF variable importances with LD Subsetting

Effect	Condition	*df*	RF Gini+Subsets		RF MDA+Subsets	
			χ^2	p	χ^2	p
MAF	$\rho = 0$	4	1147.6	$<1\times10^{-10}$	5.3	0.26
	$\rho = 0.3$	4	1194.0	$<1\times10^{-10}$	3.0	0.56
	$\rho = 0.6$	4	1190.3	$<1\times10^{-10}$	3.5	0.48
LD	$m = 0.05$	2	1.3	0.52	0.3	0.85
	$m = 0.10$	2	1.8	0.42	0.3	0.85
	$m = 0.20$	2	1.2	0.55	<0.1	0.99
	$m = 0.30$	3	1.5	0.68	3.2	0.37
	$m = 0.50$	2	2.3	0.32	0.4	0.81

Significance test results are given for the simple effect of MAF on variable importance in RF holding LD constant at the given level ρ, and for the simple effect of LD at the given level of MAF m. With family-wise Bonferroni corrections, p-values for the effect of MAF less than .0167, and p-values for the effect of LD less than .01 correspond to significance at the $\alpha = .05$ level.

Supplementary Table S2: Computational Burden for Analysis of Chromosome 22

Task	RF		GBM		ATT	
	Time	Memory	Time	Memory	Time	Memory
No Subsetting	19:33:20	3.56 GB	10:45:07	4.25 GB	00:00:07	.10 GB
Create Subsets	00:29:52	1.81 GB	00:29:52	1.81 GB	—	—
Analyze a Subset	00:43:55	.91-1.30 GB	00:06:40	.36-.44 GB	—	—
Aggregate Results	00:00:07	.37 GB	00:00:03	.32 GB	—	—
Total with Subsets	04:53:29	—	01:09:55	—	—	—

Comparison of the computational requirements to analyze Chromosome 22 (30,218 SNPs) from the empirical data ($n = 2,235$) using RF, GBM, and the ATT with and without LD subsetting. RF and GBM are implemented in R; the ATT is implemented in PLINK. Times are reported in hh:mm:ss format. The requirements to analyze a LD subset vary by subset, so the observed average time and range of RAM usage is reported. To total burden with LD subsets is computed based on the availability of 50 cores for parallel processing of $k = 300$ subsets.

Supplementary Table S3: Detection Rate for RF by Method, LD, and MAF

MAF	Method	Detection Rate (95% Confidence Interval)		
		$\rho = .9$	$\rho = .5$	$\rho = 0$
0.1	RF Gini	.58 (.52,.65)	.68 (.62,.74)	.72 (.66,.77)
	Gini Sub	.00 (.00,.01)	.00 (.00,.01)	.00 (.00,.01)
	RF MDA	.78 (.73,.83)	.74 (.68,.79)	.75 (.69,.80)
	MDA Sub	**.84 (.79,.88)**	**.86 (.81,.90)**	.85 (.80,.89)
	ATT	**.90 (.86,.94)**	**.92 (.88,.95)**	**.94 (.90,.97)**
0.3	RF Gini	.76 (.70,.81)	**.85 (.80,.89)**	**.89 (.85,.93)**
	Gini Sub	.19 (.14,.24)	.25 (.20,.31)	.20 (.16,.26)
	RF MDA	.74 (.69,.80)	.77 (.71,.82)	.72 (.66,.77)
	MDA Sub	.83 (.78,.87)	**.86 (.82,.90)**	**.84 (.78,.88)**
	ATT	**.92 (.88,.95)**	**.92 (.88,.95)**	**.90 (.85,.93)**
0.5	RF Gini	.78 (.72,.83)	**.86 (.81,.90)**	.84 (.79,.89)
	Gini Sub	.78 (.73,.83)	.81 (.76,.86)	.84 (.79,.88)
	RF MDA	.76 (.70,.81)	.65 (.59,.71)	.66 (.59,.71)
	MDA Sub	.76 (.70,.81)	.78 (.73,.83)	.81 (.76,.86)
	ATT	**.90 (.86,.94)**	**.93 (.89,.96)**	**.94 (.90,.96)**

"Detection" in each replication is defined as importance (or test statistic) for a functional SNP greater than the highest observed importance among simulated SNPs unassociated with the phenotype. Results are given for the uncorrected RF importances (Gini and MDA) and for RF importances with LD subsetting (Gini Sub and MDA Sub). Proportions are out of 250 replications, with exact confidence intervals constructed following Clopper and Pearson (1934). Values in **bold** have confidence intervals that overlap the confidence interval for the detection rate of the ATT.

