GenAMap

Feature Overview

What it does

- GWAS … but without the hassle
- Data Management
- Algorithm Execution
- Visualization

Overview (and some definitions)

- GenAMap stores instances of Data Types
- GenAMap allows users to load up instances of these data types
- GenAMap can create new instances of data types with Algorithms, using other data type instances as inputs
- GenAMap displays the data using Visualizations

Key Data Types

- Marker Data
- Trait Data
- Network Data
- Association Data

Marker Data

- Takes the form of an arbitrary data matrix
- Supposed to represent SNP values (0, 1 or 2)
- Represents the X matrix in e.g. a Lasso-based GWAS
- This data must be loaded by the user into GenAMap

Trait Data

- Takes the form of an arbitrary data matrix
- Supposed to represent either gene expression or phenotype data
- Represents the Y matrix in e.g. a Lasso-based GWAS
- This data must be loaded by the user into GenAMap
- This, together with the Marker Data, forms the basis of most analyses

Network Data

- Represents the associations between traits in a Trait Data set
- Takes the form of a sparse square matrix with dimension equal to that of the number of traits in the corresponding Trait Data set
- This can be loaded into GenAMap or created by one one of the algorithms that generate a Network Data set based on a Trait Data set (“Network Algorithms”)

Association Data

- Represents the associations between markers in a Marker Data set and traits in the Trait Data set
- Represents the beta coefficients that are the output of e.g. a Lasso-based GWAS
- Takes the form of a sparse d * t matrix with d the number of markers in a Marker Data set and t the number of traits in the Trait Data set
- This can be loaded into GenAMap or created by one of the algorithms that output an Association Data set ("Association Algorithms")

Key Data Types

Name	Description	Format	Actions
Marker Data	SNP values of samples	N x d matrix	load
Trait Data	Trait (e.g. phenotype / gene expression) values of samples	N x t matrix	load
Network Data	Represents the associations between traits in a single Trait Data set	d x d (sparse) matrix	load & create
Association Data	Represents associations between markers and traits	t x d (sparse) matrix	load & create

Network Algorithms

- Network Algorithms use a Trait Data set to find between-trait associations
- We must load up a Trait Data set before we can run a network algorithm

Network Algorithms

Name	Description
Correlation	Computes pair-wise correlation between traits
Correlation Squared	Computes pair-wise correlation squared between traits
Scale-free network (SFN)	Computes Scale-free network according to [1]
Topological Overlap Matrix (TOM)	Computes Topological Overlap Matrix according to [1]
Graphical Lasso (Glasso)	Fits an L1-penalized Gaussian Graphical Model

Association Algorithms

- Association Algorithms use a Marker Data set and a Trait Data set to find associations between traits and markers (i.e. regular GWAS)
- Some also use instances of other data types as side information. For example, the graph-guided fused Lasso uses Network Data to define its fusion penalty (hence the name …)
- We must load up a Marker Data set, a Trait Data set and load / create the side information before we can run these algorithms

Association Algorithms

Name	Description	Inputs
Lasso	Unstructured L1-penalized regression	Marker Data, Trait Data
PLINK	Wald test (qualitative traits) or chi-squared test (binary traits) as implemented by PLINK [2]	Marker, Trait
Tree-guided Lasso (TreeLasso)	Tree-guided penalty on output groups	Marker, Trait, Network Data
Graph-guided fused Lasso (Gflasso)	Fusion penalty on correlated output variables	Marker, Trait, Network
Graph-constrained fused Lasso (GcLasso)	Fusion penalty on correlated output variables	Marker, Trait, Network
Adaptive Multi-Task Lasso (AMTL)	Lasso that reweights individual betas based on SNP features	Marker, Trait, Network, Feature Data
Multi-population group Lasso (MPGL)	Lasso that allows parameters to vary between populations	Marker, Trait, Pop. Structure
PopAnal	Performs 4 analyses by population: single-SNP cross-validation, PLINK (see above), a likelihood test [3] and a t-test	Marker, Trait, Pop. Structure

Key Visualization Tools

- Chromosome Browser
  - Shows the location of SNPs along the chromosome
- Heat Map
  - Network / Association Data are viewed as a matrix of colored pixels
- JUNG view
  - Small sets of traits are shown as a ball and stick representation
- Manhatten Plot
  - Associations of specific traits are plotted against the genome

