GOTHiC - Genome Organisation Through HiC User Manual

Bori Mifsud and Robert Sugar

Monday 19th August, 2013

1 Introduction

Chromosome conformation capture (3C) is a cross-link based technique to detect spatial proximity of specific genomic distant loci [1], which has sparked a number of large-scale methods to study multiple interactions simultaneously [2]. Hi-C is a genome wide, unbiased version of 3C [3], in which all interactions are sampled by paired-end high-throughput sequencing. In an ideal Hi-C library all read-pairs would reflect real interactions and the number of read-pairs between any two genomic loci would be proportional to the frequency of the interaction in the studied cell population. Unfortunately, Hi-C has been shown to suffer from a multitude of systematic biases that confound the true signal, and complex approaches have been designed to minimize their impact [4, 5].

2 Description

2.1 Data

The data were generated by Lieberman-Aiden *et al.* and published in 2009 [3]. In this study, the authors used lymphoblastoid human cells and two different restriction enzymes (HindIII and NcoI) in the HiC experiments. The examples in this package use the mapped reads (from *Bowtie* or *HiCUP* [6], [7]) deposited in the HiCDataLymphoblastoid package, and paired reads (from our *pairReads* function) of the HindIII replicate SRR027956 for chromosome 20. The paired reads are in a *GRangesList* object.

```
> library(GOTHiC)
> data(lymphoid_chr20_paired_filtered)
```

2.2 A binomial statistic for accurately modelling Hi-C data

2.2.1 Method

First, self-ligation read-pairs and those mapping to adjacent fragments are filtered out by removing any pair of reads whose middle points are closer than 10kb (the median size of the fragments is about 2130bp in our experiment, therefore the 10kb cut-off allows for removal of read pairs resulting from incomplete

digestion as well). PCR duplicates are filtered by removing read-pairs with both ends mapping to identical coordinates (HiCUP already filters invalid read-pairs and removes duplicates).

Having done so, true interactions can be separated from spurious ligations using a binomial test. Briefly, the probability that a read pair is the consequence of a spurious ligation between two sites can be estimated as:

$$p_{j,h} = 2 * rel\_coverage_j * rel\_coverage_h \tag{1}$$

where the relative coverage of a given site or region is

$$rel\_coverage_j = reads_j / reads_{total} \tag{2}$$

and the noise fraction represents the fraction of the ligations corresponding to spurious inter-molecular ligations (having previously removed self-ligations).

The observed coverage is a complex function of a multitude of biases, including the density of restriction sites, cleavage efficiency, ligation efficiency, amplification and sequencing biases and mappability. The probability of seeing a random read-pair between two regions depends on the relative coverages of the interacting regions in a multiplicative manner since the biases affect both interacting partners independently.

Given a total number of read-pairs in the experiment of N (after excluding self-ligation read-pairs), the probability of observing $n_{j,h}$ read-pairs between two loci is given by the binomial density:

$$pval_{jh} = P(x \geq n_{jh}) = 1 - \sum_{i=0}^{n_{jh}-1} \binom{N}{i} (P_{randomintermol_{jh}})^i (1 - P_{randomintermol_{jh}})^{N-i} \tag{3}$$

This allows calculating a p-value of the chance of seeing the observed number of read-pairs by chance, as a simple function of the coverage of both sites and the total number of reads. Correcting for multiple testing using the Benjamini-Hochberg multiple testing correction we obtain a q-value that can be used directly to infer significant interactions with a desired false discovery rate from a Hi-C dataset.

In addition to a p-value, an observed-over-expected ratio can be easily calculated, what can be used as a measure of effect size or as an unbiased measure of interaction frequency or strength.

$$Obs\_exp\_ratio = n_{j,h} / (p_{j,h} * N) \tag{4}$$

The binomial method does not explicitly assume that different biases are multiplicative since it does not model the total effect of all biases as a product of their relative contribution. However, the method assumes that biases affect both read ends independently. This is a reasonable assumption given our understanding of known biases and has been used in previous methods (e.g. [4,5]).

Thus for any two genomic fragments, the binomial test provides two statistics: (i) a p-value of observing an interaction between the two loci given a random distribution of ligations; and (ii) a measure of the strength of the interaction compared with random expectation.

