Supplementary Material for:

Identifying rare variants inconsistent with identity-by-descent in population-scale whole genome sequencing data

K.E. Johnson, C.J. Adams, & B.F. Voight

Supplementary Methods

Gibbs sampling algorithm for the Bayesian hierarchical model

We begin by randomly assigning starting values of each parameter. Then, in each iteration, we sample from the full conditionals in the following steps:

1. Sample π from $\pi|K$
2. Sample β from $\beta|T,\alpha,\tilde{\alpha},\tilde{\beta}$
3. For each variant i and each potential value of $k_i \in p$, p representing the possible partitions:
    a. Sample t_i from $t_i|D_i,k_i,J_i,\alpha,\beta$
    b. Sample j_m from $j_m|D_i,k_i,t_i,\alpha,\beta$
    c. Calculate $P(k_i=p|D_i,t_iJ_i,\alpha,\beta)$
4. Sample k_i from $P(k_i=p|\ldots)$; accept corresponding t_i, j_m
5. Go back to step 1.

Choice of gamma distribution to model the TMRCA in Gibbs sampling

We tested two potentially appropriate distributions for the TMRCA in our hierarchical model, the Lognormal and Gamma distributions. For the Gamma distribution we were able to utilize conjugate priors to derive the full conditional distribution for the "sample t" step; however, for the Lognormal distribution, the "sample t" step required the use of computationally intensive rejection sampling. We found that the Gamma distribution gave comparable results to the Lognormal when applied to simulations generated from the above theoretical distribution of TMRCAs (e.g., for allele count of 2, AUC=0.90 (95% CI 0.89-0.92) for Lognormal vs. AUC=0.89 (95% CI 0.88-0.90) for gamma), and so moved forward with the Gamma distribution. Similarly, for ease of computation, we tested a range of fixed shape parameters (α) for the gamma distribution and found comparable results classifying simulated data (e.g. for allele count of 2, AUC=0.86 (95% CI 0.84-0.88) for α=20; AUC=0.87 (95% CI 0.85-0.88) for α=40; AUC=0.89 (95% CI 0.88-0.90) for α=140).

Supplementary Table Descriptions

Supplementary Table 1. Posterior probabilities of a variant being non-IBD in simulated data, for all possible partitions of allele count 4 and allele count 5 recurrent variants. AC: allele count; Partition: partition of alleles; nIBDpair: number of IBD pairs for this partition; N: number of times this partition was observed in simulations; Mean: mean posterior probability; 95CI_L: lower limit of empirical 95% confidence interval for the posterior probability; 95CI_U: upper limit of empirical 95% confidence interval for the posterior probability.

Supplementary Table 2. Areas under curve (AUC) from applying the Bayesian hierarchical model to a range of simulation types. Mean: bootstrapped mean AUC; Median: bootstrapped median AUC; 95CI_L: lower limit of bootstrapped AUC 95% confidence interval; 95CI_U: upper limit of bootstrapped AUC 95% confidence interval; Recombination: type of recombination rate used in simulations (Uniform: $r= 1x10^{-8}$, Variable: each simulation sampled from a human recombination map); AC: allele count; Selection: was selection simulated (logical value); Sample size: number of diploids sampled.

Supplementary Table 3. Areas under curve (AUC) from application of *runtc* to simulated data to identify non-IBD variants. Mean: bootstrapped mean AUC; Median: bootstrapped median AUC; 95CI_L: lower limit of bootstrapped AUC 95% confidence interval; 95CI_U: upper limit of bootstrapped AUC 95% confidence interval; AC: allele count.

Supplementary Table 4. The correlation between expected and observed fractions of non-IBD variants, for a given allele count and subset of variants. r: Pearson's correlation coefficient; P.value: P-value of r; 95%CI_L: lower limit of 95% confidence interval of r; 95%CI_U: upper limit of 95% confidence interval of r; AC: allele count; Type: variants included (all = all variants, cgt = CpG>T variants only, ncgt = all variants except CpG>T variants).

