Mixed-model association for biobank-scale datasets

To the Editor — Despite recent work highlighting the advantages of linear mixed-model (LMM) methods for genome-wide association studies (GWAS) in datasets containing relatedness or population structure[1–3], much uncertainty remains about best practices for optimizing GWAS power while controlling confounders. Several recent studies of the interim UK Biobank dataset[4] (~150,000 samples) removed >20% of samples by filtering for relatedness or genetic ancestry and/or used linear regression in preference to mixed-model association. These issues are exacerbated in the full UK Biobank dataset (~500,000 samples), in which suggested sample exclusions decrease sample size by nearly 30%[5]. Here we release a much faster version of our BOLT-LMM Bayesian mixed-model association method[3] and show that it can be applied with minimal sample exclusions and achieves greatly superior power as compared to common practices for analyzing UK Biobank data.

In analyses of 23 highly heritable UK Biobank phenotypes (Supplementary Table 1), we observed that BOLT-LMM (applied to all 459,327 European samples and ~20 million imputed variants) consistently achieved far greater association power than linear regression with principal-component (PC) covariates (on 337,539 unrelated British samples, following ref. 5), attaining an 84% increase in GWAS locus discovery (10,759 total independent loci versus 5,839; Fig. 1a and Supplementary Table 2). These gains in power were driven only partially by the increased number of samples analyzed; we observed that BOLT-LMM achieved effective sample sizes as high as ~700,000 by conditioning on polygenic predictions from genome-wide SNPs, which effectively reduces noise in an association test[2,3,6] (Fig. 1b, Supplementary Fig. 1 and Supplementary Table 3). (We estimated effective sample size by taking ratios of chi-squared statistics from BOLT-LMM versus linear regression at GWAS hits; Supplementary Note.) The large sample size of the UK Biobank—which enables BOLT-LMM to predict and condition away up to 43% of phenotypic variance (approaching h_g^2 for several traits; Fig. 1b)—is now demonstrating the full power of this approach. We also confirmed that BOLT-LMM achieved substantial gains in power on the unrelated British sample set (Supplementary Tables 2 and 3).

To verify that BOLT-LMM analyses of all European samples were robust to potential confounding due to relatedness or population structure, we performed LD score regression (LDSC) analyses[7] of association statistics computed using both BOLT-LMM (on all European samples, allowing related individuals as well as population structure) and linear regression (on unrelated British samples stringently quality controlled to minimize confounding[5]); we ran LDSC using the baselineLD model[8]. We observed that, while the value of the LD score regression intercept (previously proposed as an indicator of confounding[7]) was generally difficult to interpret owing to attenuation bias[3], which causes the intercept to rise above 1 with increasing sample size and heritability (Supplementary Fig. 2 and Supplementary Note), the 'attenuation ratio'—(LDSC intercept − 1)/(mean χ^2 − 1)—matched closely between BOLT-LMM and PC-corrected linear regression and was relatively small (Fig. 1c, Supplementary Fig. 2 and Supplementary Table 4). Across 23 traits, we observed similar mean attenuation ratios of 0.078 (standard error (s.e.) 0.006) for PC-corrected linear regression and 0.082 (0.005) for BOLT-LMM, indicating that BOLT-LMM successfully controlled for sample structure (as expected for mixed-model methods)[1–3]. In contrast, uncorrected linear regression produced a mean attenuation ratio of 0.104 (0.012), indicating confounding (Supplementary Fig. 2 and Supplementary Table 4). Similarly, PC-corrected linear regression on all European samples exhibited slightly elevated attenuation ratios (mean 0.085, s.e. 0.006; binomial $P = 0.01$ versus attenuation on unrelated British samples), indicating slight confounding due to relatedness (Supplementary Table 4), while still achieving lower power than BOLT-LMM (Supplementary Table 2). We note that attenuation ratios are broadly smaller under the LDSC baselineLD model[8], which incorporates functional and linkage disequilibrium (LD)-related genome annotations, than under the original LDSC model (Supplementary Table 5), consistent with better model fit.

