HiTC - Exploration of High Throughput ’C’ experiments

N. Servant, B. Lajoie, E. Nora, L. Giorgetti, J. Dekker, E. Heard, E. Barillot

July 16, 2012

1 Introduction

Chromosome Capture Conformation (3C) was first introduced by Dekker et al. (2002) ten years ago. The 3C technique aims in detecting physical contact between pairs of genomic loci and is now widely used to detect cis and trans interactions between genes and regulatory elements. The development of the 3C-based techniques has changed our vision of the nulcear oragnization (see de Wit and de Laat (2012) for a review).
With the development of high throughput analyses, and in particular second-generation sequencing, the 3C has been adapted to study in parallel physical interactions between many loci, and thus increase the scale at which interactions between genomic loci can be detected (4C - Circular 3C, Simonis et al. (2006), Zhao et al. (2006); 5C - 3C Carbone Copy, Dostie et al. (2006)). More recently, this technique was further extended to obtain detailed insights into the general three-dimensional arrangements of complete genomes (Hi-C, Lieberman-Aiden et al. (2009)).

While the use of high-throuput ’C’ techniques is expected to increase in the coming years, it also creates some new statistical and bioinformatics challenges. In this way, publicly available bioinformatics tools, as well as clear analysis strategy are still lacking. The my5C web browser was proposed by Lajoie et al. (2009) to visualise, transform and analyze 5C data. However, the my5C webtool is targeted to end-users and biologists to prepare their 5C experiments and to handle their data but is not dedicated to the development of new statistical algorithms.

Here we present the *HiTC* R package which has been developed to offer a bioinformatic environment to explore high-troughput ’C’ data. One advantage of this package is that it operates within the open source Bioconductor framework, and thus, offers new opportunities for futur development in this field. The current version of the package provides visualisation, transformation and normalisation functions as described in Lajoie et al. (2009). Our goal is also to provide a flexible basis for further development, aiming at the integration of new analysis algorithm that are being developped (Yaffe and Tanay (2011))

2 Getting started

This document briefly describes how to use the *HiTC* R package. The package is built on the functionality of Bioconductor packages such as *Girafe* and *GenomeIntervals*, and

provides a new class and methods to handle with high-throughput 'C' data. It is especially suited to 5C and Hi-C data handling, but can also in principle be used for 4C, though specific needs of 4C users may be best met by *r3Cseq* R package.
Even if the 5C and Hi-C approaches are derived from the same 3C technique, strong differences in their protocol can also be noticed. While 5C enables analysis of interactions between many loci, it also required an extensive number of primers, which is not suitable for a genome-wide analysis as the Hi-C. Thus, the pre-processing of these two types of data is totally different with, for instance, two different mapping strategies.

The current version of the *HiTC* package was developped to work on processed 5C, Hi-C or other high-throughput 3C data.
The *HTCexp* (High-Throughput 'C' experiment) object was defined as :

- An interaction map (i.e a *matrix*)
- Two *Genome_intervals* objects that describe each features of the interaction matrix, respectively, the x (i.e. columns) and y (i.e. rows) labels of the interaction matrix. Basically, in the context of 5C, these objects will be the forward and reverse primers, and for the Hi-C the binned genome intervals.

```
> library(HiTC)
> showClass("HTCexp")
```

```
Class "HTCexp" [package "HiTC"]

Slots:

Name:          intdata               xgi              ygi
Class:          matrix Genome_intervals Genome_intervals
```

A complete dataset is composed of a list of cis and trans *HTCexp* objects, characterised by their physical interactions.

The *HiTC* package includes two distinct dataset.
The first one is a 5C dataset (GSE35721) published by Nora et al. (2012), from which we choose two different Mouse samples, male undifferentiated ES cells (E14, GSM873935) and male embryonic fibroblasts (MEF, GSM873924). This dataset is mainly used to describe the available functionalities of the package.
The second is the Human Hi-C dataset (GSE18199) published by Lieberman-Aiden et al. (2009). The interaction map of chromosome 14 is used to illustrate the capabilities of the *HiTC* package to explore Hi-C data.