Supplementary Table 5. Expected and observed fractions of variants called non-IBD for 5-mer sequence contexts. Context: sequence context; nVars: number of variants of this allele count and context included; n-nIBD: number of variants called non-IBD; obsFrac: fraction of

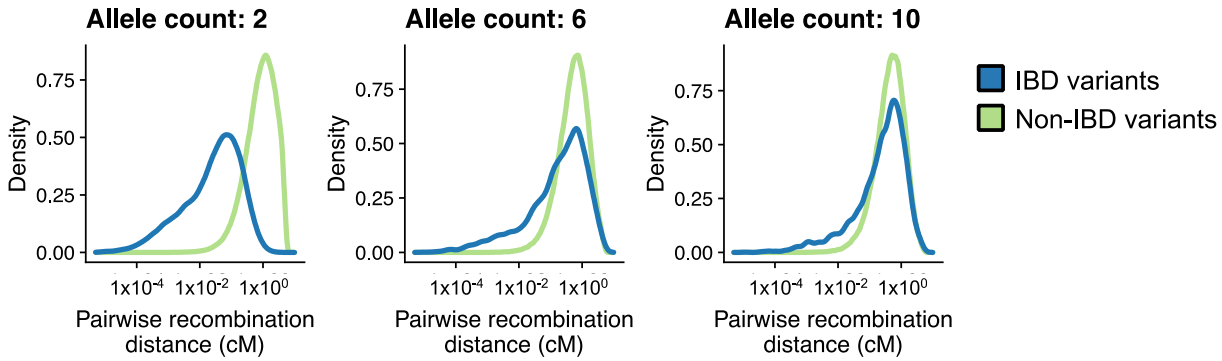
variants called non-IBD; polyProb: polymorphism probability of this sequence context; expFrac: expected fraction of non-IBD variants of this sequence context; AC: allele count.

Supplementary Table 6. Results of logistic regression predicting non-IBD variants from genomic annotations. Estimate: regression beta coefficient estimate; SD: standard deviation of beta estimate; P: P-value of beta estimate; AC: allele count; Input: variants included in model (all or CpG>T only); Annotation: genomic annotation; hsDist.z: hotspot distance z-score; rRate.z: local recombination rate z-score; B.z: McVicker's B statistic z-score; RT.z: replication timing z-score; GC.z: local GC content z-score; mOv.z: ovary CpG methylation z-score; mTes.z: testes CpG methylation z-score; DP.z: read depth z-score; VQSLOD.z: variant quality z-score; polyProb.z: 7-mer polymorphism probability.

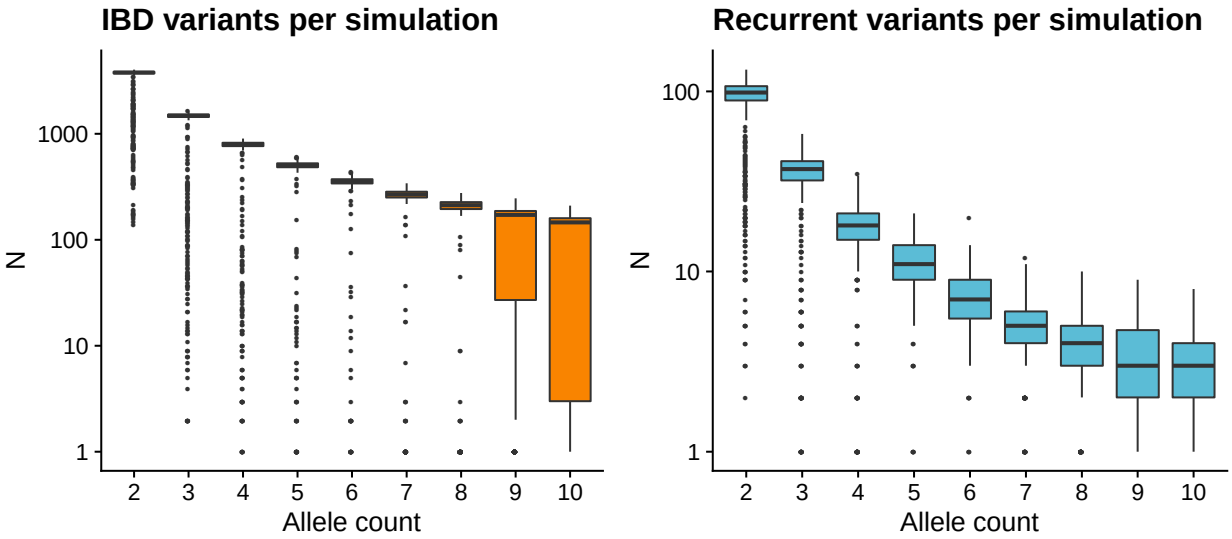
Supplementary Table 7. Results of logistic regression predicting non-IBD variants in putative gene conversion tracts vs. all other non-IBD variants, with genomic annotations as predictors. Estimate: regression beta coefficient estimate; SD: standard deviation of beta estimate; P: P-value of beta estimate; AC: allele count; Annotation: genomic annotation; hsDist.z: hotspot distance z-score; rRate.z: local recombination rate z-score; B.z: McVicker's B statistic z-score; RT.z: replication timing z-score; GC.z: local GC content z-score; mOv.z: ovary CpG methylation z-score; mTes.z: testes CpG methylation z-score; DP.z: read depth z-score; VQSLOD.z: variant quality z-score; polyProb.z: 7-mer polymorphism probability; posterior: posterior probability of being non-IBD.

