New strategies toward scaling up epistasis analysis on large-scale genomic datasets

Jia Wen, Colby T. Ford, Daniel Janies and Xinghua Shi

Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte

Introduction

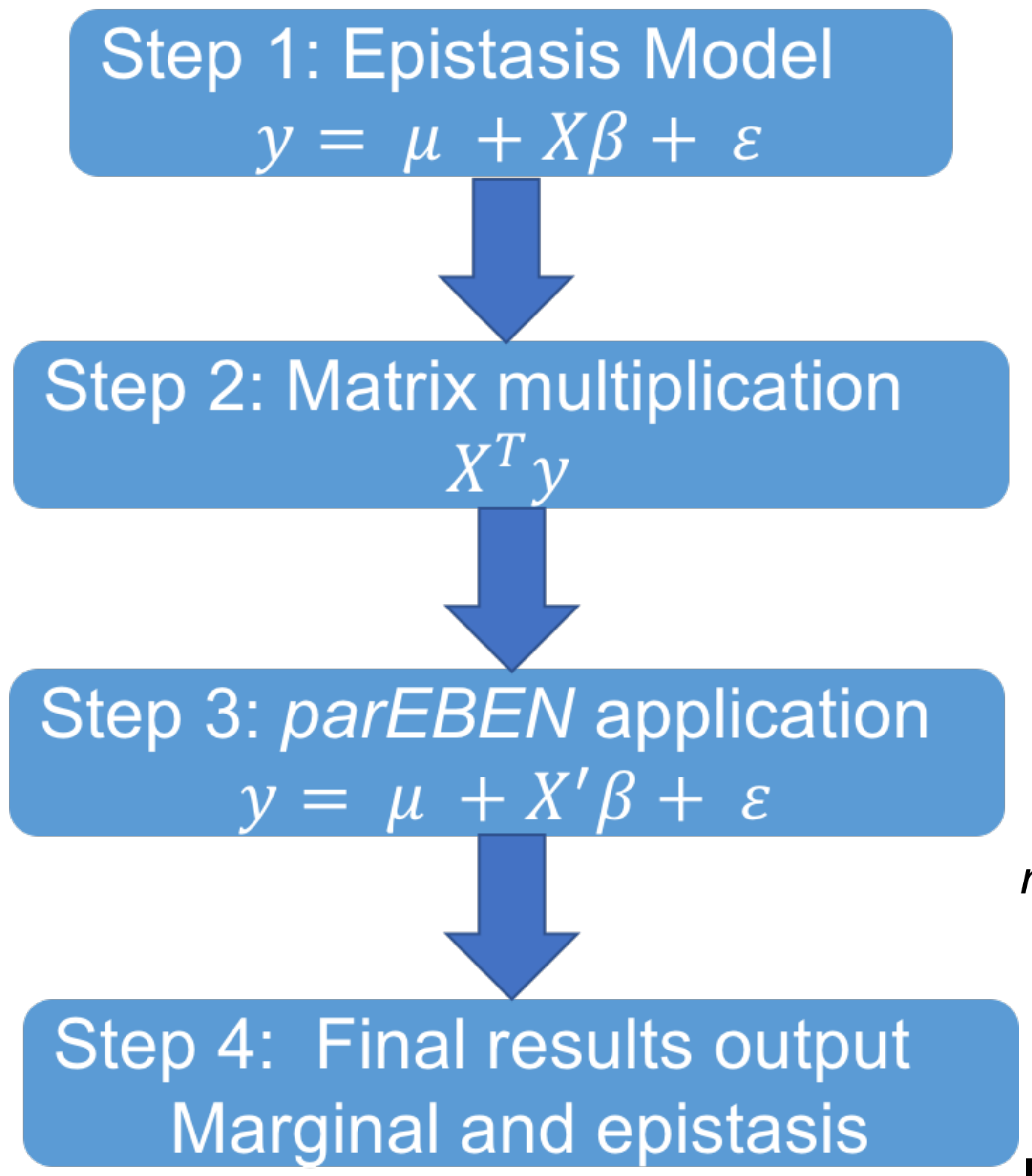
Epistasis reflects the joint effect of more than one gene or genetic variants on a particular phenotype. Identifying epistasis on large-scale genomic data is challenging. The challenge lies in two interlinked constraints that the over-saturated model demands efficient statistical methods to fit the model and the intensive computing work to solve these statistical methods for identifying both marginal and epistasis effects. Here, we develop new strategies to scale up and accelerate epistasis analysis by pre-computing and parallelization. Specifically, we first introduce a matrix strategy which pre-computes the correlation matrix using matrix multiplication to filter all the features. We then develop a parallelized version of the Empirical Bayesian Elastic Nets (*EBEN)* method [2] for faster association analysis. Test simulations and real data analysis using a yeast fitness-related phenotype, we have demonstrated the appealing performance of these strategies.