3 Load Data

The *HiTC* R package is fully compatible with the my5C web browser. The interaction counts matrices can be imported from a table or a list file and exported. The description of the genomic intervals has to be imported in the BED format (see section 9 for an example). In addition, objects can be easily created using the `HTCexp` method. The function `readBED`

is also proposed to load multi-track BED files.
For the following example, the processed 5C data of the chromosome X from Nora et al. (2012) are already included in the package.

```
> data(Nora_5C)
> ## List of HTCexp objects describing the 5C dataset
> detail(E14)
```

```
$chrXchrX
HTC object
Focus on genomic region [chrX:98831149-103425150]
CIS Interaction Map
Matrix of Interaction data: [486-504]
504  genome intervals from 'xgi' object
486  genome intervals from 'ygi' object
```

4 Quality Control

The first step after data pre-procesing is a quality control to check weither the data are likely to reflect cis and/or trans chromosomal interactions rather than just random collisions. Quality control for the percentage of reads aligned to interchromosomal and intrachromosomal interactions is available, as well as distribution of the interaction frequency against the genomic distance between two loci, and simple statistics (see Figure 1).

```
> CQC(E14)
```

```
TOTAL READS = 1e+06
NUMBER OF INTERACTIONS =  182617
AVERAGE FREQUENCY =  5.476
MEDIAN FREQUENCY =  1.043981
------------------------------
CIS READS = 1e+06
NUMBER OF CIS INTERACTIONS =  182617
------------------------------
TRANS READS = 0
NUMBER OF TRANS INTERACTIONS =  0
```

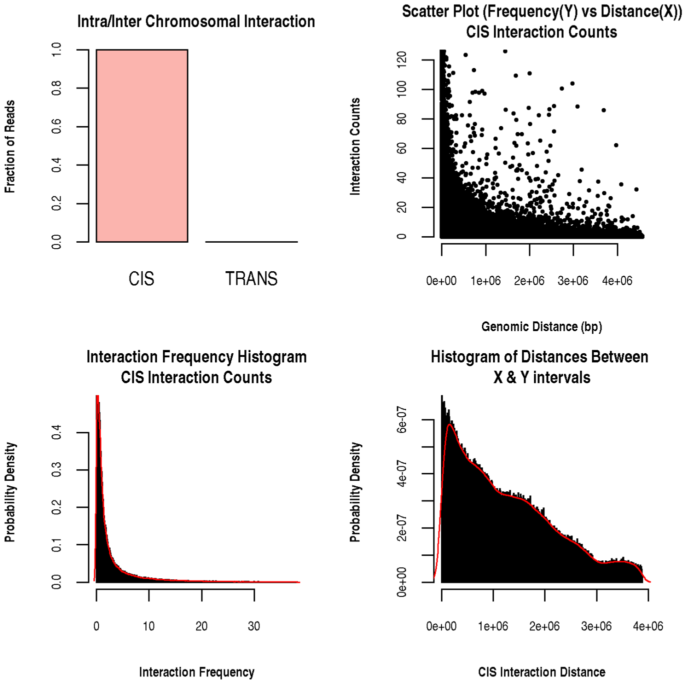

Figure 1: *Quality Control of 5C data. Top-left : proportion of inter/intra chromosomal interactions. Top-right : scatter-plot of interaction counts versus genomic distance between two loci. Bottom-rigth : histogram of interaction counts. Bottom-left : histogram of distances between two loci.*

5 Visualisation of Interaction Maps

The interaction map represents the frequency at which each pair of restriction fragments have been ligated together during the 3C procedure. The goal is to visualise at once these counts for many pairs of restriction fragments across a large genomic region. Each entry in the matrix corresponds to a count information, i.e., number of times two restriction fragments have been sequenced as a pair.
Therefore Hi-C and 5C results are typically displayed using two dimensional heatmaps. The mapC function proposes a list of options to play with data visualisation, such as contrast, color, or trimming. Two different views are provided; a square heatmap view (see Figure 2) or a triangle view. The latest is particulary useful for interaction maps comparison and alignment with genomic or epigenomic features.

```
> mapC(E14$chrXchrX)
```

```
[1] "minrange= 0.052199  - maxrange= 38.157505"
```