Supplementary Table 8. Correlation coefficients between posterior probabilities of variants being non-IBD from Bayesian hierarchical model. Comparing the model run with non-IBD alpha=20 vs. non-IBD alpha=40. r: Pearson's correlation coefficient; P: correlation coefficient P-value; N: number of variants included; AC: allele count.

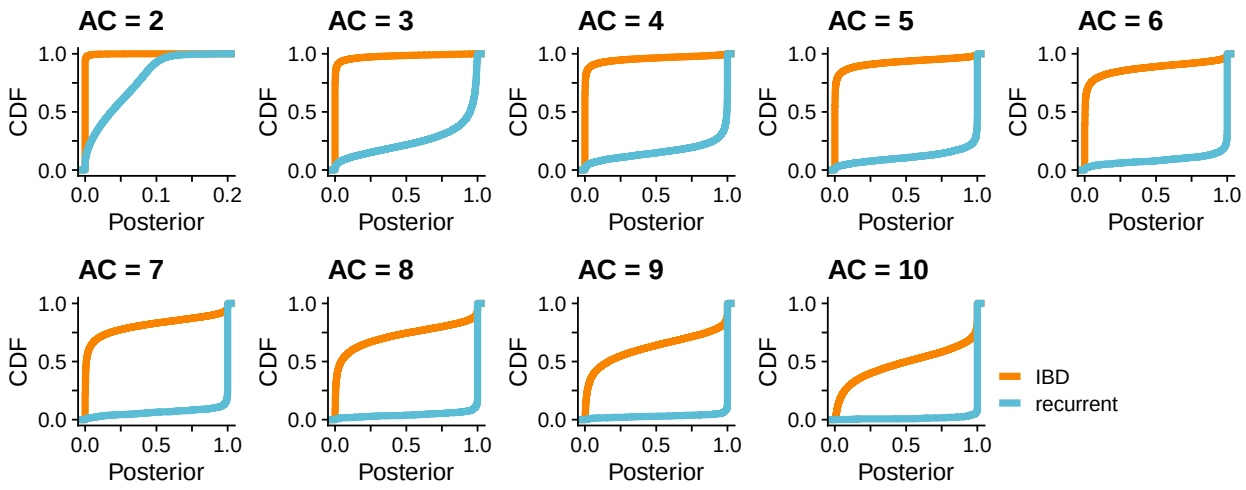
Supplementary Figures



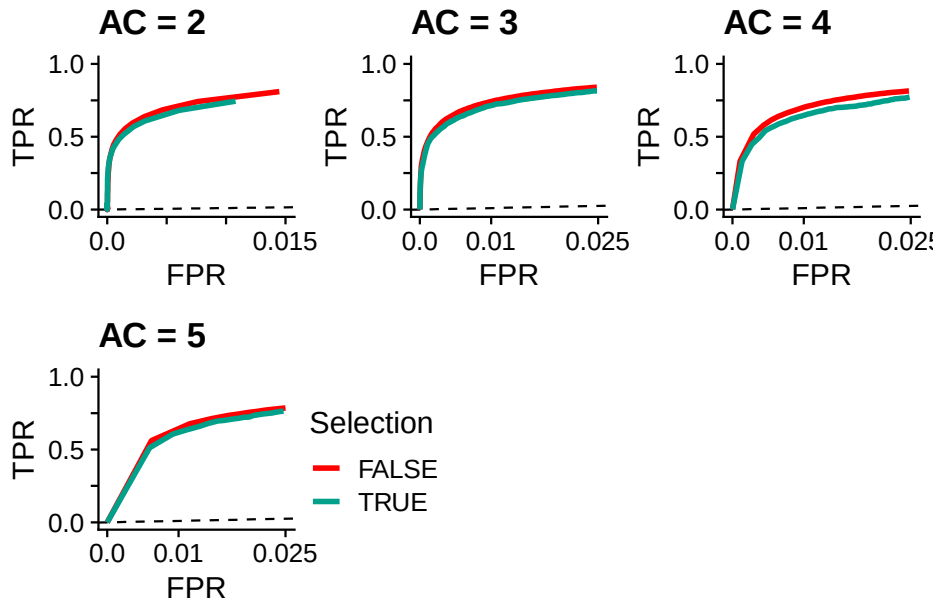
Supplementary Figure 1. Distributions of pairwise recombination distances from simulated IBD and non-IBD variants.
The density of distances to the nearest inferred recombination event for allele pairs from one million simulated variants with allele count 2, 6, or 10. Line color denotes variant status of IBD (blue) or recurrent (light green). Simulations used a uniform recombination rate, uniform mutation rate, no selection, sample size of 3,621, and demographic history as described in **Materials and Methods**.



