BOLT-LMM v2.4 User Manual

Po-Ru Loh

July 22, 2022

Contents

1 Overview

The BOLT-LMM software package currently consists of two main algorithms, the BOLT-LMM algorithm for mixed model association testing, and the BOLT-REML algorithm for variance components analysis (i.e., partitioning of SNP-heritability and estimation of genetic correlations).

We recommend BOLT-LMM for analyses of human genetic data sets containing more than 5,000 samples. The algorithms used in BOLT-LMM rely on approximations that hold only at large sample sizes and have been tested only in human data sets. For analyses of fewer than 5,000 samples, we recommend the GCTA or GEMMA software.

We also note that BOLT-LMM association test statistics are valid for quantitative traits and for (reasonably) balanced case-control traits. For unbalanced case-control traits, we recommend the SAIGE software (see section 11 for a full discussion).

1.1 BOLT-LMM mixed model association testing

The BOLT-LMM algorithm computes statistics for testing association between phenotype and genotypes using a linear mixed model (LMM) [1]. By default, BOLT-LMM assumes a Bayesian mixture-of-normals prior for the random effect attributed to SNPs other than the one being tested. This model generalizes the standard "infinitesimal" mixed model used by previous mixed model association methods (e.g., EMMAX [2], FaST-LMM [3–6], GEMMA [7], GRAMMAR-Gamma [8], GCTA-LOCO [9]), providing an opportunity for increased power to detect associations while controlling false positives. Additionally, BOLT-LMM applies algorithmic advances to compute mixed model association statistics much faster than eigendecomposition-based methods, both when using the Bayesian mixture model and when specialized to standard mixed model association. BOLT-LMM is described in ref. [1]:

> Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, Patterson N, and Price AL. Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics*, 2015.

Additionally, ref. [10] explores the performance of BOLT-LMM on the full UK Biobank data set:

> Loh P-R, Kichaev G, Gazal S, Schoech AP, and Price AL. Mixed model association for biobank-scale data sets. *Nature Genetics*, 2018.

1.2 BOLT-REML variance components analysis

The BOLT-REML algorithm estimates heritability explained by genotyped SNPs and genetic correlations among multiple traits measured on the same set of individuals. Like the GCTA software [11], BOLT-REML applies variance components analysis to perform these tasks, supporting both multi-component modeling to partition SNP-heritability and multi-trait modeling to estimate correlations. BOLT-REML applies a Monte Carlo algorithm that is much faster than eigendecomposition-based methods for variance components analysis (e.g., GCTA) at large sample sizes. BOLT-REML is described in ref. [12]:

> Loh P-R, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ, PGC-SCZ Working Group, de Candia TR, Lee SH, Wray NR, Kendler KS, O'Donovan MC, Neale BM, Patterson N, and Price AL. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance components analysis. *Nature Genetics*, 2015.

2 Download and installation

You can download the latest version of the BOLT-LMM software at:

`http://data.broadinstitute.org/alkesgroup/BOLT-LMM/downloads/`

Previous versions are also available in the `old/` subdirectory.

2.1 Change log

- Version 2.4 (July 22, 2022):
  - Fixed slow performance on UK Biobank Research Analysis Platform (RAP) when analyzing BGEN input files mounted via dxfuse (in `/mnt/project`). (This issue was caused by files not being properly closed, which does not impact performance on most filesystems but caused problems with dxfuse.)
  - Added checks for file I/O errors.
  - Added support for phased BGEN files (using 8-bit encoding).
  - Added documentation of `--predBetasFile` option for polygenic prediction.
- Version 2.3.6 (October 29, 2021):
  - Fixed bug in scaling of BETA and SE columns in linear regression output. This bug only affected effect sizes computed for phenotypes with non-unit variance by BOLT-LMM v2.3.5 in linear regression mode; effect sizes from linear mixed model analysis (`--lmm`/`--lmmInfOnly`/`--lmmForceNonInf`) were unaffected.
- Version 2.3.5 (March 20, 2021):
  - Improved efficiency of phenotype/covariate file handling (by only loading requested columns).
  - Added BETA and SE columns to output when BOLT-LMM is run in linear regression mode.
- Version 2.3.4 (August 10, 2019):
  - Prevented environment variable `MKL_NUM_THREADS` from overriding `--numThreads` parameter.
- Version 2.3.3 (August 3, 2019):
  - Added support for missing values in BGEN v1.2 data.
  - Reduced memory usage after completion of the model-fitting step (by freeing genotypes no longer needed during calculation of association tests for imputed variants).
  - Updated BLAS library to Intel MKL 2019 Update 4 and modified `bolt` executable to dynamically link Intel threading library (libiomp5.so, now distributed with BOLT-LMM).
  - Improved error-reporting and documentation; added FAQ section of this manual.
- Version 2.3.2 (March 10, 2018):
  - Added support for imputed X chromosome variants in the UK Biobank v3 imputation release (see section 10.1 for details).

  - Added `--bgenSampleFileList` option to allow multiple `bgen/sample` file pairs with differing sample files.
  - Changed the `--allowX` option to always be set (so `--allowX` is no longer needed for X chromosome analysis).
  - Added `--Nautosomes` option for non-human analyses.

- Version 2.3.1 (December 19, 2017):
  - Added check for `--covarFile` when `--covarCol` or `--qCovarCol` are specified. (Previously, these arguments were silently ignored if `--covarFile` was not specified.)
  - Added `--allowX` option for X chromosome analysis.
  - Added section of documentation on analysis of case-control traits.
  - Updated BLAS library to Intel MKL 2018 Update 1.
  - Improved error-reporting and documentation.