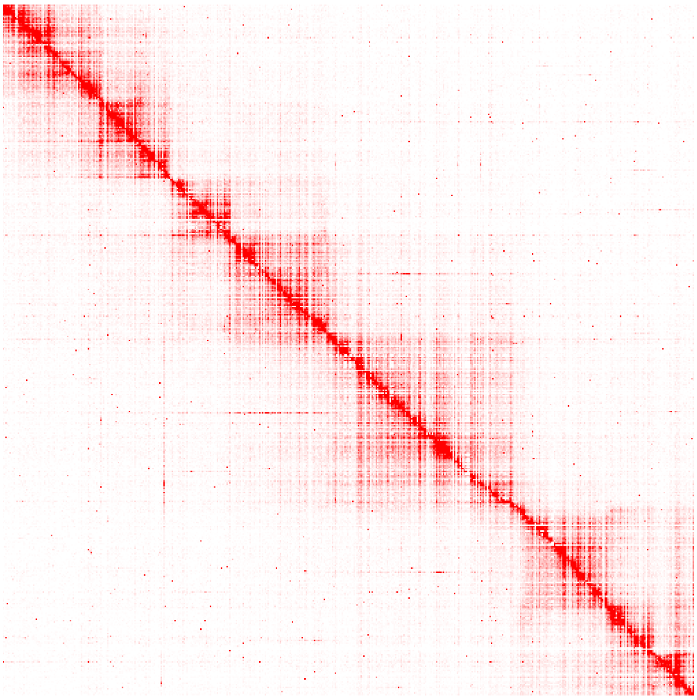

Figure 2: *5C interaction map of chromosome X.*

6 Data Transformation

6.1 Windowing

Each pixel of an interaction map can correspond either to a single restriction fragment, several restriction fragments or genomic intervals of any given size (and therefore various restriction fragment numbers). 5C allows assessing interaction frequencies for each pair of restriction fragments. The Hi-C protocol, on contrary, does not necessarily yields counts for every single pair of restriction fragments, especially when working with large genomes. Results are thus typically displayed for genomic bins of an arbitrary size.
To produce an interaction map, the genomic range of the display should be divided into appropriately size loci. This size depends on the resolution desired for the analysis. For instance, 5C data can be visualised at the primers resolution, or segmented into 100Kb or 1Mb bins that can be partially overlap or not. Such binned interaction map is symmetrical around the diagonal. For the following example, we decided to focus on a subset of the original dataset (see Figure 3).

```
> ## Focus on a subset chrX:100295000:102250000
> E14subset<-extractRegion(E14$chrXchrX, from=100295000, to=102250000)
> ## Binning of 5C interaction map
> E14.subset.binned <- binningC(E14subset, binsize=100000, step=3)
```

```
Bin size=102420 [3x34140]
```

```
> mapC(E14.subset.binned)
```

```
[1] "minrange= 0.104398  - maxrange= 21.977888"
```