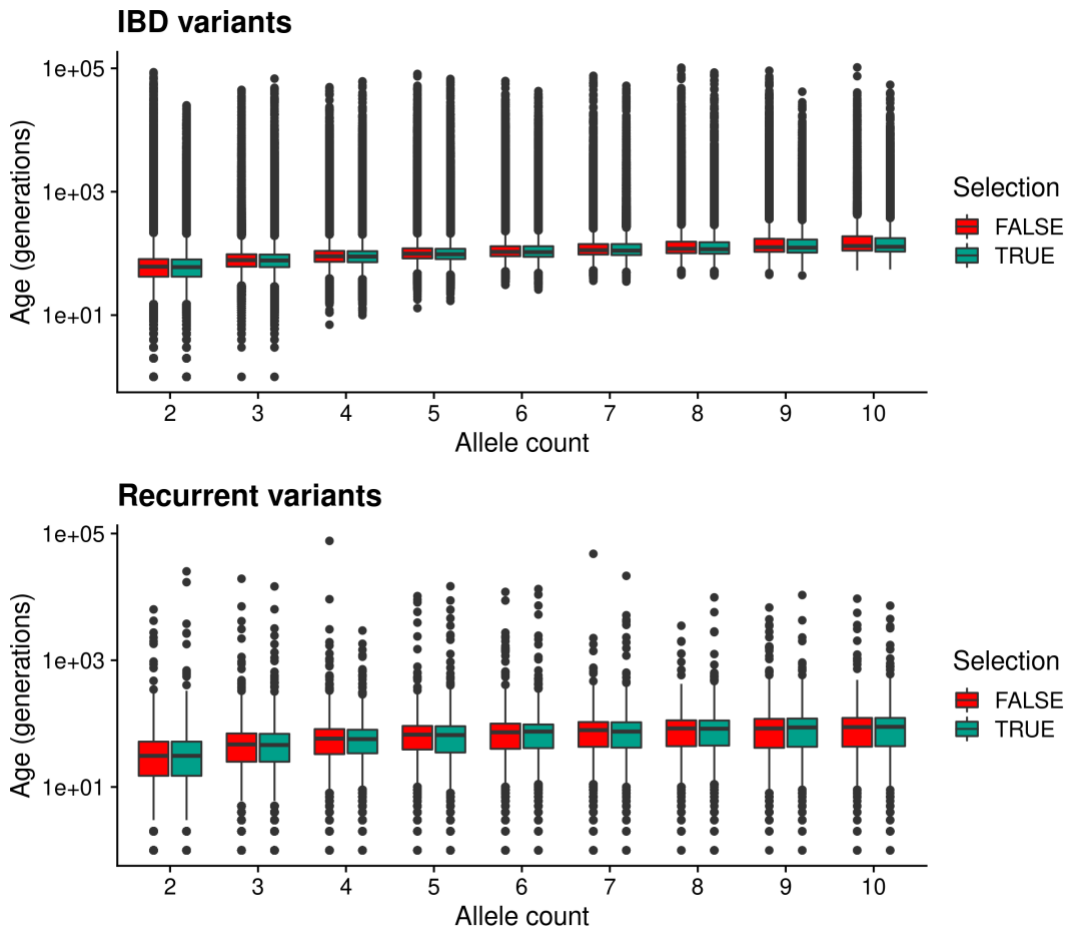
Supplementary Figure 2. Counts of IBD and recurrent variants in 1000 simulations. The counts of IBD and recurrent variants of allele count 2-10 in 1000 SLiM simulations with a uniform recombination rate, uniform mutation rate, no selection, sample size of 3,621, and demographic history as described in **Materials and Methods**.