Supplementary Table S4: Friedman test on RF variable importance for functional SNPs with LD Subsetting

Effect	Condition	RF Gini+Subsets		RF MDA+Subsets	
		χ^2	p	χ^2	p
MAF	$\rho = 0.9$	453.4	$<1\times10^{-10}$	8.9	3.1×10^{-2}
	$\rho = 0.5$	435.6	$<1\times10^{-10}$	7.3	2.6×10^{-2}
	$\rho = 0.0$	440.5	$<1\times10^{-10}$	0.7	6.9×10^{-1}
LD	$m = 0.1$	2.3	3.2×10^{-1}	1.6	4.5×10^{-1}
	$m = 0.3$	12.2	2.3×10^{-3}	5.5	6.5×10^{-2}
	$m = 0.5$	5.1	7.6×10^{-2}	1.1	5.9×10^{-1}

Significance test results are given for the simple effect of MAF on variable importances in RF at each level of LD ρ, and for the simple effect of LD on variable importances at MAF m. With family-wise Bonferroni corrections, p-values less than .017 correspond to significance at the $\alpha = .05$ level. All tests have $df = 2$.

Supplementary Table S5: Correlation between SNP importance and number of subsets

MAF	LD	GBM		RF Gini		RF MDA	
		τ	p	τ	p	τ	p
$m = .05$	$\rho = 0$	.046	.289	-.004	.925	.077	.228
	$\rho = 0.3$	-.021	.632	.071	.097	.031	.625
	$\rho = 0.6$	-.002	.964	-.047	.288	.071	.261
$m = .10$	$\rho = 0$	-.035	.419	.068	.118	-.022	.728
	$\rho = 0.3$	-.094	.029	-.035	.415	-.085	.178
	$\rho = 0.6$	-.039	.375	-.018	.674	.054	.397
$m = .20$	$\rho = 0$	.060	.172	-.059	.179	-.048	.457
	$\rho = 0.3$	.046	.286	-.034	.431	.028	.662
	$\rho = 0.6$	-.020	.647	.034	.434	-.030	.642
$m = .30$	$\rho = 0$	.020	.657	-.051	.257	-.074	.261
	$\rho = 0.3$	.062	.146	.052	.228	.118	.061
	$\rho = 0.6$	-.008	.849	.025	.561	.042	.508
	$\rho = 0.9$	.046	.320	.021	.650	-.095	.133
$m = .50$	$\rho = 0$	-.054	.240	.033	.475	-.059	.381
	$\rho = 0.3$	-.021	.633	.062	.156	-.113	.079
	$\rho = 0.6$	.007	.866	.003	.940	.044	.492

Observed correlation between SNP importance with LD subsetting and the number of LD subsets containing the given SNP for each LD and MAF condition and each SL method. Correlations are reported using Kendall's τ to account for the skewed distribution of importances. With family-wise Bonferroni correction for 16 conditions, p-values less that .003 indicate significance at the $\alpha = .05$ level for each SL method.

Supplementary Table S6: Relationship between regression coefficient and effect size for simulated SNPs

	Large Effect		Small Effect	
MAF	β	Variance Explained	β	Variance Explained
0.5	0.2000	0.02	0.1414	0.01
0.3	0.2182	0.02	0.1543	0.01
0.1	0.3333	0.02	0.2357	0.01

The table gives the unstandardized coefficient (β) and the corresponding proportion of variance explained for the simulated functional SNPs at each MAF. The phenotype y_{bin} is generated by dichotomizing the underlying normal outcome variable $y_i = \sum \beta_j x_{ij} + \epsilon_i$ where the x_i are the SNPs coded 0,1,2 and the ϵ_i are i.i.d. $N(0, \sigma^2)$. To ensure y has unit variance, we use $\sigma^2 = .73$.

Supplementary Table S7: Friedman test on GBM importance for functional SNPs with equal coefficients

Effect	Condition	GBM		GBM+Subsetting		GBM+PCVs	
		χ^2	p	χ^2	p	χ^2	p
MAF	$\rho = 0.9$	127.7	$<1\times10^{-10}$	178.1	$<1\times10^{-10}$	105.7	$<1\times10^{-10}$
	$\rho = 0.5$	143.6	$<1\times10^{-10}$	131.4	$<1\times10^{-10}$	101.5	$<1\times10^{-10}$
	$\rho = 0.0$	137.9	$<1\times10^{-10}$	180.9	$<1\times10^{-10}$	117.2	$<1\times10^{-10}$
LD	$m = 0.1$	10.3	5.8×10^{-3}	0.6	7.4×10^{-1}	0.7	6.9×10^{-1}
	$m = 0.3$	36.5	1.2×10^{-8}	3.0	2.2×10^{-1}	2.1	3.5×10^{-1}
	$m = 0.5$	54.7	$<1\times10^{-10}$	1.2	5.5×10^{-1}	1.2	5.5×10^{-1}

Significance test results for the effects of MAF and LD on the variable importance of functional SNPs with unstandardized regression coefficients $\beta = .1414$. Results are given for the simple effect of MAF at each LD ρ, and for the simple effect of LD at each MAF m. With family-wise Bonferroni corrections, p-values less than 1.7×10^{-2} correspond to significance at the $\alpha = .05$ level. All tests have $df = 2$.