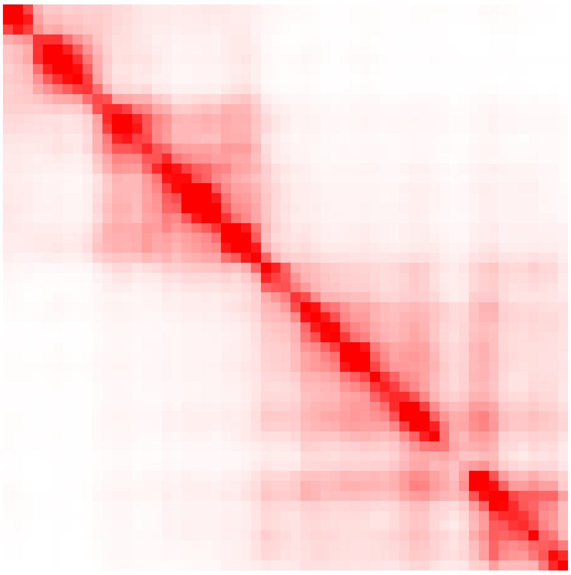

Figure 3: *Binned 5C interaction map of chrX:100295000-102250000.*

6.2 Data Normalisation

Due to the polymer nature of chromatin, at small genomic distances, pairs of restriction fragments that are close to each other in the linear genome will give higher signal than fragments that are further apart. Such property leads to strongest counts falling on the heatmap diagonal. When considering any given pair of restriction fragments, it is therefore informative to assess whether the observed counts are above what is expected given the genomic distance that separate them.
Different ways of normalisation have been proposed. In the current version of the package we propose to estimate the expected interaction counts as presented in Bau et al. (2011). The expected value is the interaction frequency between two loci that one would expect based on a sole dependency on the genomic proximity of these fragments in the linear genome. This can be estimated using a Loess regression model (see Figure 4).

```
> ## Look at exptected counts
> E14exp <- getExpectedCounts(E14subset, stdev=TRUE, plot=TRUE)
```

```
Loess fit ...
Standard deviation calculation ...
Delta= 9695.895
Calculating stdev for 1 9625 19291 38770 77526 155131 310261 620517 1241068 1939180 bps
```

Interaction frequencies can be then normalised for distance by dividing the observed value by the expected value (`normPerExpected`), or by calculating the zscore at each estimated points (`normPerZscore`). These normalisation methods can be easily applied using the methods `normPerReads`, `normPerExpected` or `normPerZscore` (see Figure 5).

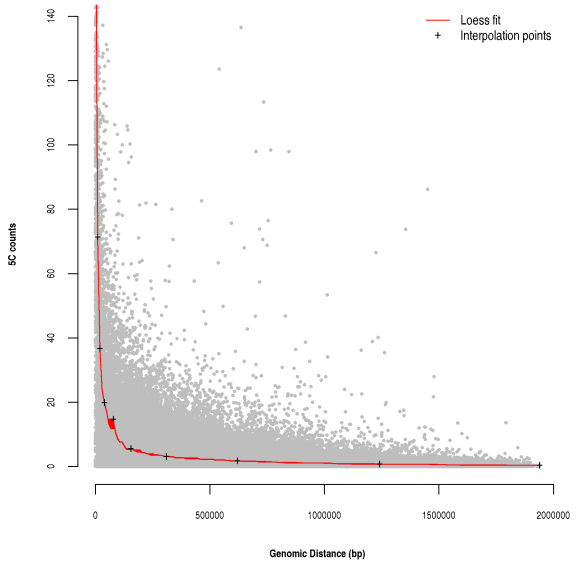

Figure 4: *Estimation of expected count using a Loess smoothing. The crosses represent the interpolation points.*

```
> E14subsetz <- normPerZscore(E14subset)
```

```
Loess fit ...
Standard deviation calculation ...
Delta= 9695.895
Calculating stdev for 1 9625 19291 38770 77526 155131 310261 620517 1241068 1939180 bps
```

```
> E14subsetz.binned <- binningC(E14subsetz, binsize=50000, step=3)
```

```
Bin size=51210 [3x17070]
```

```
> mapC(E14subsetz.binned)
```

```
[1] "minrange= 0.004854  - maxrange= 1.019847"
```