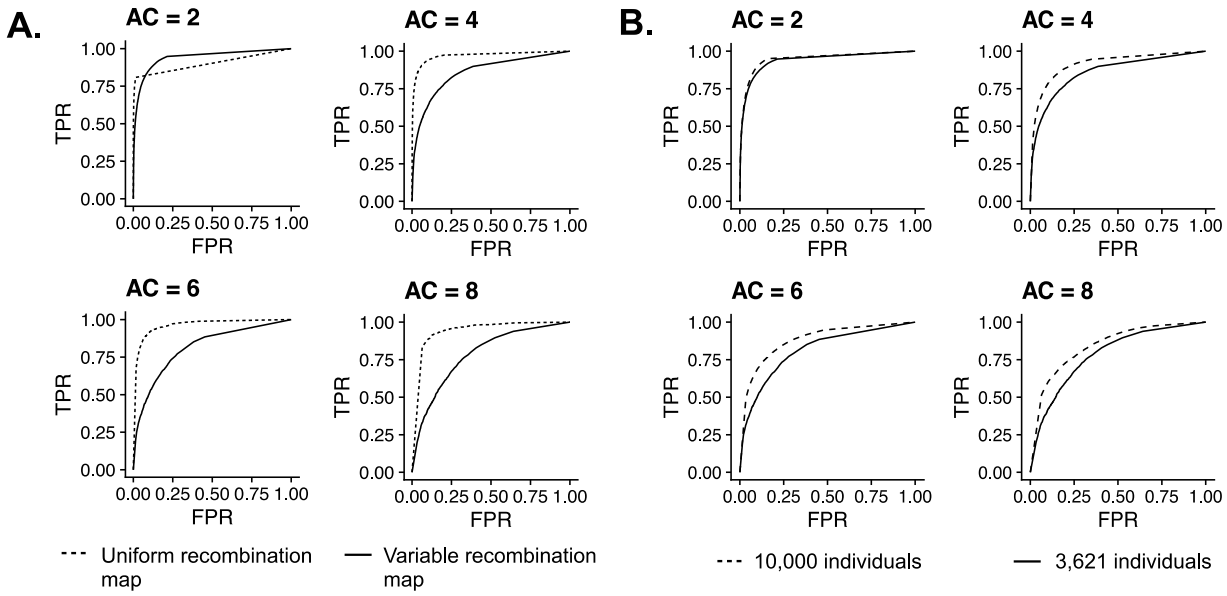
Supplementary Figure 3. Posterior probabilities of simulated IBD or recurrent variants being non-IBD.
The empirical cumulative density of posterior probabilities of a variant being non-IBD, for simulated IBD (orange) or recurrent (blue) mutations. Posterior probabilities were generated by application of our Bayesian hierarchical model to variants from simulations with a uniform recombination rate, uniform mutation rate, no selection, sample size of 3,621, and demographic history as described in **Materials and Methods**. Note the x-axis scale is not the same for AC=2 and the other allele counts.



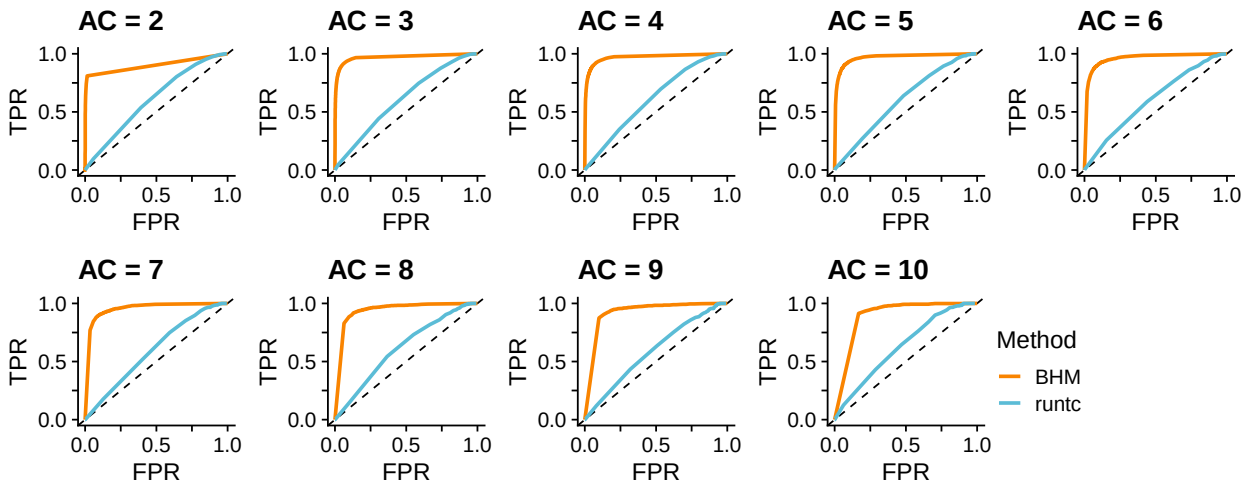
Supplementary Figure 4. ROC curves of the Bayesian hierarchical model's ability to distinguish IBD and non-IBD variants from simulations with or without selection.
ROC curves for the Bayesian hierarchical model applied to distinguish simulated recurrent and IBD mutations, with (TRUE) and without (FALSE) the presence of background selection. Each panel represents the application to variants of a given allele count (AC) 2-5. Note the x-axis scale is not the same for AC=2 and the other allele counts.



