RLMM - Robust Linear Model with Mahalanobis Distance Classifier

Nusrat Rabbee and Gary Wong

October 17, 2016

Contents

1 Introduction

The RLMM package is a genotype calling algorithm for Affymetrix SNP arrays with a classification algorithm based on multi-chip, multi-SNP approach for Affymetrix SNP arrays. This package uses a supervised method of classifying Affymetrix SNP array data, and works by using a large training sample where genotype labels are known to obtain a more accurate classification on new data. The training set contains publicly available genotypes available for 90 Centre du Etude Polymorphisme Human(CEPH) individuals from the HapMap project.

This package contains three main functions for classification and one for plotting.

1. We start out with normalization of probe data in ascii format (*.raw files*) with function **normalize_Rawfiles**, in order to get corresponding norm files (*.norm files*). These norm files are used to scale the new data to the scale of the training set. Thus the normalized probe intensities are stored in the norm files. If the user has CEL files instead, there is a section below on how to convert these files to .raw files.

2. The next step is allele summarization which involves **create_Thetafile**. This function calculates estimates of theta A and theta B for each chip, for each SNP, based on the normalized intensities stored in the *norm files*. The theta A and theta B values are summary measures of probe intensities for each allele. Only perfect match PMA and PMB values are used for calculating the theta values.

3. The last step is Classification using Mahalanobis distances. This involves calling the **Classify** function and using a *regions file* (this is an internal file that should

be downloaded from our website [1]) obtained from the training data and a *theta file* obtained from step two.

2 Instructions for Genotyping Affymetrix Mapping 100K array - Xba set

We will cover everything needed to classify probe data starting with installation and internal files that the algorithm depends on. Before even starting though, make sure that you are using *Raw files* (ascii versions of *CEL files*). Also at this very moment, this algorithm works only for 100K - Xba data set. It will work with Hind data once we obtain the data to make a training set, and corresponding internal files. Please make sure that the two internal files *Xba.CQV* and *Xba.regions* are also downloaded onto your computer, preferably in the working directory (i.e., in the directory from where you will invoke R). These files are necessary for the functions to work on Xba data set and can be obtained on our website (see foot note).

2.1 Installation

Once the package is downloaded, install the package on a UNIX machine using the following commands in the console:

R CMD INSTALL -l /dir/mylocation RLMM.tar.gz

After the -l, please put in the pathname where the package is to be stored. When installing, the part between R CMD INSTALL and RLMM.tar.gz may be disregarded if you have the permissions to write to the main library in R-Home. For more information see the R-admin pdf from CRAN on installing libraries. Currently, this package is working under Unix. (We will shortly release a version that will work under Windows) Once this package is installed onto the Unix/Linux system, it can be accessed by R console by typing:

```
R> library(RLMM)
```

To reiterate, the functions require *internal files* to work properly. We advise that you save these files in the same directory as your working directory, so that it can be referred to by the package, when doing classification and normalization.

[1] http://www.stat.berkeley.edu/users/nrabbee/RLMM

2.2 Converting CEL files to .raw files (ascii)

Affymetrix has a tool available which converts your *.CEL files to .raw files*. This tool is available for download from our website [2]. After downloading this Affymetrix tool onto your UNIX/Linux system, type at the unix system prompt:

```
$gtype_cel_to_pq -subset Xba.snpnames -cdf Mapping50K_Xba_240.CDF NA06985_Xba_B5_4000090.CEL
```

If this command works, you will get *NA06985_Xba_B5_4000090.raw* file created in your directory. Do this for all the *.CEL files* in your directory. You can download this tool and the *.CDF files* from our website. The above Affymetrix conversion tool (which has been compiled on a 64-bit AMD linux machine), *Xba.snpnames*, and /it CDF files are included in the tarfile you will download from our website.

2.3 Quick Start

This section will go through the usage of basic functions that are packaged with the library RLMM (pronounced realm) to get you up and running as fast as possible, from probe level data (intensity values) to classification of SNPs by running through a short example.

Before starting please do the following:

- Make sure that the *.raw files* are all in one directory (*probefiledir*) and both internal files (*Xba.CQV and Xba.regions*) are together in a directory (same or different than *probefiledir*). We suggest that you keep the files together in a directory and set directory to that location at the beginning of the R session.
- Load the library **RLMM** by going to R console and type:

```
R> library(RLMM) ##load the RLMM package
```

Function Calls

Reading and Normalizing the Probe Data

Probe intensity data can be read in and normalized in one step by using the **normalize_Rawfiles** function. This function will normalize (translate to the same scale as the training data), each *.raw file* and give a corresponding *.norm file*. Keep all *.raw and .norm files* in the same directory (*probefiledir*). Do not be alarmed if it takes some time, the time it takes to finish depends on the number of *raw* files.

In R:

[2] http://www.stat.berkeley.edu/users/nrabbee/RLMM

```
R> normalize_Rawfiles(cqvfile="Xba.CQV")
```

NOTE: Stating the name of the *cqvfile* (e.g., *Xba.CQV*) and location of probe data, i.e., .raw files in probefiledir. The first parameter is required. The second parameter needs to be specified only if the .raw files are not in the working directory.

Getting theta estimates (estimated probe intensity for allele A and B)

To get theta estimates for theta A and theta B for each SNP and chip, we need to use the **create_Thetafile** function. In the end, the estimates will be stored in a ascii text format of a name of your own choosing (*thetafile*). To invoke this function, run the following command below in R:

```
R> create_Thetafile(start=1,end=100,thetafile="Xba.theta")
```

The parameter *thetafile* is required to be filled in. This will create the ascii theta file where start specifies the 1st SNP and end specifies the last SNP to be processed. By default, end = -1, which makes the function process all SNPs in the .norm files (e.g., 58960 SNPs in the Xba set). *Probefiledir* is only needed if the probe files is not in the working directory.

Classification

Classification is done by a Mahalanobis distance classifier after genotype group centers and variance-covariance matrices are determined by training data. For this important step, you must call function **Classify**. Here, you will need our internal file, *Xba.regions* for the regionsfile argument. The genotypefile argument should be the name of the text file that will hold the results of Classify will produce. *Call rate* allows the change of the cut-off value to make more accurate calls. Currently, eligible call rates are: 80,82,84,86,88,90,92,94,96,98,100. If you don't specify it, the default is 100%. So, all calls are made. If you specify a ineligible call rate (e.g., 91), the call rate will be set to 80% which is the minimum. Note, with lower call rate, higher accuracy is achieved. Below is an example of how to use the function Classify with a snip of the results that you will get in the *rlmm file*.

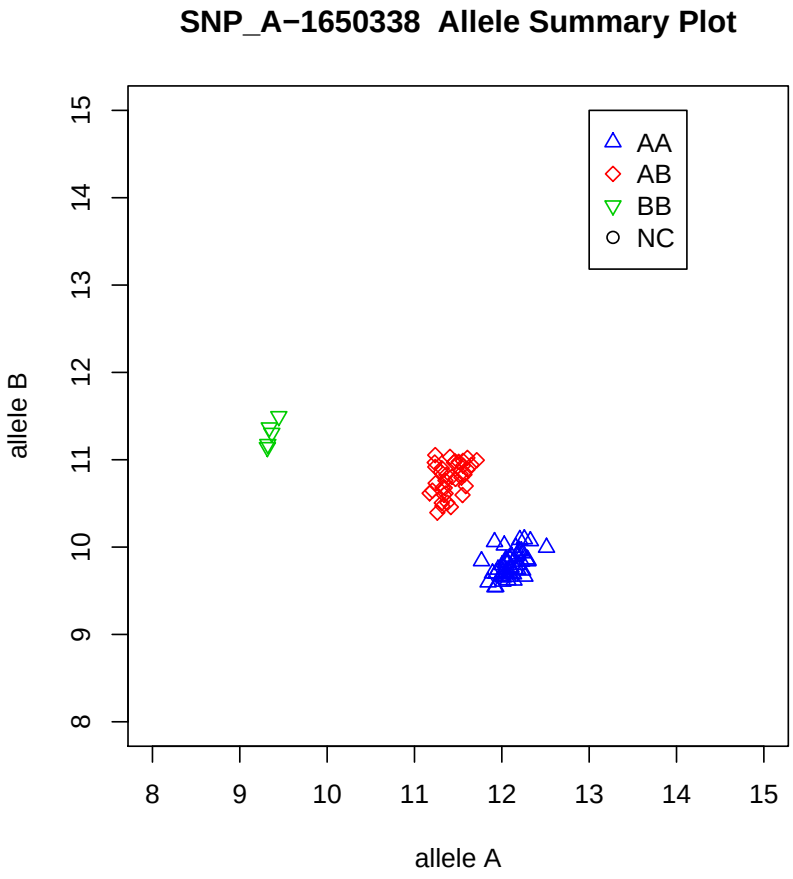
```
R> Classify(genotypefile="Xba.rlmm",regionsfile="Xba.regions",thetafile="Xba.theta")
```

```
> x<-read.table("Xba.rlmm")
> x[1:2,1:5]

             V1 V2     V3 V4     V5
1 SNP_A-1650338 AA 0.2359 AA 0.4344
2 SNP_A-1716667 AA 0.0934 AA 3.0806
```

Allele Summary Plots

This is useful for exploratory purposes and to see visually how tight each cluster of genotypes, a particular SNP exhibits. To create the theta plots we need to specify a genotype file and a theta file. In the example below, we set *Pick.Obj* equal to *FALSE* (NOTE: It should always be set to false at this point with RLMM ver 0.7). The parameter *snpsfilename* is a vector of snps we wish to plot, ideally each SNP is listed as a newline as a text file. Running the command below will save the plot in plots.ps, but if *plotfilename* is left blank "", it will display the graph onscreen.

```
R> plot_theta(genotypefile="Xba.rlmm",thetafile="Xba.theta",plotfile="",snpsfile="snps.lst")

[1] "Matching at Index"
[1] 1
```



3 Where can I learn more about RLMM?

Updated information on RLMM will be available at our website (see references). This is the main site where all information pertaining to RLMM will be stored including updates and new files for the package if necessary. For other packages and information relating to bioinformatics, check out Bioconductor [3] project.

Contact information for Nusrat Rabbee is <nrabbee@post.harvard.edu> and Gary Wong is <wongg62@gmail.com>

A Previous Release Notes

B Future Plans

- Add updates to package to handle 100K-Hind data set, 500K, 10K, etc.
- Add examples and more documentation
- Make the package more user friendly

C References

The pre-print of the manuscript containing a description of RLMM is available at the RLMM website http://www.stat.berkeley.edu/users/nrabbee/RLMM.

[3] http://www.bioconductor.org/