Data Architecture

- Data from all users is stored on COGITO
- GenAMap keeps a local copy of all your data
- It looks for the data in a folder called “data” in the startup directory
- When you log in, the GUI will query COGITO for the names of all your data sets / files.
- It will check whether it can find this data locally and download everything it can’t find
- Local Data is used to initialize visualizations

Running Algorithms

- Algorithms can be started from the GUI
- The GUI sends a message to COGITO
- Back-end of GenAMap instantiates Condor jobs to compute result
- When those jobs finish, the result is stored on Cogito and downloaded to the “data” folder

Auxiliary Data Types & Algorithms

- Beyond the core data types, GenAMap allows the loading / creation of a range of other data types, each of which has its own particular use
- Most of these data types have at least one algorithm which can be used to create instances of them from instances of other data types

Auxiliary Data Types & Algorithms

Name	Description	Format	Actions
Population Structure	Clustering of samples + list of dominant Eigenvalues	complex …	load & create
Features	SNP Features used for AMTL algorithm	d * f matrix	load
Tree	Tree-structure over traits	Linked List	load & create
Clustering	Correlation-aware ordering of traits	Permutation of {1,..,t}	load & create
Subset	Subset of traits	Subset of {1,..,t}	load & create
Module	Subset with GO annotation	Subset of {1,..,t}	create
3-way association	More later …	More later …	create

Name	Description	Output	Inputs
Structure	Standard algorithm for creating population structure	Population Structure	Marker Data
Agglomerate Hierarchical Clustering	Generates a tree structure for a trait set	Tree	Trait Data, Network
Hierarchical Clustering	Orders a trait set so that similar traits are in proximity	Clustering	Trait Data, Network
Top k connected traits	Finds the most important traits	Subset	Trait Data
Top k connected traits with neighbors	Finds the most important traits	Subset	Trait Data
Module Analysis	Greedy Dynamic Programming Algorithm for finding tightly connected trait clusters	Modules	Trait Data, Network, Clustering, Associations
gGFlasso	More later …	3-way association	More later …

3-way analysis

- The “crown jewel” of GenAMap is the 3-way visualization tool
- It is a very flexible combination of a JUNG view for gene-to-phenotype associations as well as a Manhatten plot for SNP-to-gene association
- Allows users to develop hypotheses for “3-way associations”, i.e. pairs of associations between (SNPs, genes) and (genes, phenotypes), where the gene component is equal

gGFlasso

- 3-way association data can only be created through the gGFlasso algorithm
- It uses two Trait Data sets (one representing genes, one phenotypes), two Networks (one for each Trait Data set) and an Association Data set for the genes (representing pre-computed SNP-gene associations)
- It regresses phenotypes on genes using a fusion penalty induced by the networks

Online Tools

- GenAMap allows the user to automatically query for SNPs / genes in certain web-based databases
- Clicking at a certain place in GenAMap will open up a web browser with the search result page of the respective database
- GenAMap needs to know the name of the requested SNP / gene in a format that the respective database can consume it
- Names are given as column headings when uploading a Marker Data set / Trait Data set

Online Tools

Database	Queried Object	Use Requirements	How to Access
dbSNP	SNP	SNP names must be rs#	Right-click on single SNP in chromosome browser
SGD	SNP	SNP names must have the names of the genes they are on and SGD must be able to consume these names	Right-click on single SNP in chromosome browser
UniProt	Gene	Trait names must be gene names that UniProt can consume	Right-click on Trait label in JUNG view
Google	Gene	Trait names must be such that you expect sensible results from Google	Right-click on Traits label in JUNG view

GO analysis

- GenAMap contains GO ontologies
- They were downloaded a few years ago and saved as a file into the GenAMap distribution (i.e. they don't update automatically)
- GenAMap allows users to perform GO enrichment analysis in various places
- Trait names must be following one of the conventions that are recognized in these files (check the GO files if you are not sure)

References

[1] Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology 4: article 17

[2] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81: 559–575

[3] Wu T, Chen Y, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25: 714–721.