- Version 2.3 (August 1, 2017):
  - Added multi-threaded support for BGEN v1.2 imputed file format used by UK Biobank.
  - Added `--bgenMinINFO` parameter.
  - Fixed P-value output for extremely small P-values previously rounded to zero.
  - Fixed crash caused by very long BGEN allele names.
  - Fixed bug in `--noBgenIDcheck` flag.
  - Improved error-reporting.
  - Added section at the end of this document providing recommendations for N=500K UK Biobank analyses.

- Version 2.2 (Nov 13, 2015):
  - Added support for testing imputed SNPs in BGEN format.
  - Added option to look up LD scores by base pair coordinates rather than SNP name (`--LDscoresMatchBp`).
  - Fixed bug in hg19 genetic map interpolation.
  - Fixed bug in QC filter for per-sample missing rate.
  - Improved error-checking.

- Version 2.1 (Apr 29, 2015):
  - Improved handling of IMPUTE2 files (large speedup; `INFO` output column instead of `F_MISS`; MAF filtering).

  - Fixed bug in per-sample missingness filter (`--maxMissingPerIndiv`), which was being ignored.
  - Implemented minor changes to input parameters (removed `--impute2CallThresh`, which had no effect; added `--impute2MinMAF` and `--h2gGuess`).

- Version 2.0 (Mar 13, 2015): Added BOLT-REML algorithm for estimating heritability parameters. Fixed parameter-initialization bug that prevented BOLT-LMM from running on some systems. Implemented various minor improvements to parameter-checking.

- (Dec 8, 2014): Licensed source code under GPLv3.

- Version 1.2 (Nov 4, 2014): Added support for testing imputed SNPs in 2-dosage format (Ricopili/plink2 format=2). Fixed bug causing `nan` heritability estimates.

- Version 1.1 (Oct 17, 2014): Added support for testing imputed SNPs with probabilistic dosages.

- Version 1.0 (Aug 8, 2014): Initial release.

2.2 Installation

The `BOLT-LMM_vX.X.tar.gz` download package contains a precompiled 64-bit Linux executable, `bolt`, which we have tested on several Linux systems. We recommend using this executable because it is well-optimized and no further installation is required. Note that beginning with BOLT-LMM v2.3.3, the `bolt` executable dynamically links the libiomp5.so Intel threading runtime library; this shared library is provided in the `lib/` subdirectory of the BOLT-LMM package and will be automatically loaded by the `bolt` executable from that subdirectory.

If you wish to compile your own version of the BOLT-LMM software from the source code (in the `src/` subdirectory), you will need to ensure that library dependencies are fulfilled and will need to make appropriate modifications to the `Makefile`:

- Library dependencies:
  - BLAS/LAPACK numerical libraries. The speed of the BOLT-LMM software depends critically on the efficiency of the BLAS/LAPACK implementation it is linked against. We recommend the Intel Math Kernel Library (MKL) if available (except on AMD processors); otherwise, ATLAS may be a good alternative.
  - Boost C++ libraries. BOLT-LMM links against the Boost `program_options` and `iostreams` libraries, which need to be installed after downloading and unzipping Boost.
  - NLopt numerical optimization library [13].

- `Makefile`: Paths to libraries need to be modified appropriately. Note that the released version of the `Makefile` does not set the flag `-DUSE_MKL_MALLOC`. This flag turns on the Intel MKL's fast memory manager (replacing calls to `_mm_malloc` with `mkl_malloc`),

which may improve memory performance, but we have observed crashes on some systems when using `mkl_malloc`.

For reference, the provided `bolt` executable was built on the Harvard Medical School "Orchestra 2" research computing cluster using Intel `icpc` 16.0.2 (with MKL 2019 Update 4) and the Boost 1.58.0 and NLopt 2.4.2 libraries by invoking `make linking=static-except-glibc`.

Port to Windows. Remi Daviet has created a version of BOLT-LMM that compiles in Windows: `http://remidaviet.com/software.php`

Port to FreeBSD. BOLT-LMM can be installed on FreeBSD via the FreeBSD ports system (`pkg install bolt-lmm`). Note that this installation will use only highly-portable (and potentially less fast) optimizations.

2.3 Running BOLT-LMM and BOLT-REML

To run the `bolt` executable, simply invoke `./bolt` on the Linux command line (within the BOLT-LMM install directory) with parameters in the format `--optionName=optionValue`.

2.4 Examples

The `example/` subdirectory contains a bash script `run_example.sh` that demonstrates basic use of BOLT-LMM on a small example data set. Likewise, `run_example_reml2.sh` demonstrates BOLT-REML.

- A minimal BOLT-LMM invocation looks like:

  ```
  ./bolt --bfile=geno --phenoFile=pheno.txt --phenoCol=phenoName
       --lmm --LDscoresFile=tables/LDSCORE.1000G_EUR.tab.gz
       --statsFile=stats.tab
  ```

- A minimal BOLT-REML invocation looks like:

  ```
  ./bolt --bfile=geno --phenoFile=pheno.txt --phenoCol=phenoName
       --reml --modelSnps=modelSnps.txt
  ```

  To perform multi-trait BOLT-REML (i.e., estimate genetic correlations), provide multiple `--phenoCol=`*phenoName* arguments.