Our new BOLT-LMM software release (v2.3) implements additional computational improvements that provide ~4× speed-up, achieving running times that scaled nearly linearly with sample size and were comparable to those for linear regression (a few days for UK Biobank analyses; Fig. 1d and Supplementary Table 6). BOLT-LMM v2.3 performs much faster processing of imputed genotypes (the bottleneck for analyses of extremely large imputed datasets) via fast, multithreaded test statistic computation on imputed genotypes in BGEN v1.2 format (Supplementary Note). Additionally, for analyses of very large datasets, we now recommend including PC covariates for the purpose of accelerating convergence of iterative computations performed during BOLT-LMM's model-fitting steps[3]. Projecting out top PCs improves the conditioning of the matrix computations that BOLT-LMM implicitly performs, roughly halving the iterations required for convergence (Supplementary Table 7).

Our results demonstrate the latent power that mixed-model association analysis unlocks in very large GWAS, both by reducing the need for sample exclusions and by amplifying effective sample sizes via conditioning on polygenic predictions from genome-wide SNPs. (We note that, in general, care must be taken to consider non-additive effects when retaining related individuals; however, the level of relatedness in UK Biobank was low enough not to noticeably affect the overall genetic structure of the dataset or interpretation of our results (Supplementary Note).) Our new BOLT-LMM release makes mixed-model association computationally efficient even on extremely large datasets without requiring distributed computing[9]. Our analyses also reveal subtleties in the interpretation of LD score regression intercepts as a means of differentiating polygenicity from confounding in very large GWAS; the attenuation ratio may be a more suitable metric as sample sizes increase. Finally, we note two caveats regarding mixed-model analysis of binary traits. First, chi-squared-based tests (such as BOLT-LMM) can incur inflated type I error rates when used to analyze highly unbalanced case–control traits[10]; here the binary traits we analyzed were sufficiently balanced for