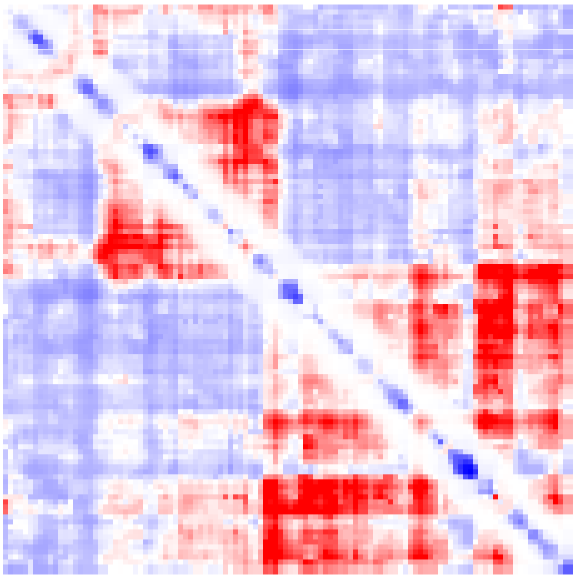

Figure 5: *Interaction map of data normalised using the Zscore approach.*

7 Annotation of Interaction Maps

The *HiTC* package contains functions for visualising genomic regions with interaction maps (see Figure 6). The annotation objects have to belong to the *Genome_intervals* class, cand can be loaded from BED files.

For instance, the following example displays the CTCF enriched regions (Kagey et al. (2010)) and RefSeq genes over the interaction map of the E14 sample.

```
> E14.binned <- binningC(E14$chrXchrX, binsize=100000, step=3)
```

```
Bin size=102087 [3x34029]
```

```
> exDir <- system.file("extdata", package="HiTC")
> Refgene <- readBED(file.path(exDir,"refseq_mm9_chrX_98831149_103425150.bed"))
```

```
Reading 1 tracks from refseq_mm9_chrX_98831149_103425150.bed
```

```
> CTCF <- readBED(file.path(exDir,"CTCF_chrX_98892125_102969775.bed"))
```

```
Reading 1 tracks from CTCF_chrX_98892125_102969775.bed
```

```
> mapC(E14.binned,
+      giblocs=list(RefSeqGene=Refgene$Refseq_Gene, CTCF=CTCF$CTCF),
+      maxrange=10, view=2)
```

```
[1] "minrange= 0.0261  - maxrange= 10"
```

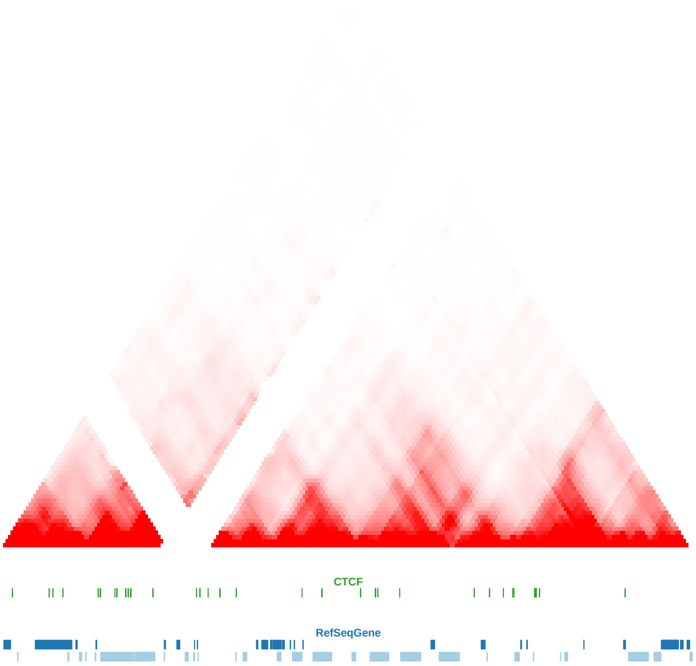

Figure 6: *Visualisation of interaction map and genomic annotations.*

8 Comparison of HTCexp objects

The *HiTC* package provides methods to perform simple operations on *HTCexp*, such as dividing, substracting two objects or extracting a genomic region.
It also proposes a graphical view to compare two 'C' experiments. In the following example, the MEF sample is compared to the E14 sample (see Figure 7).

```
> MEF.binned <- binningC(MEF$chrXchrX, binsize=100000, step=3)
```

```
Bin size=102087 [3x34029]
```

```
> mapC(E14.binned, MEF.binned,
+      giblocs=list(RefSeqGene=Refgene$Refseq_Gene, CTCF=CTCF$CTCF),
+      maxrange=10)
```

```
[1] "minrange= 0.0261  - maxrange= 10"
[1] "minrange= 0.018445  - maxrange= 10"
```