2.5 Help

To get a list of basic options, run:

```
./bolt -h
```

To get a complete list of basic and advanced options, run:

```
./bolt --helpFull
```

3 Computing requirements

3.1 Operating system

At the current time we have only compiled and tested BOLT-LMM on Linux computing environments; however, the source code is available if you wish to try compiling BOLT-LMM for a different operating system.

3.2 Memory

For typical data sets (M, N exceeding 10,000), BOLT-LMM and BOLT-REML use approximately MN/4 bytes of memory, where M is the number of SNPs and N is the number of individuals. More precisely:

- M = # of SNPs in `bim` file(s) that satisfy all of the conditions:
  - not listed in any `--exclude` file
  - passed QC filter for missingness
  - listed in `--modelSnps` file(s), if specified
- N = # of individuals in `fam` file and not listed in any `--remove` file (but pre-QC; i.e., N includes individuals filtered due to missing genotypes or covariates)

3.3 Running time

In practice, BOLT-LMM and BOLT-REML have running times that scale roughly with $MN^{1.5}$. Analyses of the full UK Biobank data set (M $\sim$ 700K SNPs, N = 500K individuals) typically take a few days using 8 threads of a single compute node; for more details, please see refs. [1, 10].

3.3.1 Multi-threading

On multi-core machines, running time can be reduced by invoking multi-threading using the `--numThreads` option.

4 Input/output file naming conventions

4.1 Automatic gzip [de]compression

The BOLT-LMM software assumes that input files ending in `.gz` are gzip-compressed and automatically decompresses them on-the-fly (i.e., without creating a temporary file). Similarly, BOLT-LMM writes gzip-compressed output to any output file ending in `.gz`.

4.2 Arrays of input files and covariates

Arrays of sequentially-numbered input files and covariates can be specified by the shorthand `{i:j}`. For example,

```
data.chr{1:22}.bim
```

is interpreted as the list of files

```
data.chr1.bim, data.chr2.bim, ..., data.chr22.bim
```

5 Input

5.1 Genotypes

The BOLT-LMM software takes genotype input in PLINK [14] binary format (`bed/bim/fam`). For file conversion and data manipulation in general, we highly recommend the PLINK/PLINK2 software [15].

If all genotypes are contained in a single `bed/bim/fam` file triple with the same file prefix, you may simply use the command line option `--bfile=prefix`. Genotypes may also be split into multiple `bed` and `bim` files containing consecutive sets of SNPs (e.g., one `bed/bim` file pair per chromosome) either by using multiple `--bed` and `--bim` invocations or by using the file array shorthand described above (e.g., `--bim=data.chr{1:22}.bim`).

5.1.1 Reference genetic maps

The BOLT-LMM package includes reference maps that you can use to interpolate genetic map coordinates from SNP physical (base pair) positions in the event that your PLINK `bim` file does not contain genetic coordinates (in units of Morgans). (The BOLT-LMM association testing algorithm uses genetic positions to prevent proximal contamination; BOLT-REML does not use this information.) To use a reference map, use the option

```
--geneticMapFile=tables/genetic_map_hg##.txt.gz
```

selecting the build (hg17, hg18, hg19, or hg38) corresponding to the physical coordinates of your `bim` file. You may use the `--geneticMapFile` option even if your PLINK `bim` file does contain genetic coordinates; in this case, the genetic coordinates in the `bim` file will be ignored, and interpolated coordinates will be used instead.

5.1.2 Imputed SNP dosages

The BOLT-LMM association testing algorithm supports computation of mixed model association statistics at an arbitrary number of imputed SNPs (with real-valued "dosages" rather than hard-called genotypes) **using a mixed model built on a subset of hard-called, PLINK-format genotypes** (typically a subset of directly genotyped SNPs). (BOLT-REML variance components analysis does not support dosage input.)

When testing imputed SNPs, BOLT-LMM first performs its usual model-fitting on PLINK-format genotypes (supplied via `--bfile` or `bed/bim/fam`) and then applies the model to scan any provided imputed SNPs. The second step requires only a modest amount of additional computation and no additional RAM, as the it simply performs a genome scan (as in GRAMMAR-Gamma [8]) of real-valued dosage SNPs against the residual phenotypes that BOLT-LMM computes during model-fitting. We currently recommend performing model-fitting on ˜500K hard-called genotypes; this approach should sacrifice almost no statistical power while retaining computational efficiency.

If you have only imputed SNP data on hand, you will need to pre-process your data set to create a subset of hard-called SNPs in PLINK format for BOLT-LMM. We suggest the following procedure.

1. Determine a high-confidence set of SNPs (e.g., based on R^2 or INFO score) at which to create an initial hard-call set.
2. Create hard-called genotypes at these SNPs in PLINK format.
3. Use PLINK to LD prune to ˜500K SNPs (via `--indep-pairwise 50 5` *r2thresh* for an appropriate *r2thresh*).
4. Run BOLT-LMM using the final hard-called SNPs as the `--bfile` (or `bed/bim/fam`) argument, specifying the imputed SNPs as additional association test SNPs using one of the formats below.

Imputed SNPs in dosage format. This input format consists of one or more `--dosageFile` parameters specifying files that contain real-valued genotype expectations at imputed SNPs. Each line of a `--dosageFile` should be formatted as follows:

```
rsID   chr   pos   allele1   allele0   [dosage = E[#allele1]] x N
```

Missing (i.e., uncalled) dosages can be specified with –9. You will also need to provide one additional `--dosageFidIidFile` specifying the PLINK FIDs and IIDs of samples that the dosages correspond to. See the `example/` subdirectory for an example.

Imputed SNPs in IMPUTE2 format. You may also specify imputed SNPs as output by the IMPUTE2 software [16]. The IMPUTE2 genotype file format is as follows:

```
snpID   rsID   pos   allele1   allele0   [p(11) p(10) p(00)] x N
```

(BOLT-LMM ignores the `snpID` field.) Here, instead of dosages, each genotype entry contains individual probabilities of the individual being homozygous for allele1, heterozygous, and homozygous for allele0. The three probabilities need not sum to 1, allowing for genotype uncertainty; if the sum of the probabilities is less than the `--impute2CallThresh` parameter, BOLT-LMM treats the genotype as missing.

To compute association statistics at a list of files containing IMPUTE2 SNPs, you may list the files within a `--impute2FileList` file. Each line of this file should contain two entries:

a chromosome number followed by an IMPUTE2 genotype file containing SNPs from that chromosome. You will also need to provide one additional `--impute2FidIidFile` specifying the PLINK FIDs and IIDs of samples that the IMPUTE2 genotypes correspond to. See the `example/` subdirectory for an example.

Imputed SNPs in 2-dosage format. You may also specify imputed SNPs as output by the Ricopili pipeline and `plink2 --dosage format=2`. This file format consists of file pairs: (1) PLINK `map` files containing information about SNP locations; and (2) genotype probability files in the 2-dosage format, which consists of a header line

```
SNP   A1   A2   [FID IID] x N
```

followed by one line per SNP in the format

```
rsID   allele1   allele0   [p(11) p(10)] x N
```

The third genotype probability for each entry is assumed to be `p(00)=1-p(11)-p(10)` (unlike with the IMPUTE2 format).

To compute association statistics at SNPs in a list of 2-dosage files, you may list the files within a `--dosage2FileList` file. Each line of this file should contain two entries: a PLINK `map` file followed by the corresponding genotype file containing probabilities for those SNPs. (As usual, if either file ends with `.gz`, it is automatically unzipped; otherwise it is assumed to be plain text.) See the `example/` subdirectory for an example.

Imputed SNPs in BGEN format. To compute association statistics at SNPs in **one or more** BGEN data files, specify the `.bgen` file(s) with `--bgenFile` and the corresponding `.sample` file with `--sampleFile`. The `--bgenMinMAF` and `--bgenMinINFO` options allows limiting output to SNPs passing minimum allele frequency and INFO thresholds. (Note: the `--bgenMinMAF` filter is applied to the full BGEN file before any sample exclusions, whereas the MAF reported in BOLT-LMM's output is computed in the subset of samples actually analyzed. Some SNPs may therefore pass the `--bgenMinMAF` filter but have lower reported MAF in the output file; if you wish to exclude such SNPs, you will need to post-process the results.)

Note that starting with BOLT-LMM v2.3, the `--bgenFile` option allows multiple BGEN files. **We have implemented multi-threaded processing for files in the BGEN v1.2 format** used in the UK Biobank N=500K release, so analyzing BGEN v1.2 data for all chromosomes within a single job is now feasible. For analyses of BGEN v1.1 data used in the N=150K release, we recommend parallelizing across chromosomes for computational convenience (using the full `--bfile` of directly genotyped PLINK data from all chromosomes in each job).

Additionally, starting with BOLT-LMM v2.3.2, you may alternatively specify a list of whitespace-separated `.bgen` / `.sample` file pairs using the `--bgenSampleFileList` option (instead of using the `--bgenFile` and `--sampleFile` options). This option enables analyses of data sets in which different BGEN files have different sample sets (e.g., the UK Biobank v3 imputation release; section 10.1).

WARNING: The BGEN format comprises a few sub-formats; we have only implemented support for the versions (and specific data layouts) used in the UK Biobank N=150K and N=500K releases. In particular, for BGEN v1.2, BOLT-LMM currently only supports the 8-bit encoding used for the UK Biobank N=500K data. (Starting with BOLT-LMM v2.3.3, missing values in BGEN v1.2 data are now allowed.)

Imputed SNPs in VCF format, exome-sequencing SNP calls in plink format, etc. BOLT-LMM does not support imputed data formats not listed above, so we recommend converting other data formats to BGEN v1.2 using PLINK2. As noted above, you will need to use the same sub-format of BGEN v1.2 used by UK Biobank:

- Specify 8-bit encoding via: `plink2 --export bgen-1.2 bits=8`
- If you wish to analyze X chromosome data, you will need to create a BGEN v1.2 file in which all genotypes are coded as diploid. By default, plink2 will code males as haploid, but you can force it to create diploid X chromosome data by setting the sex of all individuals to female before converting.

(Starting with BOLT-LMM v2.4, phased 8-bit BGEN v1.2 files are now supported, such that phase information no longer needs to be erased before converting to BGEN.)

5.1.3 X chromosome analysis

Starting with v2.3.2, BOLT-LMM accepts X chromosome genotypes for both model-fitting (via `--bfile` or `--bed/bim/fam` PLINK-format input) and association testing on imputed variants (e.g., in BGEN files). Males should be coded as diploid (as PLINK does for chromosome code 23 = X non-PAR), such that male genotypes are coded as 0/2 and female genotypes are coded as 0/1/2 (corresponding to a model of random X inactivation). There is no need to separate chrX into PAR and non-PAR; for PLINK input, you should simply merge PAR and non-PAR SNPs into a single "chromosome 23" using PLINK `--merge-x`.