Supplementary Table S8: Detection Rate for GBM for Functional SNPs with Equal Coefficients

MAF	Method	Detection Rate (95% Confidence Interval)		
		$\rho = .9$	$\rho = .5$	$\rho = 0$
0.1	GBM	.31 (.25,.37)	.44 (.38,.50)	.39 (.33,.45)
	GBM Subsets	.43 (.37,.49)	**.44 (.38,.51)**	.42 (.36,.49)
	GBM PCVs	.45 (.39,.52)	**.47 (.41,.54)**	.45 (.39,.51)
	ATT	**.60 (.54,.66)**	**.57 (.50,.63)**	**.59 (.53,.65)**
0.3	GBM	.64 (.58,.70)	**.78 (.72,.83)**	**.80 (.74,.85)**
	GBM Subsets	**.80 (.74,.85)**	**.78 (.72,.83)**	**.81 (.75,.85)**
	GBM PCVs	**.76 (.70,.81)**	**.83 (.78,.88)**	**.81 (.76,.86)**
	ATT	**.86 (.81,.90)**	**.85 (.80,.89)**	**.88 (.83,.92)**
0.5	GBM	.74 (.68,.79)	**.86 (.82,.90)**	.81 (.76,.86)
	GBM Subsets	**.87 (.82,.91)**	**.86 (.82,.90)**	**.88 (.84,.92)**
	GBM PCVs	**.84 (.79,.88)**	.82 (.77,.87)	.82 (.76,.86)
	ATT	**.91 (.87,.94)**	**.92 (.87,.95**)	**.94 (.90,.96)**

"Detection" in each replication is defined as importance (or test statistic) for a functional SNP greater than the highest observed importance among simulated SNPs unassociated with the phenotype. Results are given for the GBM, GBM with LD subsetting, GBM with subsetting and pseudocovariates (PCVs), and the ATT. Proportions are out of 250 replications, with exact confidence intervals constructed following Clopper and Pearson (1934). Values in **bold** have confidence intervals that overlap the confidence interval for the detection rate of the ATT.

Supplementary Table S9: Friedman test for effect of LD and MAF on GBM for large functional SNPs

Effect	Condition	GBM		GBM+Subsetting		GBM+PCVs	
		χ^2	p	χ^2	p	χ^2	p
MAF	$\rho = .9$	35.9	1.6×10^{-8}	4.2	1.2×10^{-1}	48.0	$<1\times10^{-10}$
	$\rho = .5$	27.9	8.9×10^{-7}	6.7	3.5×10^{-2}	94.1	$<1\times10^{-10}$
	$\rho = 0$	10.4	5.6×10^{-3}	8.2	1.7×10^{-2}	95.0	$<1\times10^{-10}$
LD	$m = .1$	72.6	$<1\times10^{-10}$	0.2	8.8×10^{-1}	1.9	3.9×10^{-1}
	$m = .3$	67.4	$<1\times10^{-10}$	0.2	8.9×10^{-1}	1.1	5.9×10^{-1}
	$m = .5$	79.9	$<1\times10^{-10}$	1.0	6.1×10^{-1}	0.5	8.0×10^{-1}

Significance test results are given for each GBM importances for the simple effect of MAF holding LD constant at a given level ρ, and for the simple effect of LD holding MAF constant at a given level m. With family-wise Bonferroni corrections, p-values less than 1.7×10^{-2} correspond to significance at the $\alpha = .05$ level. All tests have $df = 2$.

