GWAS benchmarking

Katerina Zorina-Lichtenwalter
16 December 2020

Software overview

BOLT

- -linear mixed model
- -control for relatedness
- -allows for non-Gaussian distribution of SNP effect sizes
- -steps 1 (building GRM) and 2 (association analysis) are integrated

SAIGE

- -generalised mixed model (binary and quantitative phenotypes)
- -control for relatedness
- -saddle-point approximation (SPA) controls Type I error in rare traits (MAF < 0.3)
- -steps 1 and 2 are run separately
- -slower than BOLT

REGENIE

- -mixed model with linear and Firth logistic regression
- -control for Type I error and effect size inflation for binary phenotypes with low case prevalence
- -control for relatedness
- -steps 1 and 2 are run separately
- -computational efficiency:
    - -genetic matrix is partitioned into consecutive blocks of SNPs (low memory footprint) and ridge regressions are run to make pheno predictions on these blocks before they are combined
    - -parallel analysis of multiple phenotypes

Reported performance

Software	Sample size	# SNPs (step 1)	# SNPs (step 2)	CPUs	Time elapsed (hrs)	CPU hours	CPU frq (GHz)	Max. memory (GB)
BOLT-LMM[1]	480K	300K	300K	1	177	177*	2.27	34
SAIGE[2]	400K	200K	71M	20	26	517**		10
Regenie[3]	332-407K	469K	11.4M	16	94	777***	2.1	12.9

[1] Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, Patterson N. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature genetics. 2015 Mar;47(3):284.

[2] Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature genetics. 2018 Sep;50(9):1335-41.

[3] Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, Benner C, O'Dushlaine C, Barber M, Boutkov B, Habegger L. Computationally efficient whole genome regression for quantitative and binary traits. bioRxiv. 2020 Jan 1.

*benchmarking by Mbatchou et al. 2020. : Steps 1 and 2 take 39,806 for 50 binary traits

**benchmarking by Mbatchou et al. 2020. : Step 1 takes 275,070 CPU hours and Step 2 takes 94,347 CPU hours for 50 binary traits

***for 50 binary traits

BLANCA: reduced set

Input

Sample: 8K people
Step 1 (building GRM)
125K SNPs (bed/bim/fam)

Step 2 (association analysis)
1K SNPs on chrom 1 (bgen/sample)
Phenotype: face pain (h2 = 0.009)
Covariates: age, sex, PC1-10

Parameters

```
#SBATCH --qos=preemptable
#SBATCH --nodes=1
#SBATCH --mem=50gb**
#SBATCH --cpu-freq=2100
#SBATCH --time=1-00:00
#SBATCH --ntasks=36
```

**For Regenie, step 2, when reduced to 1gb, CPU time tripled

BLANCA: reduced set results

Method	CPU time	Time Elapsed	Max Memory (gb)
BOLT-LMM	01:37:48	00:02:43	0.48
SAIGE	Step 1: 14:06:00 Step 2: 00:09:36	Step 1: 00:23:30 Step 2: 00:00:16	Step 1: 0.68 Step 2: 0.26
REGENIE(-LOOCV)	Step 1: 10:11:24 **Step 2: 00:00:36	Step 1: 00:16:59 **Step 2: 00:00:01	Step 1: 2.2 Step 2: 0

**Regenie, --mem=1gb
CPU time: 00:01:12 Time elapsed: 00:00:02 MaxRSS: 0

BLANCA: whole-genome, all Whites

Input

Sample: 435K people
Step 1 (building GRM)
330K SNPs (bed/bim/fam)

Step 2 (association analysis)
7.7M SNPs (bgen/sample)
Phenotype: face pain (h2 = 0.009)
Covariates: age, sex, PC1-10

Parameters, BOLT and Regenie

```
#SBATCH --qos=preemptable
#SBATCH --nodes=1
#SBATCH --mem=50gb
#SBATCH --cpu-freq=2100
#SBATCH --time=1-00:00
#SBATCH --ntasks=36
```

Parameters (SAIGE), step 1

```
#SBATCH --qos=blanca-ibg
#SBATCH --nodes=1
#SBATCH --mem=300gb**
#SBATCH --cpu-freq=2100
#SBATCH --time=1-00:00
#SBATCH --ntasks=36
```

**also trying mem=200gb with LOCO=F

BLANCA: whole-genome, all Whites results

Method	CPU time	Time Elapsed	Max Memory (gb)
BOLT-LMM	11-08:39:00 (max)	07:34:25 (max)	39.12 (max)
SAIGE	Step 1: *running* Step 2: *tbd*	Step 1: *running* Step 2: *tbd*	Step 1: *running* Step 2: *tbd*
REGENIE(-LOOCV)	Step 1: 18-02:59:24 Step 2: 9-09:02:24 (max)	Step 1: 12:04:59 Step 2: 06:15:04 (max)	Step 1: 125.29 Step 2: 4.17 (max)