Imputed X chromosome SNPs can also be included in BOLT-LMM association tests; again, males should be coded as diploid in one of the currently-supported formats (e.g., BGEN v1.1 or 8-bit BGEN v1.2). (BGEN v1.2 includes a data format that natively encodes a mixture of haploid and diploid SNPs, but BOLT-LMM currently does not support this format.) Chromosomes named 23, X, XY, PAR1, and PAR2 are all acceptable.

5.2 Phenotypes

Phenotypes may be specified in either of two ways:

- `--phenoUseFam`: This option tells BOLT-LMM and BOLT-REML to use the last (6th) column of the `fam` file as the phenotypes. This column must be numeric, so case-control phenotypes should be `1,0` coded and missing values should be indicated with `-9`.

- `--phenoFile` and `--phenoCol`: Alternatively, phenotypes may be provided in a separate whitespace-delimited file (specified with `--phenoFile`) with the first line containing column headers and subsequent lines containing records, one per individual. The first two columns must be `FID` and `IID` (the PLINK identifiers of an individual). Any number of columns may follow; the column containing the phenotype to analyze is specified with `--phenoCol`. Values of `-9` and `NA` are interpreted as missing data. All other values in the column should be numeric. The records in lines following the header line need not be in sorted order and need not match the individuals in the genotype data (i.e., `fam` file); BOLT-LMM and BOLT-REML will analyze only the individuals in the intersection of the genotype and phenotype files and will output a warning if these sets do not match.

5.3 Covariates

Covariate data may be specified in a file (`--covarFile`) with the same format as the alternate phenotype file described above. (If using the same file for both phenotypes and covariates; `--phenoFile` and `--covarFile` must still both be specified.) Each covariate to be used must be specified using either a `--covarCol` (for categorical covariates) or a `--qCovarCol` (for quantitative covariates) option. Categorical covariate values are allowed to be any text strings not containing whitespace; each unique text string in a column corresponds to a category. (To guard against users accidentally specifying quantitative covariates with `--covarCol` instead of `--qCovarCol`, BOLT-LMM throws an error if a categorical covariate contains more than 10 distinct values; this upper bound can be modified with `--covarMaxLevels`.) Quantitative covariate values must be numeric (with the exception of `NA`). In either case, values of `-9` and `NA` are interpreted as missing data. If groups of covariates of the same type are numbered sequentially, they may be specified using array shorthand (e.g., `--qCovarCol=PC{1:10}` for columns `PC1, PC2, ..., PC10`).

5.4 Missing data treatment

Individuals with missing phenotypes are ignored. By default, individuals with any missing covariates are also ignored; this approach is commonly used and referred to as "complete case analysis." As an alternative, we have also implemented the "missing indicator method" (via the `--covarUseMissingIndic` option), which adds indicator variables demarcating missing status as additional covariates.

Missing genotypes in plink data (`--bfile` or `bed/bim/fam`) are replaced with per-SNP averages. Imputed genotypes should not contain missing data; standard imputation software always produces genotype probability estimates even if uncertainty is high.

5.5 Genotype QC

BOLT-LMM and BOLT-REML automatically filter SNPs and individuals with missing rates exceeding thresholds of 0.1. These thresholds may be modified using `--maxMissingPerSnp` and `--maxMissingPerIndiv`. Note that filtering is **not** performed based on minor allele frequency

or deviation from Hardy-Weinberg equilibrium. Allele frequency and missingness of each SNP are included in the BOLT-LMM association test output, however, and we recommend checking these values and Hardy-Weinberg *p*-values (which are easily computed using PLINK `--hardy`) when following up on significant associations.

5.6 User-specified filters

Individuals to remove from the analysis may be specified in one or more `--remove` files listing `FID` and `IIDs` (one individual per line). Similarly, SNPs to exclude from the analysis may be specified in one ore more `--exclude` files listing SNP IDs (typically rs numbers).

Note that `--exclude` filters are **not** applied to imputed data; exclusions of specific imputed SNPs will need to be performed separately as a post-processing step.

6 Association analysis (BOLT-LMM)

6.1 Mixed model association tests

BOLT-LMM computes two association statistics, $\chi^2_{\text{BOLT-LMM}}$ and $\chi^2_{\text{BOLT-LMM-inf}}$, described in detail in our manuscript [1].

- **BOLT-LMM: Association test on residuals from Bayesian modeling using a mixture-of-normals prior on SNP effect sizes.** This approach can fit "non-infinitesimal" traits with loci having moderate to large effects, allowing increased association power.
- **BOLT-LMM-inf: Standard (infinitesimal) mixed model association.** This statistic approximates the standard approach used by eigendecomposition-based software.

6.2 BOLT-LMM mixed model association options

The BOLT-LMM software offers the following options for mixed model analysis:

- `--lmm`: Performs default BOLT-LMM analysis, which consists of (1a) estimating heritability parameters, (1b) computing the BOLT-LMM-inf statistic, (2a) estimating Gaussian mixture parameters, and (2b) computing the BOLT-LMM statistic *only if an increase in power is expected*. If BOLT-LMM determines based on cross-validation that the non-infinitesimal model is likely to yield no increase in power, the BOLT-LMM (Bayesian) mixed model statistic is not computed.
- `--lmmInfOnly`: Computes only infinitesimal mixed model association statistics (i.e., steps 1a and 1b).
- `--lmmForceNonInf`: Computes both the BOLT-LMM-inf and BOLT-LMM statistics *regardless of whether or not an increase in power is expected* from the latter.

