Generating Grey Lists from Input Libraries

Gord Brown

Edited: 2020-07-29; Compiled: October 27, 2020

Contents

1 Introduction

Many cell lines and tumour samples show anomalous signal in the input or control sample in some regions. These regions also show high signal in the corresponding ChIPs. Peak callers are not, in general, well-behaved in these regions, tending to call many spurious peaks. The purpose of this package is to identify those regions, so that reads in those regions may be removed prior to peak calling, allowing for more accurate insert size estimation and reducing the number of false-positive peaks.

As part of the ENCODE project, Anshul Kundaje identified regions that show enrichment in ChIP experiments independent of what factor is being ChIPped, or what cell line the sample comes from[1, 2]. He called those regions "signal artefact" regions, or colloquially "black lists". We call our lists of high signal grey lists, to distinguish them from ENCODE's black lists, because they are not universal, but rather cell line (or sample) specific, and because they can be tuned depending on the stringency required, and so that we can make jokes about having 50 shades of grey lists[3].

This vignette summarizes the construction of grey lists.

2 Generating a Grey List

Generating a grey list involves

1. generating a tiling of the genome,

2. counting reads from a BAM file for the tiling,
3. sampling from the counts and fitting the samples to the negative binomial distribution to calculate the read count threshold,
4. filtering the tiling to identify regions of high signal, then
5. exporting the resulting set to a bed file.

First create the `GreyList` object (this karyotype file includes just human chromosome 21, from reference genome version GRCh37):

```
> library(GreyListChIP)
> path <- system.file("extra", package="GreyListChIP")
> fn <- file.path(path,"karyotype_chr21.txt")
> gl <- new("GreyList",karyoFile=fn)
```

Normally the next step would be to count reads:

```
> gl <- countReads(gl,"myBamFile.bam")
```

but to save time we'll generate some fake data:

```
> gl@counts <- rnbinom(length(gl@tiles),size=1.08,mu=11.54)
```

Now calculate the threshold. The defaults are `reps=100,sampleSize=30000,p=0.99` but for demonstration purposes we'll use smaller values for faster results:

```
> gl <- calcThreshold(gl,reps=10,sampleSize=1000,p=0.99,cores=1)
```

This method fits the sample(s) to the negative binomial distribution, then uses the estimated parameters to identify a read-count threshold[4, 5].

Now generate the grey list itself:

```
> gl <- makeGreyList(gl,maxGap=16384)

coverage: 2871289 bp (5.97%)

> gl

GreyList on karyotype file karyotype_chr21.txt
  tiles: 94004
  size (mean): 1.12255508234841
  mu (mean): 11.8411090305001
  params: reps=10, sample size=1000, p-value=0.99
  threshold: 53
  regions: 632
  coverage: 5.97%
```

(The coverage is higher than normal due to the counts being fake. Normally a threshold of p=0.99 leads to coverage of about 1% of the genome.)

And export it to a file:

```
> export(gl,con="myGreyList.bed")
```

And that's it. If you are happy to accept the package's defaults, you can generate the list in one step (not counting the `export` step):

```
> library(BSgenome.Hsapiens.UCSC.hg19)
> gl <- greyListBS(BSgenome.Hsapiens.UCSC.hg19,"myBamFile.bam")
> export(gl,con="myGreyList.bed")
```

3 Sample Data

A sample `GreyList` object named `gl` can be obtained, once the package is attached, via:

```
> # Load a pre-built GreyList object named "gl"
> data(greyList)
> print(gl)

GreyList on karyotype file karyotype_chr21.txt
  tiles: 94004
  size (mean): 1.12255508234841
  mu (mean): 11.8411090305001
  params: reps=10, sample size=1000, p-value=0.99
  threshold: 53
  regions: 632
  coverage: 5.97%
```

This sample object covers only human chromosome 21 (from genome version hg19). The read counts are from an MCF7 input library constructed in the Carroll Lab of Cancer Research UK's Cambridge Institute. See the `greyList` man page for details of this object.

4 Obtaining Karyotypes

If a `BSgenome` object exists for your reference genome of interest, the karyotype is usually most easily obtained via that object. See the `BSgenome` package documentation for a list of available reference genomes[6].

Otherwise, if the reference genome is available via the UCSC Genome Browser[7], karyotype files can be obtained using the `fetchChromSizes` utility available on the Genome Browser's software download page[8].

Failing that, a karyotype file can be constructed by hand using a text editor. The file format is given in the `loadKaryotype` documentation. All that is needed is the names of the chromosomes, (exactly) matching the names in the BAM file, and their lengths in base pairs.

5 Acknowledgements

Thanks to Rory Stark and Tom Carroll for suggestions, advice and encouragement.

6 Session Info

```
> toLatex(sessionInfo())
```

- R version 4.0.3 (2020-10-10), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=C`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Running under: `Windows Server 2012 R2 x64 (build 9600)`
- Matrix products: default
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, stats4, utils
- Other packages: BiocGenerics 0.36.0, GenomeInfoDb 1.26.0, GenomicRanges 1.42.0, GreyListChIP 1.22.0, IRanges 2.24.0, S4Vectors 0.28.0
- Loaded via a namespace (and not attached): BSgenome 1.58.0, Biobase 2.50.0, BiocManager 1.30.10, BiocParallel 1.24.0, BiocStyle 2.18.0, Biostrings 2.58.0, DelayedArray 0.16.0, GenomeInfoDbData 1.2.4, GenomicAlignments 1.26.0, MASS 7.3-53, Matrix 1.2-18, MatrixGenerics 1.2.0, RCurl 1.98-1.2, Rsamtools 2.6.0, SummarizedExperiment 1.20.0, XML 3.99-0.5, XVector 0.30.0, bitops 1.0-6, compiler 4.0.3, crayon 1.3.4, digest 0.6.27, evaluate 0.14, grid 4.0.3, htmltools 0.5.0, knitr 1.30, lattice 0.20-41, matrixStats 0.57.0, rlang 0.4.8, rmarkdown 2.5, rtracklayer 1.50.0, tools 4.0.3, xfun 0.18, yaml 2.2.1, zlibbioc 1.36.0

References

[1] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[2] Anshul Kundaje. Blacklisted genomic regions for functional genomics analysis. https://sites.google.com/site/anshulkundaje/projects/blacklists.

[3] E. L. James. *Fifty Shades of Grey*. Arrow, 2012.

[4] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, fourth edition edition, 2002.

[5] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986.

[6] H. Pages. Infrastructure for Biostrings-based genome data packages. http://www.bioconductor.org/packages/release/bioc/html/BSgenome.html.

[7] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, June 2002.

[8] W. J. Kent. UCSC genome browser utilities download. http://hgdownload.cse.ucsc.edu/admin/exe/.