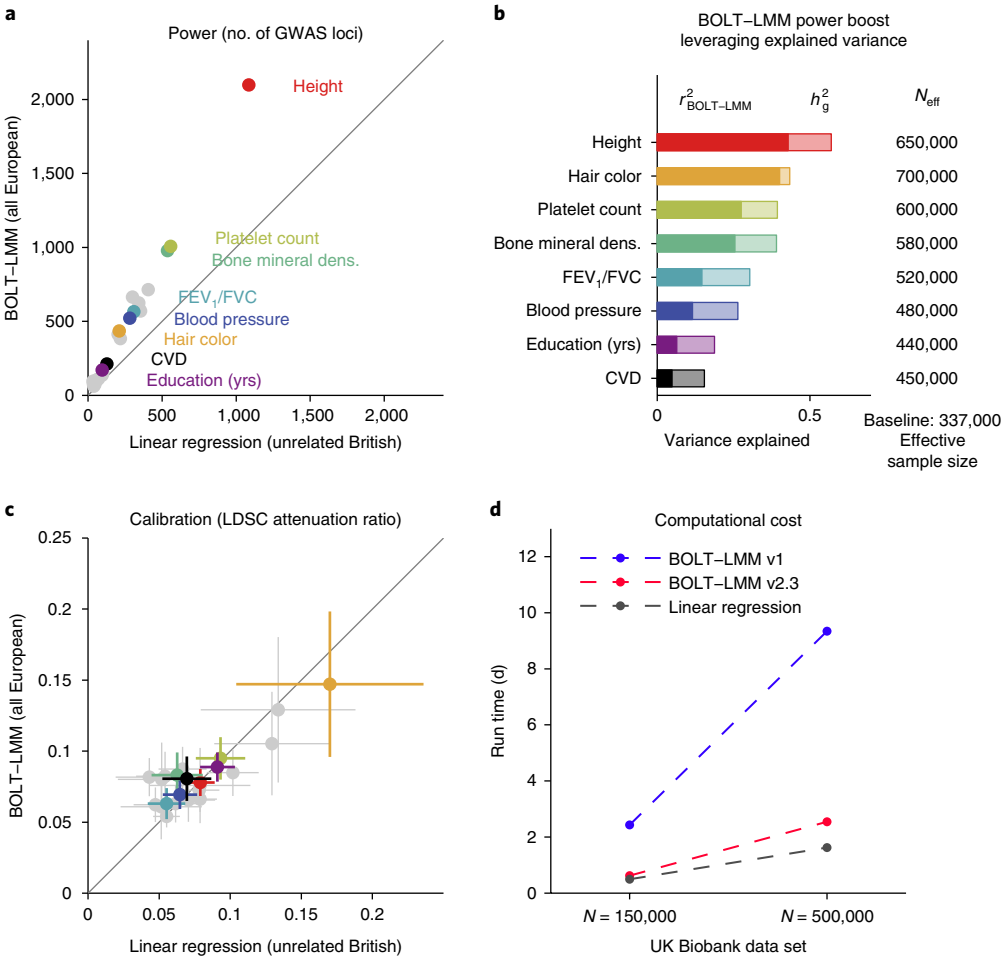

Fig. 1 | Power, calibration and speed of BOLT-LMM v2.3 in UK Biobank analyses. a, Numbers of independent genome-wide significant associations ($P < 5 \times 10^{-9}$) identified by BOLT-LMM analyses of all European-ancestry individuals ($N = 459{,}327$) versus linear regression analyses of unrelated British individuals ($N = 337{,}539$, following common practice[5]). Results for 23 phenotypes are plotted, with 8 representative phenotypes highlighted. **b**, Variance explained by genome-wide SNPs on which BOLT-LMM implicitly conditions to increase power. Conditioning on BOLT-LMM's polygenic predictions—which attain accuracy ($r^2_{\text{BOLT-LMM}}$) approaching SNP heritability (h^2_g) for some traits—achieves effective sample sizes (N_{eff}) as high as ~700,000. (We measured effective sample size by comparing χ^2 statistics at associated SNPs; Supplementary Note.) **c**, Test statistic calibration of BOLT-LMM on all European individuals versus linear regression on unrelated British individuals (using 20 PC covariates). Attenuation ratios from LD score regression[7,8] match closely between the two methods, indicating that BOLT-LMM properly controls false positives (Supplementary Fig. 2). Error bars, jackknife standard error ($N = 200$ blocks). **d**, Computational cost of association analysis using BOLT-LMM v2.3, the previous version of BOLT-LMM[3] and linear regression (implemented efficiently within the BOLT-LMM software) on the UK Biobank $N = 150{,}000$ and $N = 500{,}000$ data releases. Analyses were run on eight threads on a 2.10 GHz Intel Xeon E5-2683 v4 processor. Additional details and numerical data are provided in the Supplementary Note, Supplementary Fig. 1 and Supplementary Tables 1–9.

our results not to be impacted by this issue (which begins to arise at case fractions <10% in UK Biobank–scale sample sizes; Supplementary Note and Supplementary Table 8), but in general the saddlepoint approximation of SAIGE[10] is more robust in such scenarios. Second, conditioning on genome-wide signal can produce loss of power under case–control ascertainment[2,3]; specialized LMM methods are needed for modeling this scenario at scale. Overall, we hope that our findings provide clarity on analytical best practices for maximizing the value of large biobanks.

