1 Introduction

We provide here documentation supporting PAINTOR (V1.1). A full description of the methods can be found in:

Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Kraft P, Pasaniuc B. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics 10.10 (2014): e1004722.*

Sections 2 provides instructions for installation of the software. Sections 3 and 4 describe the input and output for PAINTOR. In section 5 we provide instructions on how to run the software and detail a recommended pipeline.

2 Installation

The early release of PAINTOR is optimized to run on a UNIX-like system. To install the software, unpack the zip file in the directory you want PAINTOR installed. CD into the PAINTOR folder and run the installation script:

> tar -xvf PAINTOR.tar.gz
> cd PAINTOR
> bash install.sh

This will unpack and compile all the necessary dependencies and create an executable PAINTOR . The two main libraries that PAINTOR uses are Eigen V3.2 and NLopt V2.4.2.

`http://eigen.tuxfamily.org/index.php?title=Main_Page`
`http://ab-initio.mit.edu/wiki/index.php/NLopt`

3 Input

For each locus of interest, there are three input files needed to run the software:

1. Z-scores
2. LD matrix
3. Annotation matrix

A directory containing of the required files to run a fine-mapping study on 40 loci is included with the software package.

3.1 File Formats

3.1.1 Z-score file

The Z-score file should contain a single column of Z-scores with no header. The z-score of a SNP is the Wald statistic ($\frac{\hat{\beta}}{SE(\hat{\beta})}$) of a marginal regression of the phenotype on the SNP.

Example: A Z score file of a locus with 4 SNPs
1.5
-3.2
6.0
4.0

3.1.2 LD matrix file

The LD file contains a symmetric matrix of Pearson correlation coefficients where entry i,j will correspond to the correlation between SNPs i and j (r_{ij}). White space must separate individual columns of the matrix. The first line in the file MUST specify the dimensions of the matrix.

Example: A potential 4 SNP LD matrix
4 4
1.0 0.5 0.5 0.2
0.5 1.0 0.3 0.1
0.5 0.3 1.0 0.9
0.2 0.1 0.9 1.0

Note: (1) Particular care must be taken when computing LD from a reference panel such the 1000 genomes. It is imperative that all the reference and alternate alleles for SNPs from which the z-scores were computed match the reference and alternate alleles of the reference panel. The output of PAINTOR will not be correct if there are mismatches of this type in the data. (2) The LD matrix may be ill-conditioned and need regularization. We recommend adding an ϵ to the diagonals of each matrix to overcome this.

3.1.3 Annotation Matrix File

There should be an annotation file for each locus that contains a matrix of annotations that are typically binary. The rows of the matrix correspond to SNPs at that locus and columns represent unique annotations. For example, if the first column of the matrix represented "coding region", and entry [2,1] of the matrix was equal to 1, this would signify that SNP 2 falls within a coding region. We note that annotation columns must be consistent across all loci. In other words, from the earlier example, column 1 of the annotation matrix should correspond to "coding region" in all the annotation files. The first line in the file MUST specify the dimensions of the annotation matrix.

Example: A locus with 4 SNPs and three potential annotations {A1, A2, A3}
4 3
1 0 1
0 1 0
1 1 1
1 0 0

3.2 File Naming conventions for input

PAINTOR is designed to run on multiple loci simultaneously. The program takes as input a single file with a list of file names that contain the z-scores for all the loci the user wants to consider jointly. By convention, these names serve as the prefix for the entire locus. The files containing the corresponding LD matrix and annotation file will have suffixes ".LD" and ".annotations" appended to them (these can modified using the appropriate flags described below). For example, if one wanted to run PAINTOR on three loci, the input

file would look like this:

Locus1
Locus2
Locus3

And the directory that contained all the files would contain the following 9 files (i.e. 3 files for each locus)

Locus1
Locus1.LD
Locus1.annotations
Locus2
Locus2.LD
Locus2.annotations
Locus3
Locus3.LD
Locus3.annotations

4 Output

4.1 Posterior Probabilities

Files that contain a column vector of posterior probabilities for SNPs to be causal are the main output of PAINTOR for each locus. The ith row of the file corresponds to the ith SNP at the locus.

Example: Posterior probabilities for a locus of 4 SNPs
0.002
0.013
0.980
0.130

Note: The default filename will be the locus name with ".PaintorProbs" appended to it. User can change name using the -OUTname flag.

4.2 Gamma Estimates

A file that has the effect size estimates for each of the annotation(s) used. PAINTOR automatically estimates the baseline annotation (A0) and will always output this values as the first line in the file.