6.2.1 Reference LD score tables

A table of reference LD scores [17] is needed to calibrate the BOLT-LMM statistic. Reference LD scores appropriate for analyses of European-ancestry samples are provided in the `tables/` subdirectory and can be specified using the option

```
--LDscoresFile=tables/LDSCORE.1000G_EUR.tab.gz
```

For analyses of non-European data, we recommend computing LD scores using the LDSC software on an ancestry-matched subset of the 1000 Genomes samples.

By default, LD scores in the table are matched to SNPs in the PLINK data by rsID. The `--LDscoresMatchBp` option allows matching SNPs by base pair coordinate.

6.2.2 Restricting SNPs used in the mixed model

If millions of SNPs are available from imputation, we suggest including at most 1 million SNPs at a time in the mixed model (using the `--modelSnps` option) when performing association analysis. Using an LD pruned set of at most 1 million SNPs should achieve near-optimal power and correction for confounding while reducing computational cost and improving convergence. Note that even when a file of `--modelSnps` is specified, all SNPs in the genotype data are still tested for association; only the random effects in the mixed model are restricted to the `--modelSnps`. Also note that BOLT-LMM automatically performs leave-one-chromosome-out (LOCO) analysis, leaving out SNPs from the chromosome containing the SNP being tested in order to avoid proximal contamination [4, 9].

6.3 Standard linear regression

Setting the `--verboseStats` flag will output standard linear regression chi-square statistics and *p*-values in additional output columns `CHISQ_LINREG` and `P_LINREG`. Note that unlike mixed model association, linear regression is susceptible to population stratification, so you may wish to include principal components (computed using other software, e.g., PLINK2 or FastPCA [18] in EIGENSOFT v6.0+) as covariates when performing linear regression. Including PCs as covariates will also speed up convergence of BOLT-LMM's mixed model computations.

7 Variance components analysis (BOLT-REML)

Using the `--reml` option invokes the BOLT-REML algorithm for estimating heritability parameters and genetic correlations.

7.1 Multiple variance components

To assign SNPs to different variance components, specify a `--modelSnps` file in which each whitespace-delimited line contains a SNP ID (typically an rs number) followed by the name of the variance component to which it belongs.

7.2 Multiple traits

To perform multi-trait variance components analysis, specify multiple `--phenoCol` parameter-value flags (corresponding to different columns in the same `--phenoFile`). BOLT-REML currently only supports multi-trait analysis of traits phenotyped on a single set of individuals, so any individuals with at least one missing phenotype will be ignored. For *D* traits, BOLT-REML estimates *D* heritability parameters per variance component and *D(D-1)/2* correlations per variance component (including the residual variance component).

7.3 Initial variance parameter guesses

To specify a set of variance parameters at which to start REML iteration (which may save time compared to the default procedure used by BOLT-REML if you have good initial guesses), use `--remlGuessStr="`*string*`"` with the following format. For each variance component (starting with the residual term, which is automatically named `env/noise`), specify the name of the variance component followed by the initial guess. For instance, a model with two (non-residual) variance components named *vc1* and *vc2* (in the `--modelSnps` file) could have variance parameter guesses specified by:

```
--remlGuessStr="env/noise 0.5 vc1 0.2 vc2 0.3"
```

Note that the sum of the estimates must equal 1; BOLT-REML will automatically normalize the phenotype accordingly.

For multi-trait analysis of *D* traits, the `--remlGuessStr` needs to specify both guesses of *D* variance proportions and *D(D-1)/2* pairwise correlations per variance component. Viewing these values as entries of an upper-triangular matrix (with variance proportions on the diagonal and correlations above the diagonal), you should specify these *D(D+1)/2* values after each variance component name by reading them off left-to-right, top-to-bottom.

7.4 Trading a little accuracy for speed

BOLT-REML uses a Monte Carlo algorithm to increase REML optimization speed [12]. By default, BOLT-REML performs an initial optimization using 15 Monte Carlo trials and then refines parameter estimates using 100 Monte Carlo trials. If computational cost is a concern (or to perform exploratory analyses), you can skip the refinement step using the `--remlNoRefine` flag (in addition to the `--reml` flag). This option typically gives 2–3x speedup at the cost of ˜1.03x higher standard errors.

8 Polygenic prediction

The BOLT-LMM software also has a `--predBetasFile` option that computes SNP effect coefficients that can be used for polygenic prediction. The prediction model uses the same model selected for association testing (i.e., if the non-infinitesimal Gaussian mixture model is selected,

it will also be used for prediction). Reported SNP effect coefficients (`BETA`) are per-allele, with `ALLELE1` indicating the effect allele.

9 Output

9.1 BOLT-LMM association test statistics

BOLT-LMM association statistics are output in a tab-delimited `--statsFile` file with the following fields, one line per SNP:

- `SNP`: rs number or ID string
- `CHR`: chromosome
- `BP`: physical (base pair) position
- `GENPOS`: genetic position either from `bim` file or interpolated from genetic map
- `ALLELE1`: first allele in `bim` file (usually the minor allele), used as the effect allele
- `ALLELE0`: second allele in `bim` file, used as the reference allele
- `A1FREQ`: frequency of first allele
- `F_MISS`: fraction of individuals with missing genotype at this SNP
- `BETA`: effect size from BOLT-LMM approximation to infinitesimal mixed model
- `SE`: standard error of effect size
- `P_BOLT_LMM_INF`: infinitesimal mixed model association test p-value
- `P_BOLT_LMM`: non-infinitesimal mixed model association test p-value

Optional additional output. To output chi-square statistics for all association tests, set the `--verboseStats` flag.