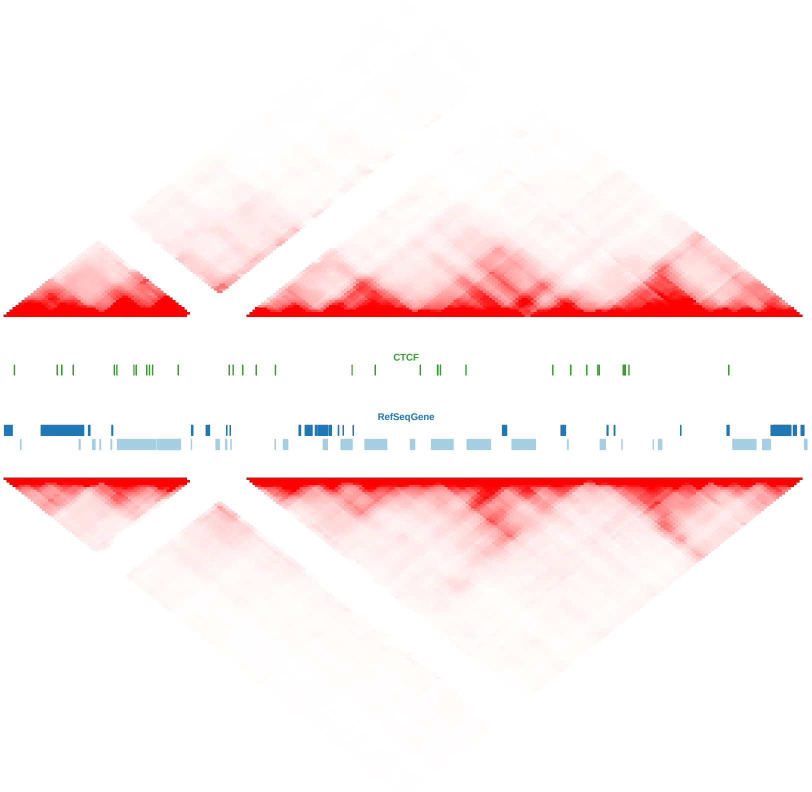

Figure 7: *Comparison of two binned interaction maps, and visualisation with genomic annotations.*

9 Application to Hi-C data

Basically, 5C and Hi-C data can be described in the same way. Thus, most of the functions described for the 5C data can be applied to the Hi-C data.
In this section, we present how, using a few command lines, we can reproduce some analyses of the Lieberman-Aiden et al. (2009) paper (see Figures 8-10).
The binned (1Mb) counts matrix of the chromosome 14 was downloaded from GEO (GSE18199).

```
> ## Load Dekker et al. Chromosome 14 data (from GEO GSE18199)
> hiC<-import.my5C(file.path(exDir,"HIC_gm06690_chr14_chr14_1000000_obs.txt"),
+                  xgi.bed=file.path(exDir,"GSE18199_gm06690_chr14_chr14_1Mb.bed"))
```

```
Reading 1 tracks from GSE18199_gm06690_chr14_chr14_1Mb.bed
Intervals files loaded
Convert my5C matrix file in HiTC object(s)
```

```
> hiC <- hiC$chr14chr14
> detail(hiC)
```

```
HTC object
Focus on genomic region [chr14:1-106368584]
```

```
Matrix of Interaction data: [107-107]
107  genome intervals from 'xgi' object
107  genome intervals from 'ygi' object
Total Reads = 755139
Number of Interactions =  7913
Median Frequency =  30
```

```
> ## Extract region of interest and plot the interaction map
> hiC <- extractRegion(hiC,from=1.8e+07, to=106368584)
> mapC(hiC, maxrange=100)
```

```
[1] "minrange= 1  - maxrange= 100"
```