Step 1: Construct the epistasis model including the marginal and epistasis effects;

Step 2: Apply a matrix multiplication strategy to pre-compute the correlations between features and dependent variable, followed by filtering features with low correlations with the dependent variable;

Step 3: Construct the reduced model using the filtered features and apply *parEBEN* to solve the reduced model.

Step 4: Output the significant marginal and epistasis features.



Step 1: Epistasis Model
$y = \mu + X\beta + \varepsilon$

Step 2: Matrix multiplication
$X^T y$

Step 3: *parEBEN* application
$y = \mu + X'\beta + \varepsilon$

Step 4: Final results output
Marginal and epistasis

Fig. 1. Our epistasis analysis strategies based on sparse parameter models and *parEBEN*.

Data and Methods

The real yeast SNP dataset used for simulations is obtained from [1], including 4,390 samples and 28,220 SNP.

The full model is:

$$y = \mu + X\beta_m + X_iX_j\beta_e + \varepsilon$$

The details of algorithm are shown in Fig. 1. y is the phenotype, μ is the population mean, X is the genetic marker matrix with sample n and dimension k, and X_i and X_j are two SNP vectors. β_m is the coefficient that represents the marginal effect of genetic variant, β_e is the coefficient that represents the epistatic effect between genetic variants i and j, and ε is residual error. In order to illustrate our work flow. We can rewrite the model as a simplified format which is used in the above workflow (Fig.1),

$$y = \mu + X\beta + \varepsilon$$

Where $X = c(X, X_iX_j), \beta = c(\beta_m, \beta_e)$.