Supplementary Table S10: Friedman test for effect of LD and MAF on large functional SNPs with RF

Effect	Condition	RF Gini		Gini+Subsetting		RF MDA		MDA+Subsetting	
		χ^2	p	χ^2	p	χ^2	p	χ^2	p
MAF	$\rho = .9$	14.5	7.1×10^{-4}	385.7	$<1\times10^{-10}$	30.0	3.1×10^{-7}	6.9	3.1×10^{-2}
	$\rho = .5$	6.7	3.6×10^{-2}	372.0	$<1\times10^{-10}$	18.6	9.1×10^{-5}	4.1	1.3×10^{-1}
	$\rho = 0$	5.5	6.4×10^{-2}	381.4	$<1\times10^{-10}$	23.8	6.8×10^{-6}	11.9	2.6×10^{-3}
LD	$m = .1$	37.9	5.9×10^{-9}	0.2	8.8×10^{-1}	2.2	3.4×10^{-1}	0.3	8.4×10^{-1}
	$m = .3$	34.7	3.0×10^{-8}	1.4	5.0×10^{-1}	1.6	4.5×10^{-1}	1.6	4.5×10^{-1}
	$m = .5$	38.5	4.3×10^{-9}	0.5	8.0×10^{-1}	1.6	4.5×10^{-1}	0.9	6.4×10^{-1}

Significance test results are given for the simple effect of MAF on each importance measure holding LD constant at a given level ρ, and for the simple effect of LD holding MAF constant at a given level m. With family-wise Bonferroni corrections, p-values less than 1.7×10^{-2} correspond to significance at the $\alpha = .05$ level. All tests have $df = 2$.

Supplementary Table S11: Detection Rate for Large Functional SNPs by Method, LD, and MAF

MAF	Method	Detection Rate (95% Confidence Interval)		
		$\rho = .9$	$\rho = .5$	$\rho = 0$
0.1	GBM	**.968 (.938,.986)**	**.988 (.965,.998)**	**1.000 (.985,1.000)**
	GBM Sub	**1.000 (.985,1.000)**	**.992 (.971,.999)**	**.996 (.978,1.000)**
	GBM PCVs	**.992 (.971,.999)**	**.992 (.971,.999)**	**1.000 (.985,1.000)**
	RF Gini	.956 (.923,.978)	**.976 (.948,.991)**	**.980 (.954,.993)**
	Gini Sub	.020 (.007,.046)	.028 (.011,.057)	.036 (.017,.067)
	RF MDA	**.984 (.960,.996)**	**.976 (.948,.991)**	**.972 (.943,.989)**
	MDA Sub	**.992 (.971,.999)**	**.992 (.971,.999)**	**.988 (.965,.998)**
	ATT	**.996 (.978,1.000)**	**.996 (.978,1.000)**	**.992 (.971,.999)**
0.3	GBM	**.984 (.960,.996)**	**.988 (.965,.998)**	**.996 (.978,1.000)**
	GBM Sub	**.980 (.954,.993)**	**.996 (.978,1.000)**	**1.000 (.985,1.000)**
	GBM PCVs	**.992 (.971,.999)**	**.992 (.971,.999)**	**.992 (.971,.999)**
	RF Gini	**.988 (.965,.998)**	**1.000 (.985,1.000)**	**.992 (.971,.999)**
	Gini Sub	.760 (.702,.812)	.740 (.681,.793)	.808 (.754,.855)
	RF MDA	**.984 (.960,.996)**	**.972 (.943,.989)**	**.980 (.954,.993)**
	MDA Sub	**.996 (.978,1.000)**	**.988 (.965,.998)**	**.992 (.971,.999)**
	ATT	**.996 (.978,1.000)**	**.992 (.971,.999)**	**.996 (.978,1.000)**
0.5	GBM	**.968 (.938,.986)**	**.984 (.960,.996)**	**.996 (.978,1.000)**
	GBM Sub	**.996 (.978,1.000)**	**1.000 (.985,1.000)**	**.984 (.960,.996)**
	GBM PCVs	**.984 (.960,.996)**	**.980 (.954,.993)**	**.980 (.954,.993)**
	RF Gini	**.984 (.960,.996)**	**.992 (.971,.999)**	**.996 (.978,1.000)**
	Gini Sub	**.984 (.960,.996)**	**.984 (.960,.996)**	**.992 (.971,.999)**
	RF MDA	.956 (.923,.978)	**.940 (.903,.966)**	.956 (.923,.978)
	MDA Sub	**.992 (.971,.999)**	**.968 (.938,.986)**	**.988 (.965,.998)**
	ATT	**.996 (.978,1.000)**	**1.000 (.985,1.000)**	**.996 (.978,1.000)**

"Detection" in each replication is defined as importance (or test statistic) for a functional SNP greater than the highest observed importance among simulated SNPs unassociated with the phenotype. Results are given for the RF importances with and without LD subsetting, and for the GBM importance with no correction, LD subsetting only, and subsetting with pseudocovariates (PCVs). Proportions are out of 250 replications, with exact confidence intervals constructed following Clopper and Pearson (1934). Values in **bold** have confidence intervals that overlap the confidence interval for the detection rate of the ATT.