9.2 BOLT-REML output and logging

BOLT-REML output (i.e., variance parameter estimates and standard errors) is simply printed to the terminal (`stdout`) when analysis finishes. Both BOLT-LMM and BOLT-REML write output to (`stdout` and `stderr`) as analysis proceeds; we recommend saving this output. If you wish to save this output while simultaneously viewing it on the command line, you may do so using

```
./bolt [... list of options ...] 2>&1 | tee output.log
```

10 Recommendations for analyzing N=500K UK Biobank data

Many users of BOLT-LMM wish to analyze UK Biobank data. Here are a few tips for computing association statistics on N=500K UK Biobank samples (see also ref. [10]):

- Computing genome-wide association test statistics for a single phenotype should typically take a few days to a week with multi-threading; we recommend using 8+ threads.
- The computation will require up to 100GB of memory depending on the number of directly genotyped SNPs included in the model (with `--bed`/`--bim`/`--fam`). If computational cost is a concern, running time and RAM can be reduced by specifying a subset of SNPs to use in the model with `--modelSnps` (e.g., by filtering on MAF or missingness or by LD-pruning).
- The model-fitting step of the analysis can also be sped up by including principal components as covariates (because projecting out top eigenvectors improves the conditioning of the kernel matrix). Note that in order to achieve this speedup, principal components must be computed using the sets of samples and SNPs used in the model (e.g., using `plink2 --pca approx` or EIGENSOFT v6.0+ `fastmode`). Pre-computed principal components computed on different sets of samples and/or SNPs (e.g., those provided by UK Biobank) tend not to provide much speedup.
- After model-fitting, BOLT-LMM computes association statistics on imputed SNPs. This computation is now multi-threaded for BGEN v1.2 data and should be fast enough to include all chromosomes in a single job, but parallelizing analyses across jobs analyzing subsets of chromosomes is of course allowable as well. Setting `--bgenMinMAF`/`--bgenMinINFO` will reduce output file size and improve speed.
- You will need to create a version of the `--fam` file that has numeric values in its 6th column, and you will also need to `--remove` the individuals in the plink data but not in the imputed data. (If thre is a mismatch, BOLT-LMM will generate a list of such individuals to `--remove`.)
- An example command line is below:

```
./bolt \
    --bed=ukb_cal_chr{1:22}_v2.bed \
    --bim=ukb_snp_chr{1:22}_v2.bim \
    --fam=ukb1404_cal_chr1_v2_CURRENT.fixCol6.fam \
    --remove=bolt.in_plink_but_not_imputed.FID_IID.976.txt \
    --remove=sampleQC/remove.nonWhite.FID_IID.txt \
    --exclude=snpQC/autosome_maf_lt_0.001.txt \
    --exclude=snpQC/autosome_missing_gt_0.1.txt \
    --phenoFile=ukb4777.phenotypes.tab \
    --phenoCol=height \
```

```
--covarFile=ukb4777.covars.tab.gz \
--covarCol=cov_ASSESS_CENTER \
--covarCol=cov_GENO_ARRAY \
--covarMaxLevels=30 \
--qCovarCol=cov_AGE \
--qCovarCol=cov_AGE_SQ \
--qCovarCol=PC{1:20} \
--LDscoresFile=tables/LDSCORE.1000G_EUR.tab.gz \
--geneticMapFile=tables/genetic_map_hg19.txt.gz \
--lmmForceNonInf \
--numThreads=8 \
--statsFile=bolt_460K_selfRepWhite.height.stats.gz \
--bgenFile=ukb_imp_chr{1:22}_v2.bgen \
--bgenMinMAF=1e-3 \
--bgenMinINFO=0.3 \
--sampleFile=ukb1404_imp_chr1_v2_s487406.sample \
--statsFileBgenSnps=bolt_460K_selfRepWhite.height.bgen.stats.gz
--verboseStats
```

10.1 UK Biobank v3 imputation release

The BGEN files in the UK Biobank v3 imputation release (7th March 2018) have the same format as the previous v2 release but now include files for chromosomes X and XY (= PAR1 + PAR2). These files are coded in the same BGEN sub-format (8-bit encoding, males coded as diploid) as the rest; however, they contain slightly fewer samples than the autosomal files (corresponding to the `in.Phasing.Input.chrX` and `in.Phasing.Input.chrXY` fields of the sample QC file).

If you wish to analyze both the autosomal and X chromosome data, you may do either of the following:

1. Run BOLT-LMM on all files at once by using the `--bgenSampleFileList` option to specify a list of whitespace-separated `.bgen` / `.sample` file pairs. Note that you will need to `--remove` all samples not present in **any** of the BGEN files. (If BOLT-LMM detects missing samples, it will report an error and write a list of such samples for you to `--remove`.)

2. Analyze the autosomal and chrX variants in two separate BOLT-LMM runs (using all autosomal and chrX typed variants in both runs as PLINK input for model-fitting).

The first approach is slightly more convenient than the second, at the expense of removing slightly more samples ($\sim$1000 additional samples) compared to the second.