Real data-based simulation

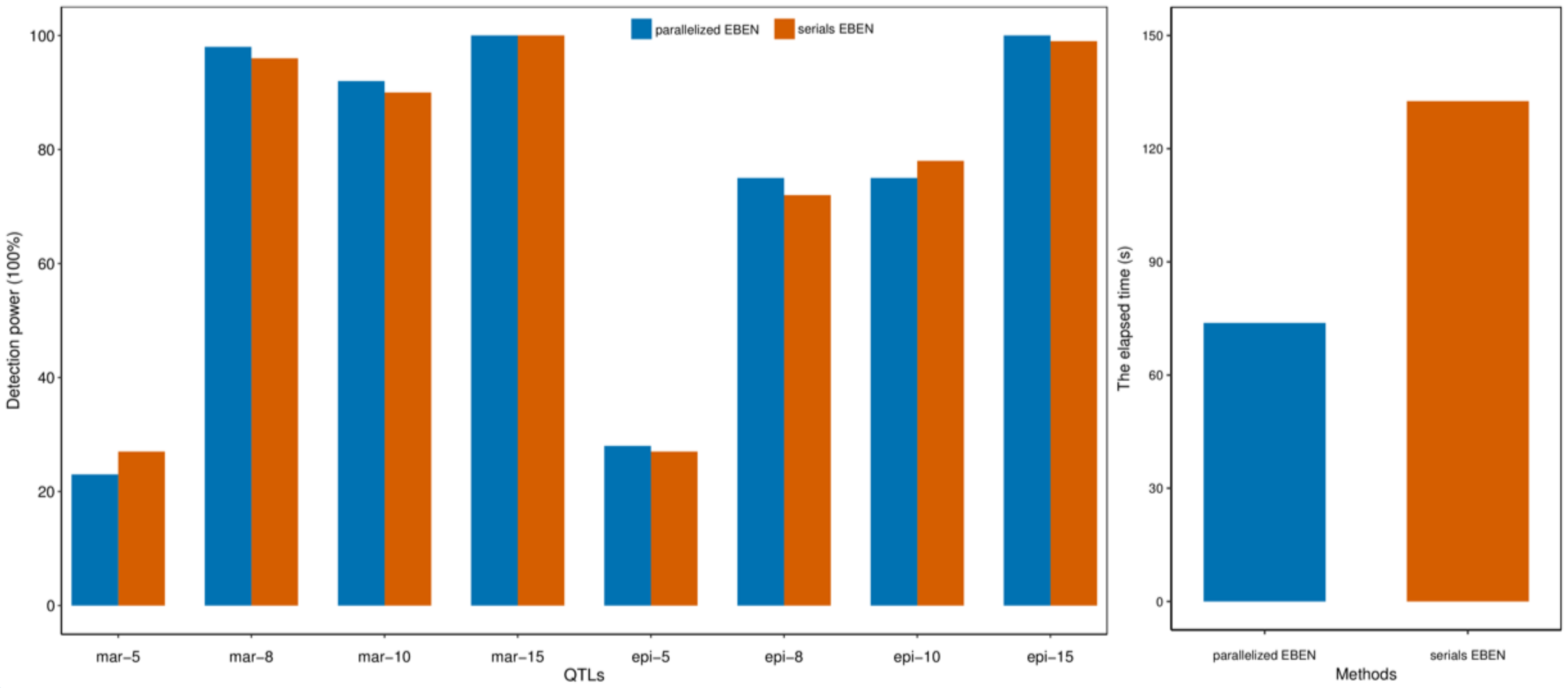


Fig. 2A. The detection power of marginal and epistatic effects with different heritabilities and elapsed time between EBEN and parEBEN (sim I).

The simulations show that a drastic time reduction in most cases can be seen on the resulting EBEN models by parallelizing the iterations of the cross-validation over multiple CPU cores or multiple machines of a computing cluster (Fig. 2).

