Statistical fine-mapping of 94 complex diseases and traits in UK Biobank

Dec. 3rd, 2019

Common human diseases and traits are often caused by thousands of small effect genetic variants. Identifying causal common genetic variants is difficult due to a lack of recombination between sets of nearby variants, resulting in correlation typically known as linkage disequilibrium (LD). Fine-mapping is a Bayesian approach used to jointly model genetic signals and LD to statistically identify causal genetic variants from these associations.

Genetic association was performed as follows for 94 heritable traits in the UK Biobank:

1. Up to 361,194 individuals of white British ancestries with available phenotypes were included in GWAS, as determined by the PCA-based sample selection criteria.
2. Variants with INFO > 0.8, MAF > 0.01% (except for rare coding variants with MAC > 0), and HWE p-value > 1e-10 were included in association studies.
3. Quantitative traits were inverse rank normal transformed (excluding ordinal phenotypes: MCP, morning person, and insomnia), and BOLT-LMM (Loh et al. 2018 Nature Genetics) was used to estimate associations controlling for top 20 PCs, sex, age, age^2, sex * age, sex * age^2, and dilution factor where appropriate (infinitesimal model). For binary traits, SAIGE (Zhou et al. 2018 Nature Genetics) was used to estimate associations using the same covariates. Exceptions including LOY and Alzheimer's disease (see Thompson et al. 2019 bioRxiv and Hujoel et al. 2019 bioRxiv).

Fine-mapping was performed as follows for 94 heritable traits in the UK Biobank:

1. Regions for fine-mapping were defined by greedily starting with the most significantly associated (highest chi-square) variant, including all genome-wide significant (p < 5e-8) variants within a window of 3 Mb centered at the variant, and merging overlapping regions.
2. Beta and SE(Beta) were obtained from BOLT-LMM and SAIGE summary statistics.
3. In-sample dosage LD was estimated using LDStore v2.0b.
4. Fine-mapping was conducted using FINEMAP v1.3.1 (Benner et al. 2016 Bioinformatics, 2018 bioRxiv) and SuSiE v0.8.1.0521 (Wang et al. 2018 bioRxiv) with the inputs of summary statistics, in-sample dosage LD, sample size, and variance of phenotype. The maximum number of causal variants for each locus was specified as 10.

Post-processing:

1. Variants in the MHC region (chr6: 25–36 Mb) were excluded.
2. Variants and 95% credible sets containing variants with MAC < 100 were excluded.
3. Variants in moderate LD (R^2 > 0.6) with variants that failed HWE (p < 1e-12) in white British individuals but were used by the UKBB for imputation were identified using UK10K LD (computed in Hail 0.2) and are flagged as lower confidence (see our blog post).
4. Variants in strong LD (R^2 > 0.8) with common structural variants (SVs) were identified using gnomAD LD (computed in Hail 0.2, see Collins et al. 2019 bioRxiv) and are flagged as lower confidence.
5. Variants in 95% CSs or with PIPs > 0.001 are included in the primary .tsv or bed files. All fine-mapped regions and variants (including ~1% that failed) are listed in the secondary region .bed file.

Column descriptions (UKBB_94traits_release1.{tsv|bed}.gz):

1. chromosome: chromosome in hg19 coordinates (autosomes only)
2. start: start position of variant in hg19 coordinates (0-indexed)
3. end: end position of variant in hg19 coordinates (0-indexed)
4. variant: unique variant identifier (chr:pos:ref:alt)
5. rsid: rsID identifier
6. allele1: reference allele in hg19 coordinates
7. allele2: alternative allele in hg19 coordinates
8. minorallele: minor allele in cohort
9. cohort: GWAS cohort
10. model_marginal: type of regression model used
11. method: fine-mapping method used
12. trait: abbreviation for phenotype used for genetic association tests
13. region: region of the genome fine-mapping in hg19 coordinates
14. maf: allele frequency of the minor allele in cohort
15. beta_marginal: marginal association effect size from linear mixed model (effect allele: alternative)
16. se_marginal: standard error on marginal association effect size from linear mixed model
17. chisq_marginal: test statistic for marginal association
18. pip: posterior probability of association from fine-mapping
19. cs_id: ID of 95% credible set (-1 indicates that variant is not in a 95% CS)
20. beta_posterior: posterior expectation of true effect size (effect allele: alternative)
21. sd_posterior: posterior standard deviation of true effect size
22. LD_HWE: indicator that the variant is in LD (R^2 > 0.6) with a variant that failed Hardy Weinberg equilibrium (p < 10^-12) that was also used in phasing based upon UK10K LD

(http://www.nealelab.is/blog/2019/9/17/genotyped-snps-in-uk-biobank-failing-hardy-weinberg-equilibrium-test)

23. LD_SV: indicator that the variant is in LD (R^2 > 0.8) with a common structural variant based upon European samples from gnomAD (Collins et al. bioRxiv 2019)

Column descriptions (UKBB_94traits_release1_regions.bed.gz):

1. chromosome: chromosome in hg19 coordinates (autosomes only)
2. start: start position of variant in hg19 coordinates (0-indexed)
3. end: end position of variant in hg19 coordinates (0-indexed)
4. cohort: GWAS cohort
5. trait: abbreviation for phenotype used for genetic association tests
6. region: region of the genome fine-mapping in hg19 coordinates
7. variant: unique variant identifier (chr:pos:ref:alt)
8. success_finemap: if FINEMAP successfully completely
9. success_susie: if SuSiE successfully completely

Contacts:

* Jacob Ulirsch (julirsch@broadinstitute.org)
* Masahiro Kanai (mkanai@broadinstitute.org)
* Pardis Sabeti (pardis@broadinstitute.org)
* Hilary Finucane (finucane@broadinstitute.org)

Funding:

This work was supported by NHGRI 5UM1HG009435 and NHGRI 1K99HG010669. M.K. was supported by a Nakajima Foundation Fellowship and the Masason Foundation. J.U. was supported by NIH training grant NRSA 5T32GM007226.

Acknowledgements:

UK Biobank analyses were conducted via application 31063. We thank BM Neale, MJ Daly, and their colleagues for providing scripts and resources for the UK Biobank analyses.

Change log:

- Dec 3, 2019: corrected small formatting errors. changed primary format to tsv. added secondary bed format.
- Oct 17, 2019: added a description of effect alleles for betas