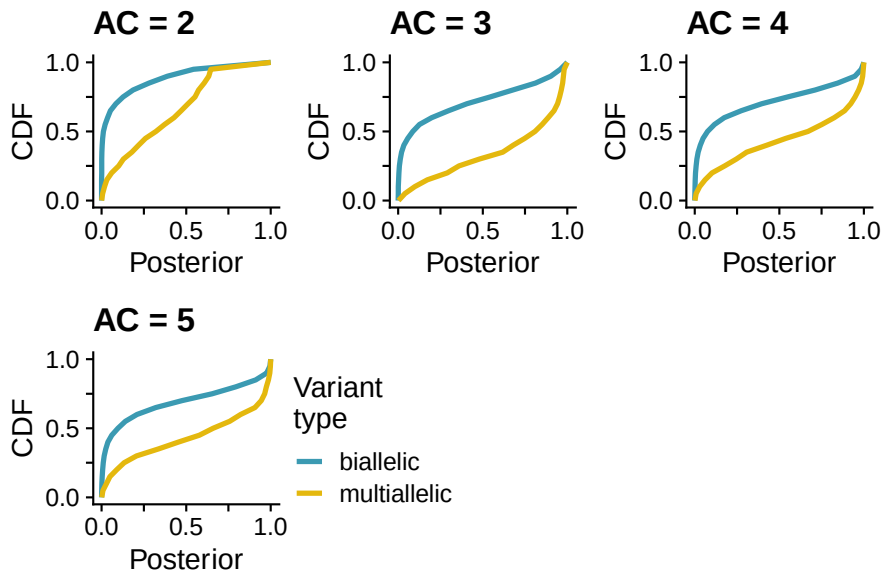
Supplementary Figure 5. The distribution of allele ages for IBD or recurrent mutations in simulations with and without selection.
For each allele count, each boxplot describes the distribution of allele ages for simulated IBD (top) or recurrent (bottom) variants, with (TRUE) or without (FALSE) selection present. The recurrent mutation ages are for each independent mutation event, and thus are on average more recent than IBD variants of the same allele count.



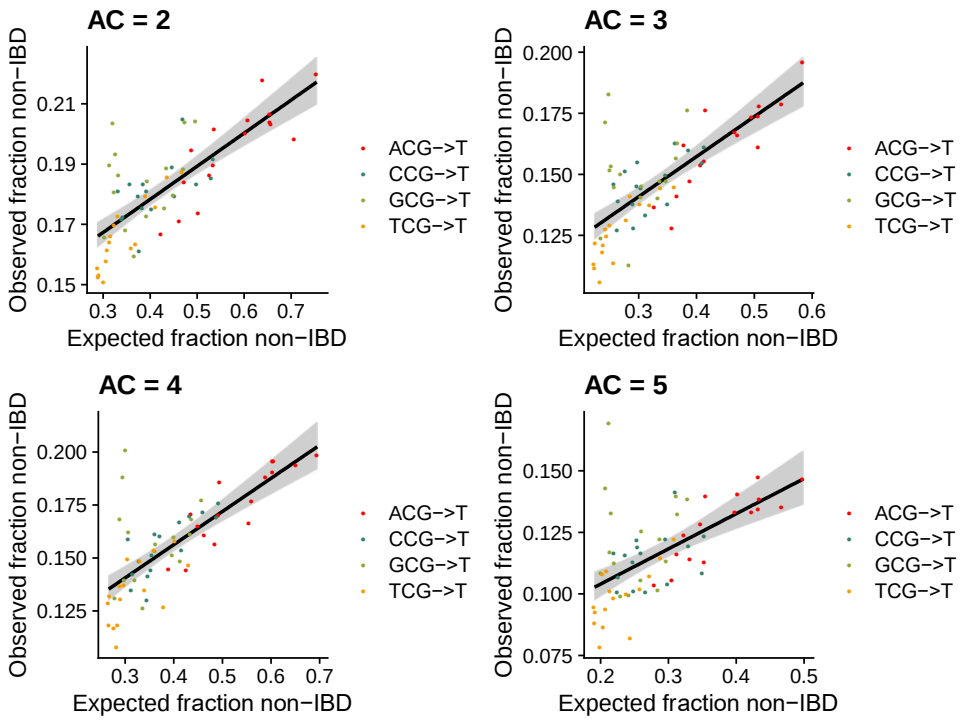
Supplementary Figure 6. ROC curves of the Bayesian hierarchical model's ability to distinguish IBD and non-IBD variants from simulations with variable recombination rates. **(A)** ROC curves for the Bayesian hierarchical model applied to distinguish simulated recurrent and IBD mutations, comparing simulations with variable recombination rate sampled from a human recombination map (solid line) vs. a uniform recombination map (dashed line). **(B)** ROC curves for the Bayesian hierarchical model applied to distinguish simulated recurrent and IBD mutations, comparing simulations with variable recombination rate and sample size 3,621 (solid line) vs. sample size 10,000 (dashed line). The solid lines in **(A)** and **(B)** represent the same simulations. AC: allele count, FPR: false positive rate, TPR: true positive rate.