Example: Estimates for 2 annotations (+1 baseline) {A0, A1, A2}

5.2
-1.3
2.0

These effect sizes can be converted to probabilities using the expit transformation. For the preceding output the corresponding prior probabilities can be calculated as follows:

The baseline prior probability for any SNP in the fine-mapping dataset to be causal is obtained as:

$$\begin{aligned} \frac{1}{1+exp(\gamma_0)} &= \frac{1}{1+exp(5.2)} \\ &= 0.0055 \end{aligned}$$

The prior probability for a SNP in annotation A1:

$$\begin{aligned} \frac{1}{1+exp(\gamma_0+\gamma_1)} &= \frac{1}{1+exp(5.2+(-1.3))} \\ &= 0.01984 \end{aligned}$$

The relative probability for a SNP to be causal given that it is in A1 is simply computed as

$$\frac{0.01984}{0.0055} = 3.6$$

Note: The default filename is GammaEstimate.txt. User can change name using the -Gname flag

4.3 Final log-likelihood

A file that has the final log likelihood of the PAINTOR model. This can be used in subsequent steps to conduct a likelihood ratio test (LRT) for significance of the annotation effect sizes. To test the marginal significance of a single annotation (A1) one would first fit a PAINTOR model with just the baseline annotation A0 (M0) then fit a joint model with both annotations A0, A1 (M1). The resultant log likelihoods for each model can be used to compute an LRT statistic which will be distributed asymptotically χ^2 with degrees of freedom = 1 (under the null).

Example: Testing significance of annotation A1

Model 1 log likeilhood with only baseline annotation (A0)
-10039
Model 2 log likeilhood with both annotations(A0, A1)
-10036

$$\begin{aligned} LRT &= -2[ln(likelihood(M0)) - ln(likelihood(M1))] \\ &= -2[-10039-(-10036)] \\ &= 6 \end{aligned}$$

$\sim \chi^2$ (df=1) p-value= 0.0143

5 Running software

5.1 PAINTOR

PAINTOR computes posterior probabilities for SNPs to be causal and quantifies enrichment of causal variants within the given annotations.

Usage:
-input (required) specifies the name of the input file that lists the names of all the z-score files

-d (required) specifies the input directory for the three file types (Z-scores, LD, Annotations)

-o (required) specifies the output directory

-c (required) specifies the number of potential causal snps to consider at each locus

-i (recommended) specifies the index of the annotations to be considered for the model. Note: the index is 0-based so if one wants to consider the first column of the annotation matrix the corresponding flag would be (-i 0). To consider multiple annotations separate successive indices by commas. For example to consider the second and fifth annotations the required flag is (-i 1,4). If -i flag is not specified the default is to fit the model with only the baseline annotation.

-Gname (optional) filename of gamma estimates. Default: EstimatedGamma.txt

-Lname (optional) filename of final likelihood file. Default: Likelihood.txt

-LDname (optional) specify the suffix of LD file name. Default: LD

-ANname (optional) specify the suffix of Annotation file name. Default: annotations

-OUTname (optional) specify the suffix of the output probabilities. Default: PaintorProbs

-MI (optional) maximum number of iterations to run algorithm. Default: 10

-m (optional) 1 = fast computation of posterior probabilities under the assumption of a single causal variant with no annotations. 0 = default

Example: To run PAINTOR on ten loci and consider up to three causal variants per locus using the first annotation.

> ./PAINTOR -input input.files -d /Input/ -o /Output/ -c 3 -i 0

5.2 Suggested Pipeline

In order to determine which annotations are relevant to the phenotype being considered, we recommend running PAINTOR on each annotation independently.

Example: 10 Loci with 100 annotations considering up to three causal variants per locus

> ./PAINTOR -input input.files -d /Input/ -o /Output/ -c 3 -Gname Baseline -Lname BaseLikeli
> ./PAINTOR -input input.files -d /Input/ -o /Output/ -c 3 -i 0 -Gname Annot1 -Lname Likeili1
> ./PAINTOR -input input.files -d /Input/ -o /Output/ -c 3 -i 1 -Gname Annot2 -Lname Likeili2
> ./PAINTOR -input input.files /Input/ -o /Output/ -c 3 -i 2 -Gname Annot3 -Lname Likeili3
.
.
.
> ./PAINTOR -input input.files -d /Input/ -o /Output/ -l 10 -c 3 -i 99 -Gname Annot99 -Lname100

Likeili100

After obtaining the output for all of the annotations marginally. Compute likelihood ratio statistics using the baseline as the null model to determine the most significantly associated annotations . Then use those annotations in a final model to compute trait-specific posterior probabilities for causality:

> ./PAINTOR -input input.files -d /Input/ -o /Output/ -c 3 -i 1,10,53