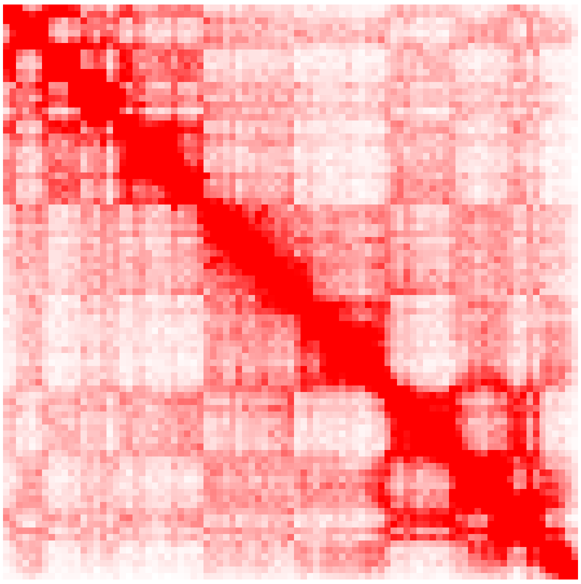

Figure 8: *Hi-C interaction map of chromosome 14*

```
> ## Data Normalisation by Expected number of Counts
> hiCnorm <- normPerExpected(hiC)
```

```
Loess fit ...
```

```
> mapC(hiCnorm, log.data=TRUE)
```

```
[1] "minrange= 0.014971  - maxrange= 2.518636"
```

```
> ## Correlation Map of Chromosome 14
> mapC(cor(intdata(hiCnorm)), maxrange=1, minrange=-1,
+      col.pos=c("black", NA, "red"), col.neg=c("black",NA, "blue"))
```

```
[1] "minrange= -1  - maxrange= 1"
```

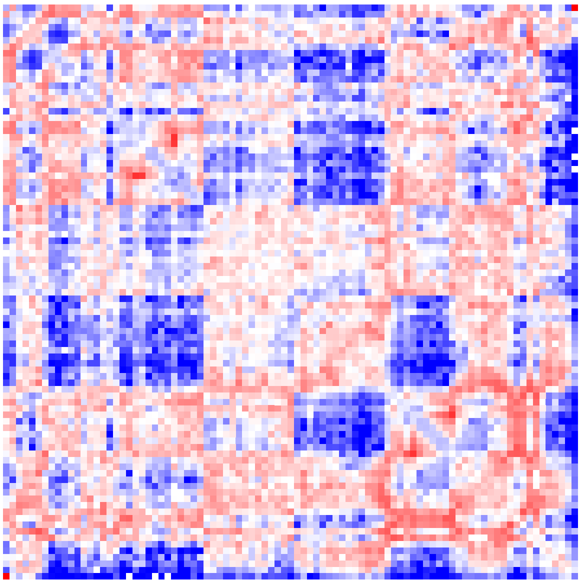

Figure 9: *Interaction map of data normalised by the expected interaction counts*