Code and data availability

BOLT-LMM v2.3 is open-source software freely available at http://data.broadinstitute.org/alkesgroup/BOLT-LMM/. Access to the UK Biobank resource is available via application (http://www.ukbiobank.ac.uk/). BOLT-LMM association statistics computed in this study are currently available for public download at http://data.broadinstitute.org/alkesgroup/UKBB/ and have been submitted to the UK Biobank Data Showcase. ❐

Po-Ru Loh[1,2]*, Gleb Kichaev[3], Steven Gazal[2,4], Armin P. Schoech[2,4,5] and Alkes L. Price[2,4,5]*

[1]*Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA.* [2]*Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.* [3]*Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, USA.* [4]*Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.* [5]*Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.*
**e-mail: poruloh@broadinstitute.org; aprice@hsph.harvard.edu*

Published online: 11 June 2018
https://doi.org/10.1038/s41588-018-0144-6

References

1. Yu, J. et al. *Nat. Genet.* **38**, 203–208 (2006).
2. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. *Nat. Genet.* **46**, 100–106 (2014).
3. Loh, P.-R. et al. *Nat. Genet.* **47**, 284–290 (2015).
4. Sudlow, C. et al. *PLoS Med.* **12**, 1–10 (2015).
5. Bycroft, C. et al. Preprint at *bioRxiv* https://www.biorxiv.org/content/early/2017/07/20/166298 (2017).
6. Listgarten, J. et al. *Nat. Methods* **9**, 525–526 (2012).
7. Bulik-Sullivan, B. K. et al. *Nat. Genet.* **47**, 291–295 (2015).
8. Gazal, S. et al. *Nat. Genet.* **49**, 1421–1427 (2017).
9. Canela-Xandri, O., Rawlik, K. & Tenesa, A. Preprint at *bioRxiv* https://www.biorxiv.org/content/early/2017/08/16/176834 (2017).
10. Zhou, W. et al. Preprint at *bioRxiv* https://www.biorxiv.org/content/early/2017/11/15/212357 (2017).

Acknowledgements

We are grateful to H. Finucane and Y. Reshef for helpful discussions. This research was conducted using the UK Biobank Resource under application 10438 and was supported by US National Institutes of Health grants R01 HG006399, R01 GM105857 and R01 MH107649 (A.L.P.), a Burroughs Wellcome Fund Career Award at the Scientific Interfaces and the Next Generation Fund at the Broad Institute of MIT and Harvard (P.-R.L.), and a Boehringer Ingelheim Fonds fellowship (A.P.S.). Computational analyses were performed on the Orchestra High-Performance Compute Cluster at Harvard Medical School, which is partially supported by grant NCRR 1S10RR028832-01.

Author contributions

P.-R.L. and A.L.P. designed the study. P.-R.L., G.K., S.G. and A.P.S. performed analyses. All authors wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0144-6.

Subfunctionalization versus neofunctionalization after whole-genome duplication

To the Editor: The question of what the predominant evolutionary fate is of genes after duplication events has been intensely debated for decades[1,2]. Two articles in *Nature* (Lien et al.[3]) and *Nature Genetics* (Braasch et al.[4]) investigated the regulatory fate of gene duplicates after the salmonid-specific (Ss4R) and teleost-specific (Ts3R) whole-genome duplication (WGD) events, respectively.