Supplementary Figure 7. ROC curves comparing the performance of the Bayesian hierarchical model to age estimates from *runtc*.
ROC curves comparing the performance of our Bayesian hierarchical model (BHM; orange) vs. variant age estimates from *runtc* (blue) to distinguish IBD and recurrent mutations in simulations.

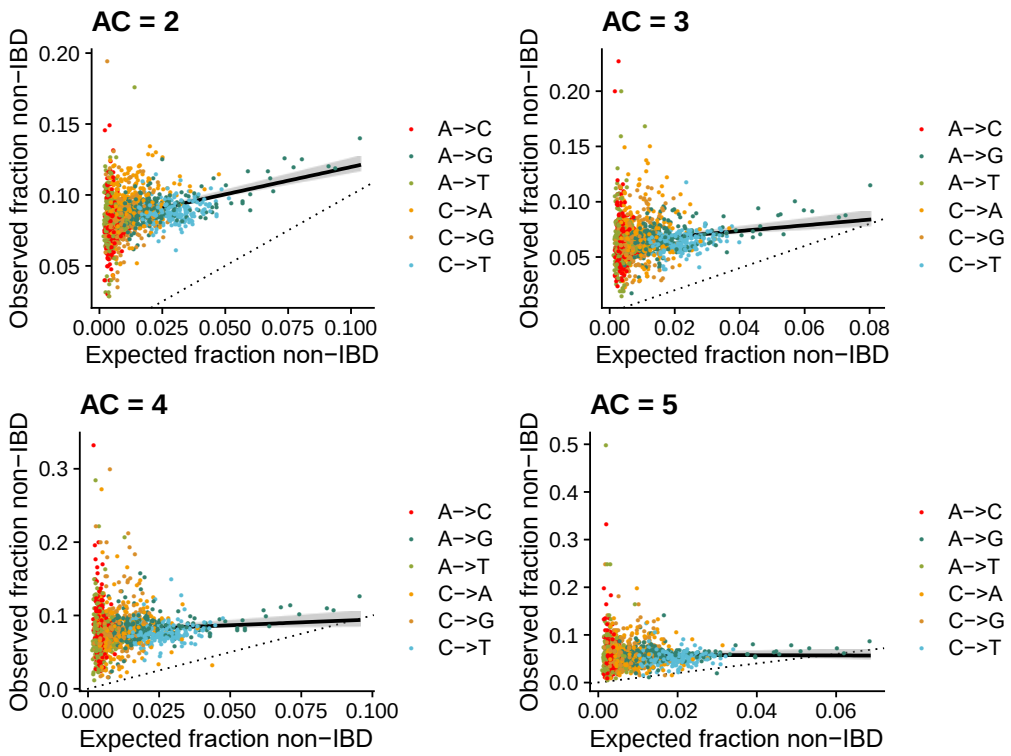


Supplementary Figure 8. The distribution of posterior probabilities of being non-IBD for biallelic vs. multiallelic variants from UK10K.
Posterior probabilities were calculated by applying our hierarchical model to biallelic (blue) or multiallelic (orange) sites from the UK10K dataset. Each panel represents variants of a given allele count (AC). CDF: cumulative distribution function.

