GenAMap on the Web: Intuitive and Scalable Machine Learning for Structured Association Mapping

Benjamin J. Lengerich*, Haohan Wang*, Min Kyung Lee1, Eric P. Xing1,2,4
Carnegie Mellon University: (1) Computational Biology Department, (2) Language Technologies Institute, (3) Center for Machine Learning and Health, (4) Machine Learning Department

Abstract
The proliferation of genomic data has increased the usefulness of complex machine learning algorithms for structured association mapping. Such methods can effectively relate genetic polymorphisms with phenotypes, but correct use requires algorithmic expertise to run code and domain expertise to analyze results. To overcome these challenges, the GenAMap software platform was developed and released in 2010[1]. Since then, the sizes of available biological data have continued to increase exponentially. To address this challenge, GenMap is redesigned for scalability and updated with state-of-the-art methods for efficient calculations on human genome-scale data. The user experience is enhanced as an intuitive web application with a focus on simplicity and ease of use. GenAMap is available at http://sailing.cs.cmu/main/genamap, with source code at http://github.com/blengerich/genamap.

State-of-the-Art Genomic Software
Association mapping techniques have evolved over the past decade from simple genome-wide association studies (GWAS) to complex machine learning algorithms. While GWAS has success in discovering single nucleotide polymorphisms of interest, the polygenic variability in complex traits is more often masked by a mixture of variance. To capture this variability, a number of methods have been proposed, and corresponding software packages have been offered to enhance the use of these algorithms. Here we briefly review some of the most widely used association mapping software toolkits.

Plink[2] is a free, open-source whole genome association analysis toolset introduced in 2007. It offers a range of basic, large-scale analyses in a computationally efficient manner. Specifically, it highlights:
- Summary Statistics: such as genotype rates, allele and genotype frequencies, Hardy-Weinberg equilibrium tests, and single SNP Mendelian error summaries.
- Confounding Analysis offers methods to correct confounding effects due to stratification and nonrandom genotyping failure.
- Association Analysis provides a variety of standard statistical tests, including case/control allele test, Cochran-Armitage trend test, Fisher’s exact test, Cochran-Mantel-Haenszel tests, and others.
- Identical By Descent (IBD) Estimation provides options to estimate genome-wide IBD-sharing coefficients between seemingly unrelated individuals from whole genome data. While PLINK excels at statistical standard qualities, it lacks machine learning methods for structured association mapping.

Matkap[3] which stands for Marker-Trait Association Platform And Explorer is an online platform designed to associate genotypes with traits to perform for key organisms. It was built with the belief that previous approaches for controlling population and family structure. For result interpretation, the program allows for linkage disequilibrium statistics to be analyzed.

SNPTEST application which requires programming expertise to use.
- Increase the size of datasets by analyzing different populations together
- Key Assumption: Same phenotype across different populations is likely to be affected by the same genotypes
- Solves the optimization problem:

\[
\text{arg min} \sum (y_i - X\beta)^2 + \lambda\beta
\]

Adaptive Multi-task Lasso
- Key Assumption: Prior knowledge of SNP influences can help the model understand the importance of SNP for structured association mapping.
- Knowledge of SNPs is scaled by a factor that is automatically learned.
- Solves the optimization problem:

\[
\text{arg min} \sum (y_i - X\beta)^2 + \lambda\beta
\]

Graph-guided fused Lasso
- Uses prior knowledge of relatedness of phenotypes
- Key Assumption: Similar phenotypes are affected by the same genotypes
- Solves the optimization problem:

\[
\text{arg min} \sum (y_i - X\beta)^2 + \lambda\beta
\]

Structured Association Mapping
Motivation
While standard GWAS methods are sufficient to discover individual SNP relationships, phenotypes of interest are often caused by a complex interaction of changes. Structured association mapping learns a joint model that considers the effect of all SNPs. This allows these methods to consider the interactions between SNPs. Furthermore, structured association methods can effectively incorporate prior knowledge about the structure of the data to increase accuracy.