Summit, reduced set

Input

Sample: 8K people
Step 1 (building GRM)
125K SNPs (bed/bim/fam)

Step 2 (association analysis)
1K SNPs on chrom 1 (bgen/sample)
Phenotype: face pain (h2 = 0.009)
Covariates: age, sex, PC1-10

Parameters, BOLT

```
#SBATCH --partition=ssky-preemptable
#SBATCH --nodes=1
#SBATCH --mem=170gb
#SBATCH --cpu-freq=2500
#SBATCH --time=1-00:00
#SBATCH --ntasks=24
```

Parameters, step 1

```
#SBATCH --partition=shas-testing
#SBATCH --nodes=1
#SBATCH --mem=5gb
#SBATCH --cpu-freq=2500
#SBATCH --time=30:00
#SBATCH --ntasks=24
```

Parameters, step 2

```
#SBATCH --partition=shas
#SBATCH --nodes=1
#SBATCH --mem=5gb
#SBATCH --cpu-freq=2500
#SBATCH --time=1-00:00
#SBATCH --ntasks=24
```

Summit, results (reduced set)

Method	CPU time	Time Elapsed	Max Memory (gb)
BOLT-LMM	01:10:00	00:02:55	0.40
SAIGE	Step 1: 11:30:48 Step 2: 00:13:12	Step 1: 00:28:47 Step 2: 00:00:33	Step 1: 0.61 Step 2: 0.26
REGENIE(-LOOCV)	Step 1: 06:24:00 Step 2: 00:02:24	Step 1: 00:16:00 Step 2: 00:00:06	Step 1: 2.18 Step 2: 0

Compare to BLANCA

Method	CPU time	Time Elapsed	Max Memory (gb)
BOLT-LMM	01:37:48	00:02:43	0.48
SAIGE	Step 1: 14:06:00 Step 2: 00:09:36	Step 1: 00:23:30 Step 2: 00:00:16	Step 1: 0.68 Step 2: 0.26
REGENIE(-LOOCV)	Step 1: 10:11:24 **Step 2: 00:00:36	Step 1: 00:16:59 **Step 2: 00:00:01	Step 1: 2.2 Step 2: 0

Summit, whole genome, all Whites set results

Input

Sample: 435K people
Step 1 (building GRM)
330K SNPs (bed/bim/fam)

Step 2 (association analysis)
7.7M SNPs (bgen/sample)
Phenotype: face pain (h2 = 0.009)
Covariates: age, sex, PC1-10

Parameters, BOLT

```
#SBATCH --partition=ssky-preemptable
#SBATCH --nodes=1
#SBATCH --mem=170gb
#SBATCH --cpu-freq=2500
#SBATCH --time=1-00:00
#SBATCH --ntasks=24
```

Parameters (SAIGE, Regenie), step 1

```
#SBATCH --partition=ssky-preemptable
#SBATCH --nodes=1
#SBATCH --mem=170gb**
#SBATCH --cpu-freq=2500
#SBATCH --time=1-00:00
#SBATCH --ntasks=24
```

**Regenie failed with out-of-memory error when --mem=110gb

Parameters (SAIGE, Regenie, step 2

```
#SBATCH --partition=ssky-preemptable
#SBATCH --nodes=1
#SBATCH --mem=50gb
#SBATCH --cpu-freq=2500
#SBATCH --time=1-00:00
#SBATCH --ntasks=24
```

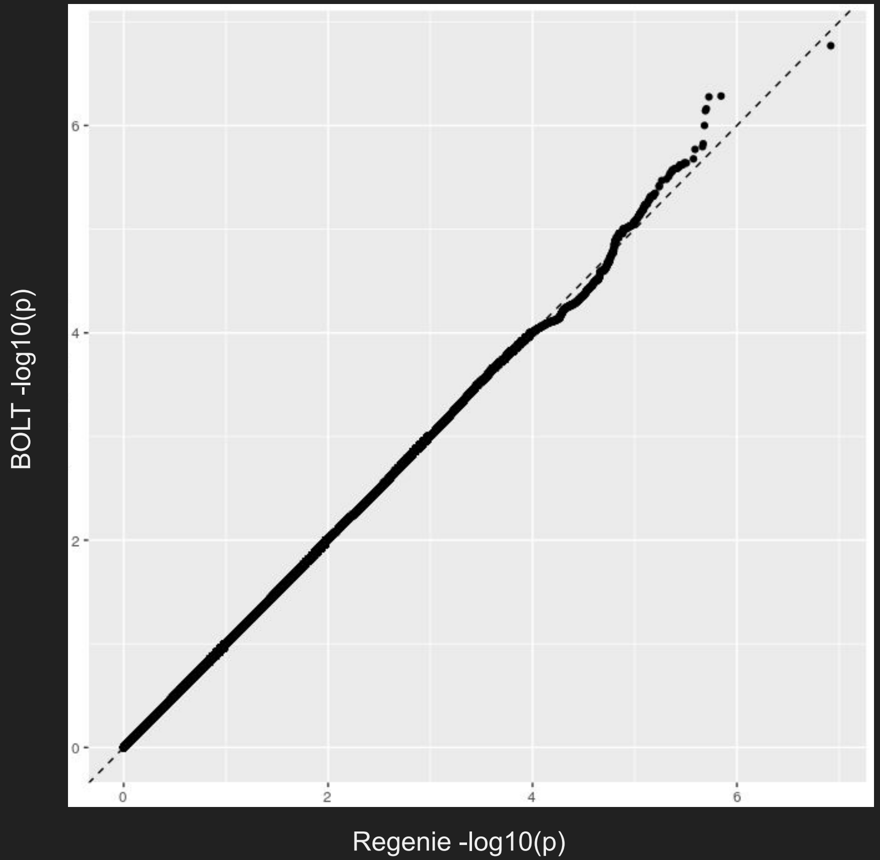
Summit results (whole-genome)

Method	CPU time	Time Elapsed	Max Memory (gb)
BOLT-LMM	6-21:53:12	06:54:43	39.07
SAIGE	Step 1: *running* Step 2: *tbd*	Step 1: *running* Step 2: *tbd*	Step 1: *running* Step 2: *tbd*
REGENIE(-LOOCV)	Step 1: 14-15:38:00 Step 2: 3-10:25:12 (max)	Step 1: 14:39:05 Step 2: 03:26:03 (max)	Step 1: 125.28 Step 2: 3.14 (max)

Compare to BLANCA

Method	CPU time	Time Elapsed	Max Memory (gb)
BOLT-LMM	11-08:39:00 (max)	07:34:25 (max)	39.12 (max)
SAIGE	Step 1: *running* Step 2: *tbd*	Step 1: *running* Step 2: *tbd*	Step 1: *running* Step 2: *tbd*
REGENIE(-LOOCV)	Step 1: 18-02:59:24 Step 2: 9-09:02:24 (max)	Step 1: 12:04:59 Step 2: 06:15:04 (max)	Step 1: 125.29 Step 2: 4.17 (max)

GWAS results comparison: BOLT and Regenie



Type I error trend:
BOLT outputs more near-significant p-values

...but this is just one trait, not enough to conclude anything

Summary

BOLT

-can run using preemptable queue on Blanca and Summit

-fastest GWAS method of these 3

-Type I error inflation with class imbalance reported for rare traits

SAIGE

-cannot run using preemptable queue on Blanca or Summit

-may run on blanca-ibg nodes with 7-day limit, if not using LOCO option

-LOCO option has not been extensively tested

-recommendation: wait for software development to stabilise or use BOLT for common traits (0.1 prevalence) and Regenie for rare traits (<0.1 prevalence)

Regenie

-can run using preemptable queue on blanca and summit

-slower than BOLT (but maybe faster if leave-one-out cross-validation is not used, LOOCV=F)

-Firth regression implemented for rare binary traits (advantage over BOLT)

-good for parallel multi-trait GWAS

-does not seem to have a low memory footprint as claimed in the paper!

General:

-jobs on Blanca appear to take what memory they need, regardless of --mem flag, whereas on Summit the --mem requested is a hard limit

Recommended GWAS parameters (blanca)

BOLT

```
#SBATCH --qos=preemptable
#SBATCH --nodes=1
#SBATCH --mem=170gb**
#SBATCH --ntasks=36
#SBATCH --time=1-00:00
```

**Minimum: 45gb

Regenie, step 1

```
#SBATCH --qos=preemptable
#SBATCH --nodes=1
#SBATCH --mem=130gb
#SBATCH --ntasks=36
#SBATCH --time=1-00:00
```

Regenie, step 2

```
#SBATCH --qos=preemptable
#SBATCH --nodes=1
#SBATCH --mem=50gb**
#SBATCH --cpu-freq=2500
#SBATCH --time=1-00:00
#SBATCH --ntasks=24
```

**Minimum: 5gb

Recommended parameters (summit)

BOLT

```
#SBATCH --partition=ssky-preemptable
#SBATCH --nodes=1
#SBATCH --mem=170gb**
#SBATCH --time=1-00:00
#SBATCH --ntasks=24
```

**Minimum: 45gb

Regenie, step 1

```
#SBATCH --partition=ssky-preemptable
#SBATCH --nodes=1
#SBATCH --mem=130gb
#SBATCH --time=1-00:00
#SBATCH --ntasks=24
```

Regenie, step 2

```
#SBATCH --partition=ssky-preemptable
#SBATCH --nodes=1
#SBATCH --mem=50gb**
#SBATCH --time=1-00:00
#SBATCH --ntasks=24
```

**Minimum: 5gb