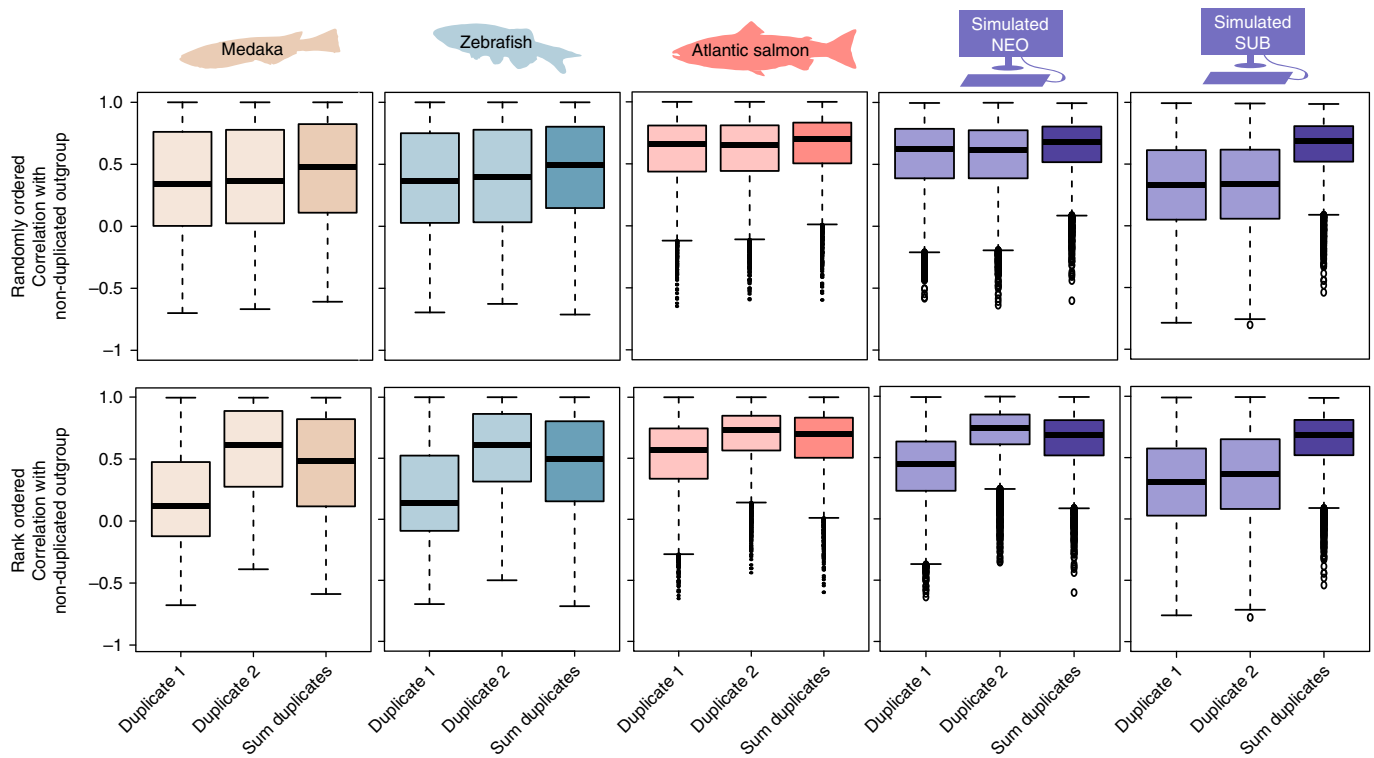

Fig. 1 | Tissue expression divergence in real and simulated data. Tissue expression correlation between duplicates in medaka or zebrafish and the corresponding orthologs in spotted gar (1,606 and 1,315 triplets, respectively) and between duplicates in Atlantic salmon and the orthologs in Northern pike (8,070 triplets). In the upper row, duplicated genes are assigned labels 'duplicate 1' and 'duplicate 2' randomly, while in the lower row the duplicates are ranked so that duplicate 1 has the lowest correlation with the ortholog and duplicate 2 has the highest. 'Sum duplicates' represents the correlation between the summed expression of the two duplicates and the ortholog in the unduplicated species. All correlations were computed using the Pearson correlation coefficient on the original expression data from the two publications in the first three columns and simulated data in the two last columns. All pairwise comparisons were statistically significant ($P < 5 \times 10^{-11}$, Wilcoxon signed-rank test, two-sided), with the exception of comparisons between duplicate 1 and duplicate 2 in the upper row (randomly ordered). Box plots were produced using the function 'boxplot' in R with default settings. The boxes indicate upper and lower quartiles with the horizontal lines marking the medians. The lines extending vertically from the boxes (whiskers) indicate the maximum and minimum values excluding outliers. Outliers are plotted as open circles. The expression data are available in the Supplementary Data.