2.3 Function description

2.3.1 Pairing aligned paired NGS reads

The *pairReads* function takes bowtie output files, pairs the reads, only keeps those where both ends mapped, filters for perfect duplicates to avoid PCR bias, and saves and returns a *GenomicRangesList* object that contains the paired_reads_1 and paired_reads_2 *GenomicRanges* with the genomic coordinates of each end of the read pairs respectively.

```
> dirPath = system.file("extdata", package="HiCDataLymphoblast")
> fileName1 = list.files(dirPath, full.names=TRUE)[1]
> fileName2 = list.files(dirPath, full.names=TRUE)[2]
> paired = pairReads(fileName1, fileName2, sampleName="lymphoid_chr20",
+ DUPLICATETHRESHOLD = 1, fileType="Table")
```

2.3.2 Mapping aligned and paired reads to the restriction fragments

The *mapReadsToRestrictionSites* function takes mapped paired NGS reads in the format of a GenomicRangesList object where the two end of the reads are in the *GenomicRanges*paired_reads_1 and paired_reads_2. It prepares the digestion file from the genome supplied to it with the given restriction enzyme and specificity and maps the reads to the fragments. It returns a *GenomicRangesList* containing two textitGenomicRanges objects containing the start of the fragment where each end of the read pairs is mapped respectively (parallel version of findOverlaps from the GenomicRanges was adapted from Kasper Daniel Hansen).

```
> data(lymphoid_chr20_paired_filtered)
> mapped=mapReadsToRestrictionSites(filtered, sampleName="lymphoid_chr20",
+ BSgenomeName="BSgenome.Hsapiens.UCSC.hg18",
+ genome=BSgenome.Hsapiens.UCSC.hg18,
+ restrictionSite="A^AGCTT", enzyme="HindIII", parallel=FALSE, cores=1)
```

2.3.3 Binomial Test for detecting significant interactions in Hi-C data

The *GOTHiC* or *GOTHiChicup* functions perform a cumulative binomial test to detect interactions between distal genomic loci that have significantly more reads than expected by chance in Hi-C experiments. They take either mapped paired NGS reads, or the mapped and filtered reads from HiCUP as input and give back the list of significant interactions for a given bin size in the genome. They return a data.frame where each line contains the genomic coordinates of each en of a pair (chromosome and start of the bin), the observed number of reads, the relative coverage of each region, the expected frequency to observe this interaction by random, the binomial p-value and q-value (Benjamini correction of the p-value). They also generate a density plot from the p-value distribution Figure 1, which helps with assessing the quality of the HiC experiments. Poor quality experiments tend to have a p-value distribution that

is close to the one from a random sample (uniform 0 to 1). Good quality experiments have a clear peak around 0 (very low coverage samples can also have bimodal distribution even if the quality is lower, the number of identified significant interaction is an additional indicator of quality that would be low in low coverage samples).

```
> dirPath = system.file("extdata", package="HiCDataLymphoblast")
> fileName1 = list.files(dirPath, full.names=TRUE)[1]
> fileName2 = list.files(dirPath, full.names=TRUE)[2]
> binom=GOTHiC(fileName1,fileName2, sampleName="lymphoid_chr20",
+ res=1000000, BSgenomeName="BSgenome.Hsapiens.UCSC.hg18",
+ genome=BSgenome.Hsapiens.UCSC.hg18,
+ restrictionSite="A^AGCTT", enzyme="HindIII" ,cistrans="all", filterdist=10000,
+ DUPLICATETHRESHOLD=1, fileType="Table", parallel=FALSE, cores=NULL)

> dirPath <- system.file("extdata", package="HiCDataLymphoblast")
> fileName <- list.files(dirPath, full.names=TRUE)[4]
> restrictionFile <- list.files(dirPath, full.names=TRUE)[3]
> binom=GOTHiChicup(fileName, sampleName='lymphoid_chr20', res=1000000,
+ restrictionFile, cistrans='all', parallel=FALSE, cores=NULL)
```
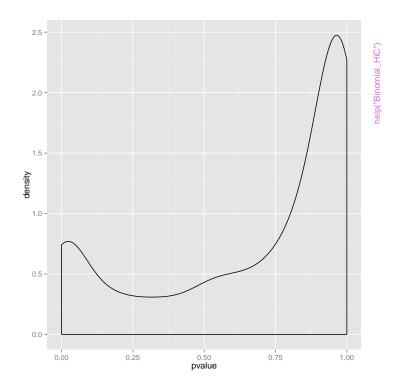
Figure 1: **P-value distribution of human chromosome 20 interactions in lymphoblastoid cells.**

2.4 Aknowledgements

We are grateful to Elodie Darbo for helping with putting the script together in a package format, and Philip East for testing and providing useful comments.

References

[1] Dekker J, Rippe K and Dekker M and Kleckner N: **Capturing chromosome conformation.** *Science* 2002, **295**:1306–11.

[2] de Wit E and de Laat W: **A decade of 3C technologies: insights into nuclear organization.** *Genes & development* 2012 **26** (1):11–24.

[3] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo P, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke Andreas, Stamatoyannopoulos J, Mirny LA, Lander ES and Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009 **326**:289–93.

[4] Yaffe E and Tanay A: **Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture.** *Nature genetics* 2011 **43** (11):1059–65.

[5] Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J and Mirny LA: **Iterative correction of Hi-C data reveals hallmarks of chromosome organization.** *Nature methods* 2012 **9** (10):999–1003.

[6] Langmead B, Trapnell C, Pop M and Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biology* 2009 **10**:R25.

[7] Wingett et al.: **http://www.bioinformatics.babraham.ac.uk/projects/hicup/overview/**.