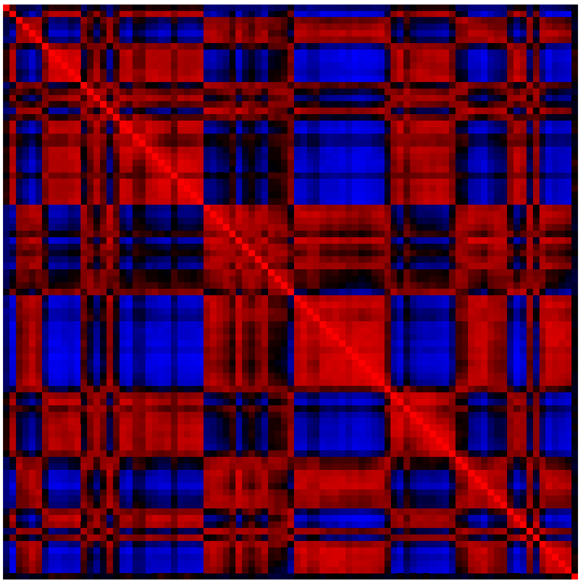

Figure 10: *Correlation map of chromosome 14*

10 A word about speed

For improving the run time on machines with multiple processors, some of the functions in the *HiTC* package have been implemented to make use of the functionality in the *multicore* package. If *multicore* has been attached and initialised before calling these functions, some functions will make use of `mclapply` instead of the normal `lapply`.

Package versions

This vignette was generated using the following package versions:

- R Under development (unstable) (2012-06-24 r59622), `x86_64-unknown-linux-gnu`
- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, utils
- Other packages: BiocGenerics 0.3.0, Biostrings 2.25.6, genomeIntervals 1.13.2, GenomicRanges 1.9.31, girafe 1.9.0, HiTC 1.1.1, intervals 0.13.3, IRanges 1.15.20, lattice 0.20-6, latticeExtra 0.6-19, RColorBrewer 1.0-5, Rsamtools 1.9.21, ShortRead 1.15.9
- Loaded via a namespace (and not attached): Biobase 2.17.6, bitops 1.0-4.1, BSgenome 1.25.3, hwriter 1.3, stats4 2.16.0, tools 2.16.0, zlibbioc 1.3.0

Acknowledgements

Many thanks to Joern Toedling and Pierre Gestraud for useful discussion and help in developping this R package. Special thanks to Julien Gagneur and Joern Toedling for writing the *genomeIntervals* and *girafe* packages.

References

D. Bau, A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, J. Dekker, and M. A. Marti-Renom. The three-dimensional folding of the Îś-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol*, 18(1):107–114, Jan 2011. doi: 10.1038/nsmb.1936. URL `http://dx.doi.org/10.1038/nsmb.1936`. 6

E. de Wit and W. de Laat. A decade of 3c technologies: insights into nuclear organization. *Genes Dev*, 26(1):11–24, Jan 2012. doi: 10.1101/gad.179804.111. URL `http://dx.doi.org/10.1101/gad.179804.111`. 1

J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, Feb 2002. doi: 10.1126/science.1067799. URL `http://dx.doi.org/10.1126/science.1067799`. 1

J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome Res*, 16(10):1299–1309, Oct 2006. doi: 10.1101/gr.5571506. URL `http://dx.doi.org/10.1101/gr.5571506`. 1

M. H. Kagey, J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. A. Young. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, Sep 2010. doi: 10.1038/nature09380. URL `http://dx.doi.org/10.1038/nature09380`. 8

B. R. Lajoie, N. L. van Berkum, A. Sanyal, and J. Dekker. My5c: web tools for chromosome conformation capture studies. *Nat Methods*, 6(10):690–691, Oct 2009. doi: 10.1038/nmeth1009-690. URL `http://dx.doi.org/10.1038/nmeth1009-690`. 1

E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct 2009. doi: 10.1126/science.1181369. URL `http://dx.doi.org/10.1126/science.1181369`. 1, 2, 10

E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. BlÃijthgen, J. Dekker, and E. Heard. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, Apr 2012. doi: 10.1038/nature11049. URL http://dx.doi.org/10.1038/nature11049.

M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nat Genet*, 38(11):1348–1354, Nov 2006. doi: 10.1038/ng1896. URL http://dx.doi.org/10.1038/ng1896. 1

E. Yaffe and A. Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, 43(11):1059–1065, Nov 2011. doi: 10.1038/ng.947. URL http://dx.doi.org/10.1038/ng.947. 1

Z. Zhao, G. Tavoosidana, M. SjÃűlinder, A. GÃűndÃűr, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, and R. Ohlsson. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, 38(11):1341–1347, Nov 2006. doi: 10.1038/ng1891. URL http://dx.doi.org/10.1038/ng1891.