Structured Association Mapping
In GenAMap
- Lasso
  - Key assumption: there is only a small subset of the SNPs that are responsible for the phenotype of interest
  - Solves the optimization problem:

\[
\text{arg min} (y_i - X\beta)^2 + \lambda\beta
\]
- Tree Lasso
  - Uses the prior knowledge that phenotypes have a cluster structure which can be represented as a tree with leaf nodes as output and internal nodes as clusters of the outputs at multiple granularities
  - Solves the optimization problem:

\[
\text{arg min} (y_i - X\beta)^2 + \lambda\beta
\]
- Multi-Population Lasso
  - Increase the size of datasets by analyzing different populations together
  - Key Assumption: Same phenotype across different populations is likely to be affected by the same genotypes
  - Solves the optimization problem:

\[
\text{arg min} (y_i - X\beta)^2 + \lambda\beta
\]
- Adaptive Multi-task Lasso
  - Key Assumption: Prior knowledge of SNP influences can help the model understand the importance of SNP for structured association mapping.
  - Knowledge of SNPs is scaled by a factor that is automatically learned.
  - Solves the optimization problem:

\[
\text{arg min} (y_i - X\beta)^2 + \lambda\beta
\]

- Graph-guided fused Lasso
  - Uses prior knowledge of relatedness of phenotypes
  - Key Assumption: Similar phenotypes are affected by the same genotypes
  - Solves the optimization problem:

\[
\text{arg min} (y_i - X\beta)^2 + \lambda\beta
\]

- Performance

![Performance](image)

GenAMap automates algorithms for you. It collects analysis requests from the frontend, and schedules them on a distributed backend. The parallelization enables the machine learning algorithms to scale to large datasets, including human genome.

Interactive Visualizations
GenAMap offers a number of interactive visualizations for exploratory and summary analysis. Current visualizations include:
- Manhattan Plot
- Matrix View
- Dendrogram

Visualizations in the desktop version include:
- Transcriptional Regulatory Network Visualization
- Co-expression Network Visualization
- Population Segmentation

Data Management
GenAMap organizes data into projects, which contain marker (SNP) and trait data. Data can either be uploaded from your computer, or linked directly from the NIH Genome Data Commons.

Conclusions
While many genomics platform offer simple association mapping, the graphical interfaces fall short for complex traits. GenAMap has been constructed to fill this gap, allowing biologists without programming expertise to run machine learning methods and visualize the results.

- Sign up for GenAMap updates at sailing.cs.cmu/main/genamap.
- More information and server available at sailing.cs.cmu.edu/main/genamap.
- Please use our server for small jobs.
- Please use our code to set up your own server for large analyses.
- Open source at github.com/blengerich/genamap.

Features of GenAMap
- Algorithm Automation
- State-of-the-Art Genomic Software
- Structured Association Mapping
- Interactive Visualizations
- Data Management

References
2. Paschal, S. B. Expectation, the procedure for linkage disequilibrium statistics to be calculated and visualized graphically. Other features include analyzing models, calculating diversity statistics, integrating phenotypic and genotypic data, imputing missing, and calculating principal components. Unfortunately, Tassel is only available as a standalone desktop application which requires programming expertise to use.
3. SNPTEST performs analysis of individual SNP associations in GWAS. The tools implemented include Binary or multiple quantitative phenotypes, Bayesian and frequentist tests, and conditioning upon covariates. Unfortunately, the only interface is a command line, and there is no visualization. Furthermore, the utility of analyzing individual SNPs decreases when examining complex traits.

Features of GenAMap
- Algorithm Automation
- State-of-the-Art Genomic Software
- Structured Association Mapping
- Interactive Visualizations
- Data Management

References
2. Paschal, S. B. Expectation, the procedure for linkage disequilibrium statistics to be calculated and visualized graphically. Other features include analyzing models, calculating diversity statistics, integrating phenotypic and genotypic data, imputing missing, and calculating principal components. Unfortunately, Tassel is only available as a standalone desktop application which requires programming expertise to use.
3. SNPTEST performs analysis of individual SNP associations in GWAS. The tools implemented include Binary or multiple quantitative phenotypes, Bayesian and frequentist tests, and conditioning upon covariates. Unfortunately, the only interface is a command line, and there is no visualization. Furthermore, the utility of analyzing individual SNPs decreases when examining complex traits.

Acknowledgements
Thanks to Jingkang Kang, Beilin Li, Dylan Steele, Bryan Yan, Danielle Hu, Clarence Ngan, and Joey Gibelli for software development.

This work is funded by: