Introduction To affyCustomCdf

Ernur Saka[*]
Eric C. Rouchka[*]

October 15, 2017

Abstract

Commercially developed Affymetrix® GeneChip® microarrays have been widely used for understanding differential expression changes on a genome-wide scale. One main drawback of microarray technology involves the static selection of probes based on available genomic knowledge and gene annotation information at the design stage. As the genomic and transcriptomic knowledge evolved, the need for a dynamically changing approach to microarray expression analysis has become apparent. The package affyCustomCdf provides a method to create custom CDFs (Chip Description File) via removing nonspecific probes, updating probe target mapping based on the supplied genome information and grouping probes into region (exon, UTR), gene and transcript levels. This document explains the use of the package as well as the data preparation. For the explanation of the method please see our paper http://www.biorxiv.org/content/early/2017/04/11/126573.

Contents

1 Input Data and Preparations 2
1.1 Obtaining probe mapping file . . . 2
1.1.1 Recommended probe mapping text file creation . . . 2
1.2 Obtaining annotation file . . . 3
1.3 Obtaining original CDF . . . 3

2 Selecting Parameters 3

3 Running The affyCustomCdf Tool 5

4 Interpreting Results 5
4.1 CDF File . . . 6
4.1.1 Data analysis via Custom CDFs . . . 6
4.2 Report File . . . 7
4.3 Missing Probes File . . . 7

References 7

[*]University of Louisville

1 Input Data and Preparations

1.1 Obtaining probe mapping file

The affyCustomCdf expects probe mapping as a text file. The probe mapping file consists of one line per probe, each containing five columns of data. Columns must be tab delimited without a header line. Columns are in the following order: x location of a probe, y location of a probe, sense/antisense of a probe (-/+), chromosome name and chromosomal starting location. The following demonstrates the first 10 lines of the Rat 230 2.0 probe mapping file. The file is included inside the scripts directory of the github repository branch affyCustomCdfFull. https://github.com/UofLBioinformatics/affyCustomCDF/tree/affyCustomCdfFullaffyCustomCdf

```
335 337 -   8   22161426
99  499 -   8   22161407
726 583 -   8   22161386
386 415 -   8   22161292
101 649 -   8   22161280
180 413 -   8   22161263
13  31  -   8   22161145
258 663 -   8   22160989
392 57  -   8   22160974
```

1.1.1 Recommended probe mapping text file creation

The DNA sequences for perfect match (PM) probes is obtained from the Affymetrix® Netaffx^{TM} web site in a FASTA file format. PM probes are aligned to the genome using Bowtie version 1.0.1 (Langmead et al. 2009) with the parameters -v 0 and -m 1 which returns the alignment results for the probes that align to one genomic location with 100% identity. Note that Bowtie version 1 is best at aligning shorter sequences (25-50 bp) as found with microarray probes while the most recent versions of bowtie are optimized for long sequence reads (>50 bp). Also, the –suppress parameter can be used to eliminate some output for clarity. The following command line is an example for aligning Rat 230 2.0 PM to Rat assembly rn6. It suppresses the 5th, 6th and 7th default Bowtie outputs.

```
$ ./bowtie -v 0 -best -m 1 rnRnor6Chromosome -f Rat230_2.fa rnRnorchromosome6.txt
--suppress 5,6,7 --quiet --max maxrnRnorChromosome6.txt
--un unalignedrnRnorChromosome6.txt
```

If the index of the intended genome of an organism is not supplied by Bowtie, the assembled unmasked genomic DNA sequences of every chromosome can be downloaded from one of the online repositories such as Ensembl (Cunningham et al. 2015) and NCBI and indexes can be created from DNA sequences via Bowtie-Build version 1.0.1 with default parameters. The following is the one line of mapping result file.

```
Rat230_2:1367453_at; 335; 337; 1100; Antisense; -   8   22161426
```

The file can be formatted via an R script to obtain the proper column arrangement. More script examples and data files can be found under the scripts folder of the github repository branch affyCustomCdfFull with the file name ProbeFileClear.R.

```
fileName = "rat.txt"
probes = read.csv2(fileName, header = FALSE, sep = "\t", quote = "\"",
dec = ",", fill = TRUE, col.names = c("names","sense","chromosome", "start",
"empty") ,comment.char = "",stringsAsFactors = FALSE)

X = unlist(lapply(strsplit(probes[,1],";"), "[[" , 2))
Y = unlist(lapply(strsplit(probes[,1],";"), "[[" , 3))

probeTable = data.table(X = as.numeric(X), Y = as.numeric(Y),
direction=unlist(probes$sense), chromosome=unlist(probes$chromosome),
start=unlist(probes$start))

write.table(probeTable, "Rat230-2Rn6ChrAligned.txt", quote = FALSE,
row.names = FALSE, col.names = FALSE, sep="\t")
remove(probeTable,X,Y,probes)
```

1.2 Obtaining annotation file

The affyCustomCdf tool accepts annotations as a General/Gene Transfer Format (GTF). In the GTF file, fields must be tab delimited. The GTF format consist of nine columns of data per feature. The columns are seqname, source, feature, start, end, score, strand, frame and attribute. Comment lines in the file must start with #. The feature column is being used to classify probes for gene features. Therefore region features must contain transcript, EXON, CDS and UTR key words (not case sensitive). If separation between 3′ UTR and 5′ UTR is desired, features must include the direction information such as three_prime_utr or five_prime_utr. Otherwise annotation will be performed without classificatio of 3′ and 5′. For more information please check the Ensembl GFF/GTF file format definition. http://www.ensembl.org/info/website/upload/gff.html. A sample GTF file can be found in the scripts directory of the github repository branch affyCustomCdfFull.

1.3 Obtaining original CDF

To obtain the probe information of a chip, the affyCustomCdf tool uses the original CDF file provided by Affymetrix®. A specific CDF file can be obtained from the Affymetrix® Netafffx^{TM} web site or from the Gene expression omnibus (GEO) (Barrett et al. 2013). They are usually included in the library files of a chip.

2 Selecting Parameters

The affyCustomCdf tool takes thirteen parameters. This includes the original CDF (Section 1.3), probe alignment file (Section 1.1) and annotation (GTF) file (Section 1.2) which must be supplied by the user. The remaining parameters are optional, with default value set.

The parameters are:

- **orginalCdfName:** String type. The original CDF file of the selected Affymetrix® GeneChip® technology. It can be obtained from Affymetrix® NetAffx^{TM} (Liu et al. 2003) web site.

  ```
  Rat230_2.cdf file for the Rat 230 2.0
  ```

- **probeAlignmentFile:** String type. The tab separated probe alignment file. Please see section 1.1 for more details.

- **gtfFileName:** String type. The General/Gene Transfer Format (GTF) file name. Please see section 1.2 for more details.

- **newCDFName:** String type. Name of the created custom CDF. If the user does not provide a name, a name will be created based on the template provided in the flowing line.

  ```
  orginalCdfName_type_min_minProbeSetNumber_D_SD.cdf
   * type the value of the type parameter
   * min stands for minimum
   * minProbeSetNumber indicates the minimum number of probe per probe set
   * D stands for direction
   * SD indicates the value of direction parameter
  ```

- **reportFile:** String type. Name of the report file name. If the user does not provide a name, a name will be created based on the template provided in the flowing line.

  ```
  report_orginalCdfName.txt
  ```

- **controlProbeSetNumber = 0:** Number type. The number of control probe sets in the original CDF. Control probe sets were being designed to check the quality of the experiment. They can be detected in the CDF based on the probe set names which usually starts with AFFX prefix such as AFFX_ratb2/X14115_at. If the number of control probe set is given, they will be included in the custom CDF without changes otherwise they will not be included. The default value is 0 (not included).

- **minProbeSetNumber = 3:** Number type. The minimum number of probes in a probe set. Probe sets with less than the minimum number will not be included in the custom CDF. The default value is three.

- **probeLength = 25:** Number type. Number of bases in a probe. In most cases length of a probe is twenty-five. For different sizes, it can be obtained via checking a sequence of a probe belongs to an Affymetrix® GeneChip®. The default value is 25.

- **SD = 1:** Number type. Sense/antisense relationship between probes and annotations. Possible values are 0 (no direction), 1 (same direction) and 0 (opposite direction). The default value is 1 (Same direction).
  - When SD is 0, direction will not be considered during probes to annotations mapping.
  - When SD is 1, mapping is performed between same directions such as the sense probes are being mapped to the sense annotations.
  - When SD is 0, mapping is performed between opposite directions such as the sense probes are being mapped to the anti-sense annotations.

- **type = "regionG":** String type. The type of the created custom CDF. Options are regionG, gene, transcript. Default value is regionG.

  - regionG: When regionG option is selected, probe sets are designed to target a specific region (exon, UTR) of a gene and consist of probes which map to the same region of a gene.
  - gene: When regionG option is selected, probe sets are designed to target genes and consist of probes which map to the same gene.
  - transcript: When transcript option is selected, probe sets are designed to target transcripts and consist of probes which map to the same transcript.

- **transcriptShare = FALSE:** Bool type. It defines whether to allow probe sharing between transcripts of a gene. It can only be used when transcript type is selected. Possible values are FALSE for not allowing probe share and TRUE for allowing probe share. The default value is FALSE.

- **junction = FALSE:** Bool type. It defines whether we are adding probes to probe sets that map onto a junction of a specific region of a gene. It can be used when regionG or gene option is selected. Options are FALSE for no junction and TRUE to allow junctions. The default value is FALSE.

- **uniqueProbe = FALSE:** Bool type. It defines whether to use probes that map onto more than one annotations. It is an option for the user to examine unique probes to one specific annotation. It can only be used with the regionG type. Possible values are FALSE for not unique, TRUE for unique. The default value is FALSE.

3 Running The affyCustomCdf Tool

To create a custom CDF via affyCustomCdf tool, one must call the createAffyCustomCdf function after installation of the tool.

```
#Creates region based CDF for the Rat 230 2.0 with defult values
createAffyCustomCdf("Rat230_2.cdf","rat230Probes.txt",
"Rattus_norvegicus.Rnor_6.0.85.gtf")

#Creates gene based CDF for the Rat 230 2.0
createAffyCustomCdf("Rat230_2.cdf","rat230Probes.txt",
"Rattus_norvegicus.Rnor_6.0.85.gtf",
controlProbeSetNumber = 57, minProbeSetNumber = 2, type = "gene")

#Creates transcript based CDF for the HG-133 Plus 2
createAffyCustomCdf("HG-U133_Plus_2.cdf","human133Probes.txt",
"Homo_sapiens.GRCh38.77.gtf", controlProbeSetNumber = 62, minProbeSetNumber = 2,
type = "transcript")
```

4 Interpreting Results

createAffyCustomCdf function creates three files. They are custom CDF file, report file and missing probes file.

4.1 CDF File

Produced custom CDF file consists of regrouped probes into probe sets based on specific annotations of interest. It can be used for data analyses of Affymetrix® GeneChip® data.

4.1.1 Data analysis via Custom CDFs

1- Make custom CDF package and install the package

```
cdfFileName = "RAT230_2_regionG_min_1_D_0.cdf"
cdfFilePath = "C:/Users/Ernur/Documents/MicroArrayProject/affyCustomCdf/analyze"
specy = "Rat"
pkgpath = tempdir()
make.cdf.package(cdfFileName, cdf.path= cdfFilePath, compress=FALSE,
species = specy, package.path = pkgpath)
```

2- Go to directory and run the following commands from the command prompt

```
cmd
run R CMD INSTALL RAT2302regionGmin1D0cdf
```

3- Write an R script to analyze data produced via the Affymetrix® GeneChip$^{\circledR}$ of customized CDF. The following script can be taken as an example.

```
source("http://bioconductor.org/biocLite.R")
biocLite("GEOquery","affy"."limma","simpleaffy")
library(GEOquery)
library(R.utils)
library(affy)
library(limma)
library(simpleaffy)
cdf="RAT2302regionGmin1D0cdf"
library(cdf,character.only = TRUE)
#Download raw data from GEO
gse_number = "GSE72551"
gse <- getGEO(gse_number, GSEMatrix = TRUE)
filePaths = getGEOSuppFiles("GSE72551")
COMPRESSED_CELS_DIRECTORY = gse_number
untar( paste( gse_number , paste( gse_number , "RAW.tar" , sep="_") , sep="/" ),
       exdir=COMPRESSED_CELS_DIRECTORY)
cels = list.files( COMPRESSED_CELS_DIRECTORY , pattern = "[gz]")
sapply( paste( COMPRESSED_CELS_DIRECTORY , cels, sep="/") , gunzip )
#Read cel files based on custom CDF
celData = ReadAffy( celfile.path = gse_number,cdfname=cdf)
#Create sample inforamtion file such as
#ArrayCode  SampleInformation   ExperimentalCondition
#LA06012401.CEL Rat 01 Naive Replicate 01   0
datafile = "SampleInformation.txt"
pData(rawData2) = read.table(datafile, header=T,row.names=1,sep="\t")
pData(rawData2)
normalizedData = rma(rawData2)
library(genefilter)
normalizedData.filtered = featureFilter(normalizedData, require.entrez=FALSE,
                                        require.GOBP=FALSE, require.GOCC=FALSE,
                                        require.GOMF=FALSE,
                                        require.CytoBand=FALSE,
                                        remove.dupEntrez=FALSE,
                                        feature.exclude="^AFFX")
design <- model.matrix(~ -1+factor(c(1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,3,3,3)))
colnames(design) <- c("C0", "C7","C14")
contrast.matrix <- makeContrasts(C7-C0,C14-C0,levels=design)
fit = lmFit(normalizedData.filtered,design)
fit2=contrasts.fit(fit,contrast.matrix)
fit2 = eBayes(fit2)
arrangedD1 = topTable(fit2,coef = 1, adjust = "fdr", n = length(fit2),
                      sort.by='P')
arrangedD4 = topTable(fit2,coef = 2, adjust = "fdr", n = length(fit2),
                      sort.by='P')
#Save results
write.table(arrangedD1 ,file="C7vsC0.xls",row.names=T, sep="\t",
            col.names=NA)
write.table(arrangedD4 ,file="C14vsC0.xls",row.names=T, sep="\t",
            col.names=NA)
```

4.2 Report File

Report file consists of multiple information. They are:

- Parameter values sent to the CreateAffyCustomCdf function.
- Distribution of number of probes per probe set table.
- The histogram of the number of probes per probe set.

4.3 Missing Probes File

In some cases FASTA file of probes may include some probes that is not used in the original CDF. These probes are detected and placed into a text file to inform users. Name of the file starts with the original CDF name with postfix missingProbes. i.e Rat230_2.cdf_missingProbes.txt

References

Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, et al. 2013. "NCBI Geo: Archive for Functional Genomics

Data Sets—update." Journal Article. *Nucleic Acids Research* 41 (Database issue): D991–D995. doi:10.1093/nar/gks1193.

Cunningham, Fiona, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, et al. 2015. "Ensembl 2015." Journal Article. *Nucleic Acids Research* 43 (D1): D662–D669. doi:10.1093/nar/gku1010.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short Dna Sequences to the Human Genome." Journal Article. *Genome Biology* 10 (3): R25. doi:10.1186/gb-2009-10-3-r25.

Liu, G., A. E. Loraine, R. Shigeta, M. Cline, J. Cheng, V. Valmeekam, S. Sun, D. Kulp, and M. A. Siani-Rose. 2003. "NetAffx: Affymetrix Probesets and Annotations." Journal Article. *Nucleic Acids Research* 31. doi:10.1093/nar/gkg121.