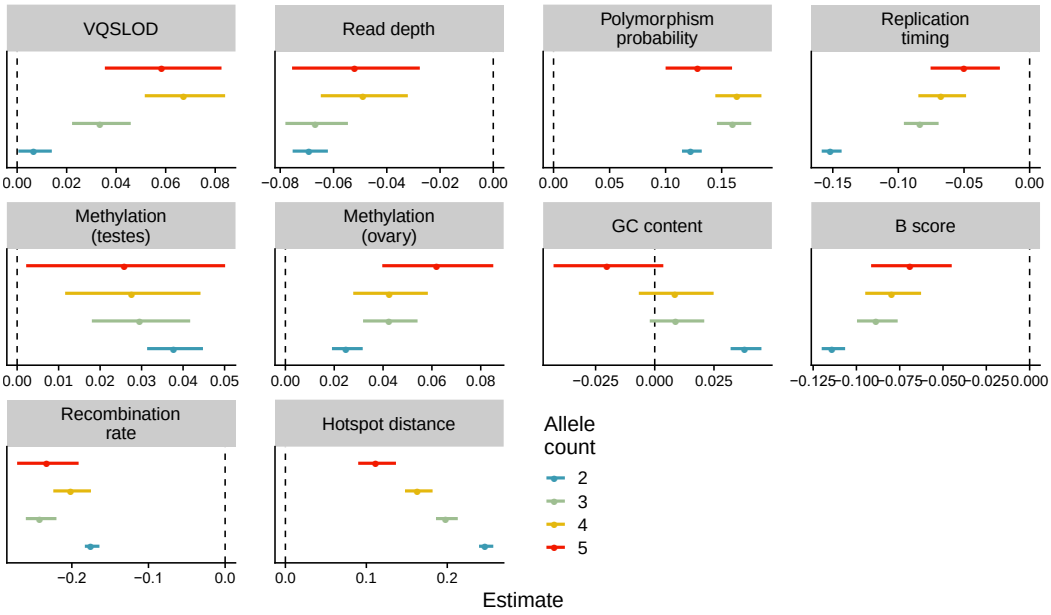


Supplementary Figure 9. The expected and observed fraction of sites called non-IBD for UK10K variants at CpG sites.

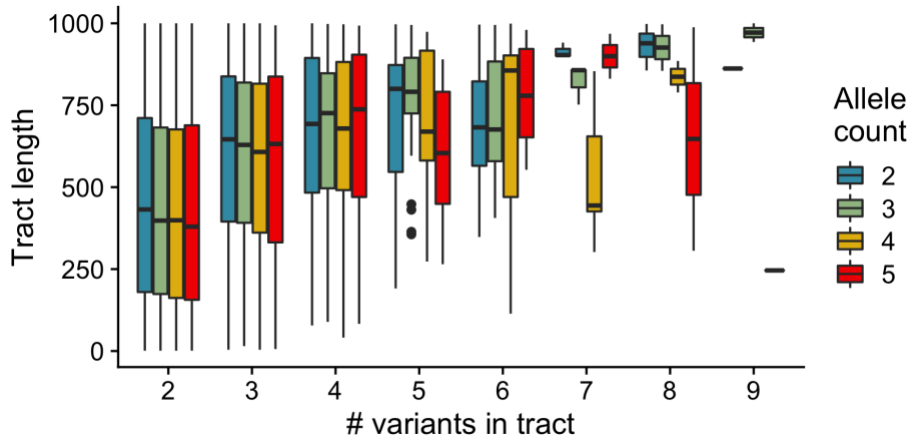
Each dot represents a 5-mer sequence context. The expected fraction is proportional to the polymorphism probability. Dot colors correspond to 3-mer sequence contexts.



Supplementary Figure 10. The expected and observed fraction of sites called non-IBD for UK10K variants at non-CpG sites.
Each dot represents a 5-mer sequence context. The expected fraction is proportional to the polymorphism probability. Dot colors represent the 1-mer sequence context.



Supplementary Figure 11. Results of logistic regression using genomic annotations to predict non-IBD variant calls for CpG->T variants.
Logistic regression of genomic annotations (predictor variables) vs. non-IBD variant calls (outcome) for CpG->T sites only, grouped by allele count. Dot colors represent allele count. Each dot's position denotes its beta coefficient estimate, with error bars representing beta ± 1.96*standard error. The vertical dashed line represents a beta estimate of zero. Hotspot distance: physical distance to nearest recombination hotspot z-score; Recombination rate: local recombination rate z-score; B score: McVicker's B statistic z-score; Replication timing: replication timing z-score; GC content: local GC content z-score; Methylation (ovary): ovary CpG methylation z-score; Methylation (testes): testes CpG methylation z-score; Read depth: read depth z-score; VQSLOD: variant quality z-score.



Supplementary Figure 12. Distribution of putative gene conversion tract lengths.
Each box plot represents the distribution of tract lengths for putative gene conversion tracts for a given allele count and number of variants in the tract. Note that our heuristic approach to identify putative gene conversions required a maximum length of 1000 base pairs.