11 Association analysis of case-control traits

While the mathematical derivations underlying BOLT-LMM are based on a quantitative trait model, BOLT-LMM can be also applied to analyze case-control traits (simply by treating the binary trait as a quantitative trait). However, an important caveat to be aware of is that BOLT-LMM test statistics can become miscalibrated for unbalanced case-control traits (resulting in false positive associations at rare SNPs), as noted in the SAIGE paper [19].

11.1 Guidelines for case-control balance

The extent to which BOLT-LMM P-values can suffer miscalibration for binary traits is a function of three variables: **sample size, minor allele frequency, and case-control ratio**. Specifically, miscalibration occurs when the minor allele count (MAC) multiplied by the case fraction is relatively small (corresponding to the conventional wisdom that chi-square test statistics break down when expected counts are small). (The same is true for P-values from any other regression-based chi-square statistic.) We also note that an analogous issue can arise when analyzing quantitative traits if a quantitative trait has extreme outliers.

In the revised version of our preprint exploring BOLT-LMM performance on UK Biobank N=500K data [10], we have included a suite of simulations that vary the three key parameters (sample size, minor allele frequency, and case fraction) that affect type I error control. **For analyses of the full UK Biobank data, we determined that for traits with a case fraction of at least 10%, BOLT-LMM test statistics are well-calibrated for SNPs with MAF>0.1%.** More extreme case-control imbalances can also be tolerated if the minimum MAF is increased. Full results of these simulations are presented in Supplementary Table 8 of ref. [10], which we recommend consulting to decide whether BOLT-LMM is appropriate for a particular binary trait analysis.

For highly unbalanced case-control settings in which BOLT-LMM analysis is inappropriate, we recommend using SAIGE [19] (which overcomes the problem of deviation from asymptotic normality by using a saddlepoint approximation).

11.2 Estimation of odds ratios

Because BOLT-LMM still uses a linear model (rather than a logistic model) when analyzing case-control traits, a transformation is required in order to convert SNP effect size estimates ("betas") on the quantitative scale to traditional odds ratios. A reasonable approximation is:

log OR = β / $(\mu * (1 - \mu))$, where μ = case fraction.

Standard errors of SNP effect size estimates should also be divided by $(\mu * (1 - \mu))$ when applying the above transformation to obtain log odds ratios.

Alternatively, a more sophisticated transformation is described here: `http://cnsgenomics.com/shiny/LMOR/`

12 Frequently asked questions

The most common question users ask is what to do when BOLT-LMM reports an error arising from a heritability estimate close to 0 or 1. Older versions of BOLT-LMM reported "ERROR: Invalid heritability estimate; cannot continue analysis"; newer versions attempt to clarify the issue:

- *"ERROR: Heritability estimate is close to 0; LMM may not correct confounding. Instead, use PC-corrected linear/logistic regression on unrelateds."*

  When a heritability estimate reaches 0, then linear mixed model association tests (including BOLT-LMM and other methods) all degenerate to simple linear regression, hence the error message. This situation is dangerous because the "mixed model" will no longer correct for population stratification and relatedness.

  You can of course still run an association test under these circumstances, but you will need to prune to an unrelated set of samples (if your sample set contains related individuals) and include principal component covariates.

  You can perform linear regression using BOLT-LMM by running it without the `--lmm` option (and with the `--verboseStats` option).

- *"ERROR: Heritability estimate is close to 1; algorithm may not converge. Analysis may be unsuitable due to low sample size or case ascertainment."*

  This error most frequently arises when sample size is low, resulting in estimated heritability having a very large standard error (perhaps even greater than 1) such that the estimate could be anywhere in the range 0 to 1 and might hit one of the boundaries. BOLT-LMM is not recommended for analyses of smaller samples; in this situation, we recommend trying other software packages such as GEMMA or GCTA.

13 Website and contact info

Software updates will be posted here and at the following website:

`http://www.hsph.harvard.edu/alkes-price/software/`

If you have comments or questions about the BOLT-LMM software, please contact Po-Ru Loh, `poruloh@broadinstitute.org`.

14 License

BOLT-LMM is free software under the GNU General Public License v3.0 (GPLv3).

References

1. Loh, P.-R. *et al.* Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics* **47**, 284–290 (2015).

2. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348–354 (2010).

3. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833–835 (2011).

4. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nature Methods* **9**, 525–526 (2012).

5. Listgarten, J., Lippert, C. & Heckerman, D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nature Genetics* **45**, 470–471 (2013).

6. Lippert, C. *et al.* The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific Reports* **3** (2013).

7. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821–824 (2012).

8. Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components-based method for whole-genome association analysis. *Nature Genetics* **44**, 1166–1170 (2012).

9. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* **46**, 100–106 (2014).

10. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nature Genetics* **50**, 906–908 (2018).

11. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* **88**, 76–82 (2011).

12. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance components analysis. *Nature Genetics* **47**, 1385–1392 (2015).

13. Johnson, S. G. The NLopt nonlinear-optimization package. URL `http://ab-initio.mit.edu/nlopt`.

14. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559–575 (2007).

15. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 1–16 (2015).

16. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genetics* **5**, e1000529 (2009).

17. Bulik-Sullivan, B. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291–295 (2015).

18. Galinsky, K. J. *et al.* Fast principal-component analysis reveals convergent evolution of *ADH1B* in Europe and East Asia. *American Journal of Human Genetics* **98**, 456–472 (2016).

19. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics* **50**, 1335–1341 (2018).