Camelot (CAusal Modeling with Expression Linkage for cOmplex Traits) provides a framework to integrate genotype and gene expression data to model and predict phenotype. Camelot aims to identify potential causal factors to explain the phenotype and therefore achieve accurate prediction for the trait. Briefly, Camelot builds linear regression model for each trait, using genotype and gene expression data as features. The selection of genotype and/or transcript predictors is based powered by regularized regression, bootstrapping and a causality test (called triangle test).

In addition, Camelot also provides a function (zoom-in score) to prioritize the causal potential of genes residing within a linked locus. Because a locus can span over a large chromosomal region, many genes can reside in a linked locus. To facilitate identification of causal gene/causal allele, Camelot calculates a Bayesian-based score (zoom-in score) to rank genes residing in a region.

For more details, please see Chen BJ., Causton H.C., Mancenido D., Goddard N.L., Perlstein E.O., Pe'er D. Harnessing gene expression to identify the genetic basis of drug resistance. Mol Syst Biol. 2009;5:310. Epub 2009 Oct 13.

For any questions, please contact camelot.program@gmail.com.

You will need Matlab software to run Camelot. This package provides a main function (camelot_main.m) to assess all the components of Camelot. It also contains a modified version of laren code, which was originally implemented by Karl Sjöstrand (kas@imm.dtu.dk).

Prepare_structure_script.m will help users build a few main data structures that are needed to run Camelot. Camelot_main.m provides an interface to run Camelot. Please notice that the number of permutation testing and number of bootstrap sampling will affect the time needed for running the program. For large-scale data, you might want to divide your computation loads into several machines.

Camelot_main.m starts by setting parameters needed for all components of Camelot. Below is the description of each parameter.

General

dataset	Matlab data file name
algo	Regression algorithm. Users can choose ‘lasso’ and/or ‘elastic’ (elastic net)
featureset	Features used to build regression model. It can be genotype (‘G’) and/or a set of gene expression (‘R’). When ‘R’ is specified, dataset must contain a ‘regulators’ field to specify the indexes of transcripts in dataset.expression. Users may also use ‘E’ to use all the expression data.
numphenocluster	Number of clusters for the phenotype data. This is used to cluster the phenotype data (kmeans) when

users do not specify clustering of the phenotype data.

numexpcluster
Number of clusters for the gene expression data. This is used to cluster the gene expression data (kmeans) when users do not specify clustering of the expression data.

regulatorvalid_stdthres
Filter for gene expression data when a set of regulators ('R') is specified. Only those with standard deviation >= regulatorvalid_stdthres will be considered as features for regression model.

expvalid_stdthres
Filter for gene expression data when all expression ('E') is used for features in regression model. Only those with standard deviation >= expvalid_stdthres will be considered as features.

btthres
Threshold for selecting confident features. When a feature is chosen with a frequency >= this threshold during bootstrapping, the feature will be then chosen.

savebt
If true, the results from bootstrapping are saved.

paraportion
Portion of data used for selecting parameters for regression. Default: 2/3 of data.

paracv
Number of cross-validation used to select parameters.

cv
Number of cross-validation used to assess the models.

bt
Number of bootstrapping runs.

seed
Seed number for randomization. Used in bootstrapping and cross-validation. Keep this number constant to reproduce results.

Triangle test

tri_doset
This should be a subset of 'featureset' and it should involve expression data. For example, if featureset is {'G', 'RG', 'EG'}, then only 'RG' and 'EG' are available here. This set tells Camelot to run triangle test on the transcripts that are chosen during regression modeling, to establish the significance of the transcript's causal role.

tri_mergemarkercorthres
Merge genotypes that share correlation coefficient >= tri_mergemarkercorthres during triangle testing.

tri_mergemarkerindexthres
Merge genotypes that are closed by (approximated by the indexes), assuming genotype data are sorted

relatively to chromosomal locations. Combination of this threshold and tri_mergemarkercorthres allows users to merge near-by and highly correlated genotype data during triangle test.

tri_expset Gene expression dataset, Matlab file.

tri_numperm Number of permutations for triangle test.

tri_FDR FDR control for triangle test.

tri_edgepvalthres p-value cutoff to define each triplet of genotype, transcript and phenotype for triangle test. Triangle test is only applied to the triplet only when all three edges pass this threshold.

tri_loosemergemarkercorthres A Loose threshold of correlation coefficient for merging genotype during triangle testing. This threshold should be lower than tri_mergemarkercorthres for more stringent testing for a transcript's causal role.

Model revision

rev_baseset The feature set used to build the basis of model. Default: 'G' (genotype).

rev_refset The feature set used to revise the model. Default: 'RG' (expression and genotype are used).

rev_minbtthres A relaxed threshold for bootstrapping frequency. This is to allow upstream causal genotype to enter the final model.

Zoom-in

zoom_algo Algorithms used to detect association. Default: 'elastic' and 'QTL' for elastic net regression and ranksum test.

zoom_doset Feature sets used for detecting association.

zoom_window Size of window used to expand the associated locus. Default: 30000 bps.

zoom_genelocdb Data file that contains location of genes.

zoom_detailmarker Data file that contains genotype data.

zoom_markerset Field in the strucutre (from zoom_detailmarker) that contains the genotype data.

zoom_consv Data file that contains conservation score among species.

zoom_knowngene Known causal genes. Specify any known causal genes will allow them to have high prior probability as a causal gene for the phenotype of interest.

zoom_numperm Number of permutations for zoom-in score calculation.