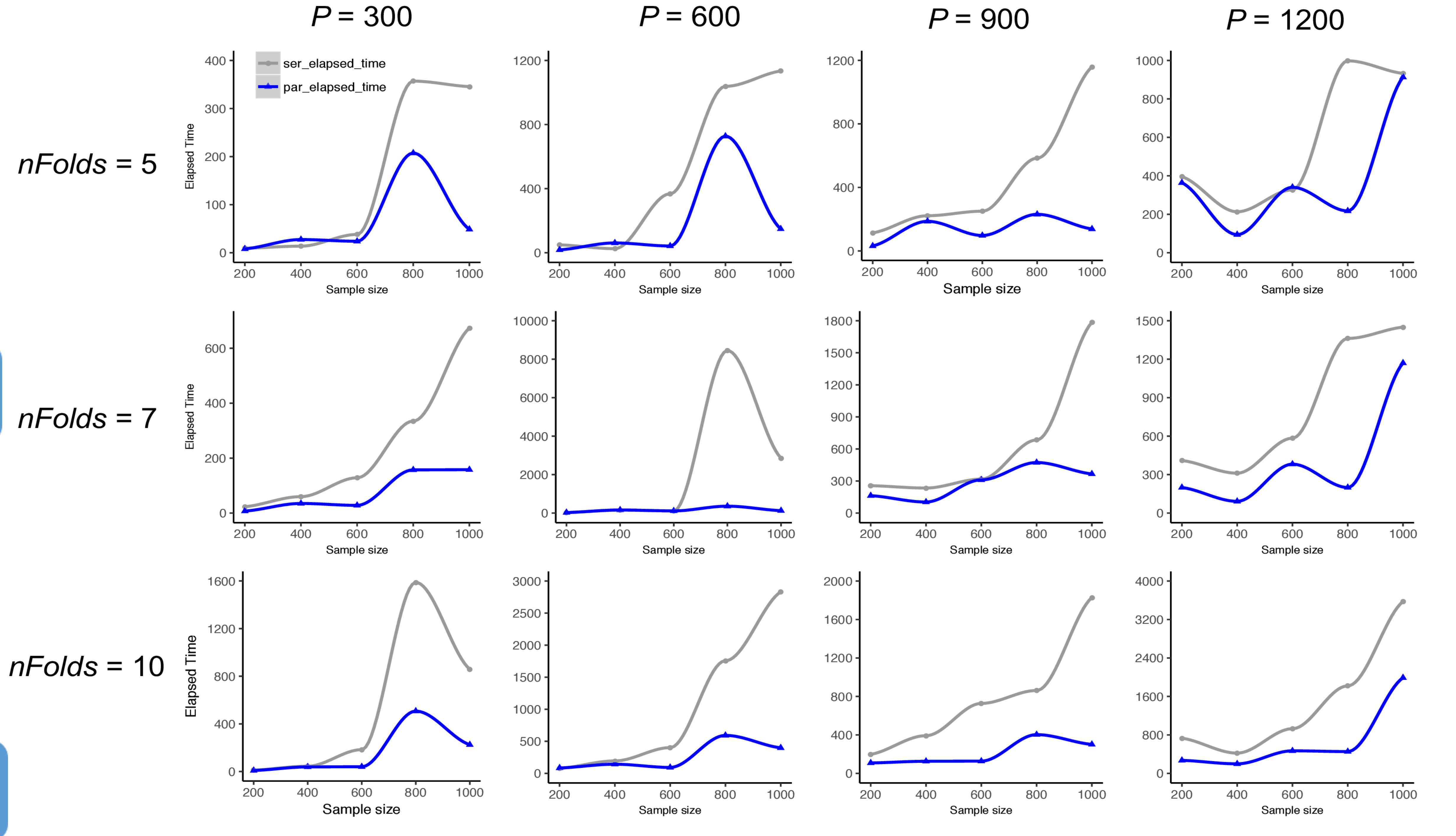


Fig. 2B. The elapsed time comparison between serials EBEN and parallelized EBEN (sim II). Three fold numbers 5, 7 and 10 are corresponding to each row, and four sample sizes 300, 600, 900 and 1200 are corresponding to each column. The grey lines denote the elapsed time from serials EBEN; the blue lines denote the elapsed time from *parEBEN*.

Empirical yeast data analysis

Findings from the yeast data analysis [1]:

- **SNP chrIV: 998,628** with marginal effect [3]
- **SNP chrVIII:114,144** with both marginal effect and epistatic effects with chrIII:191,932
  - Reported to be related with indoleacetic acid level in previous study [1,3]
  - Maps to gene *GPA1* can affect the yeast response to mating pheromone, whose corresponding pathway of yeast pheromone response further affects the fitness of indole acetic acid [1, 3]
- **SNP chrIII:191,932** is in high LD with **SNP chrIII:198,615**
  - Reported to be as radical-QTL and co-regulate the indoleacetic acid (IAA-network) with the hub-QTL chrVIII:114,144 [3]

Conclusion

To accelerate epistasis analysis, a specialized matrix multiplication strategy is developed along with the *parEBEN* package. By reducing the computation time, regression models on larger, more complex data can be completed without such a delay.

This study also provides an opportunity for larger datasets to be analyzed as opposed to limiting the research due to time and computing resource constraints. Thus, parallelizing the cross-validation and hyperparameter sweep of EBEN models will prove to be greatly beneficial in future research using cross-validated Bayesian elastic nets.

The *parEBEN* package and relevant data in this study is available at https://github.com/colbyford/parEBEN.

References:

[1] Bloom, J. S., et al. (2015). Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nature communications*, 6, 8712.

[2] Huang, A., et al. (2015). Empirical Bayesian elastic net for multiple quantitative trait locus mapping. Heredity, 114(1), 107.

[3] Forsberg, S. K., et al. (2016). Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. bioRxiv, 059485

Supported by NIH R15HG009565