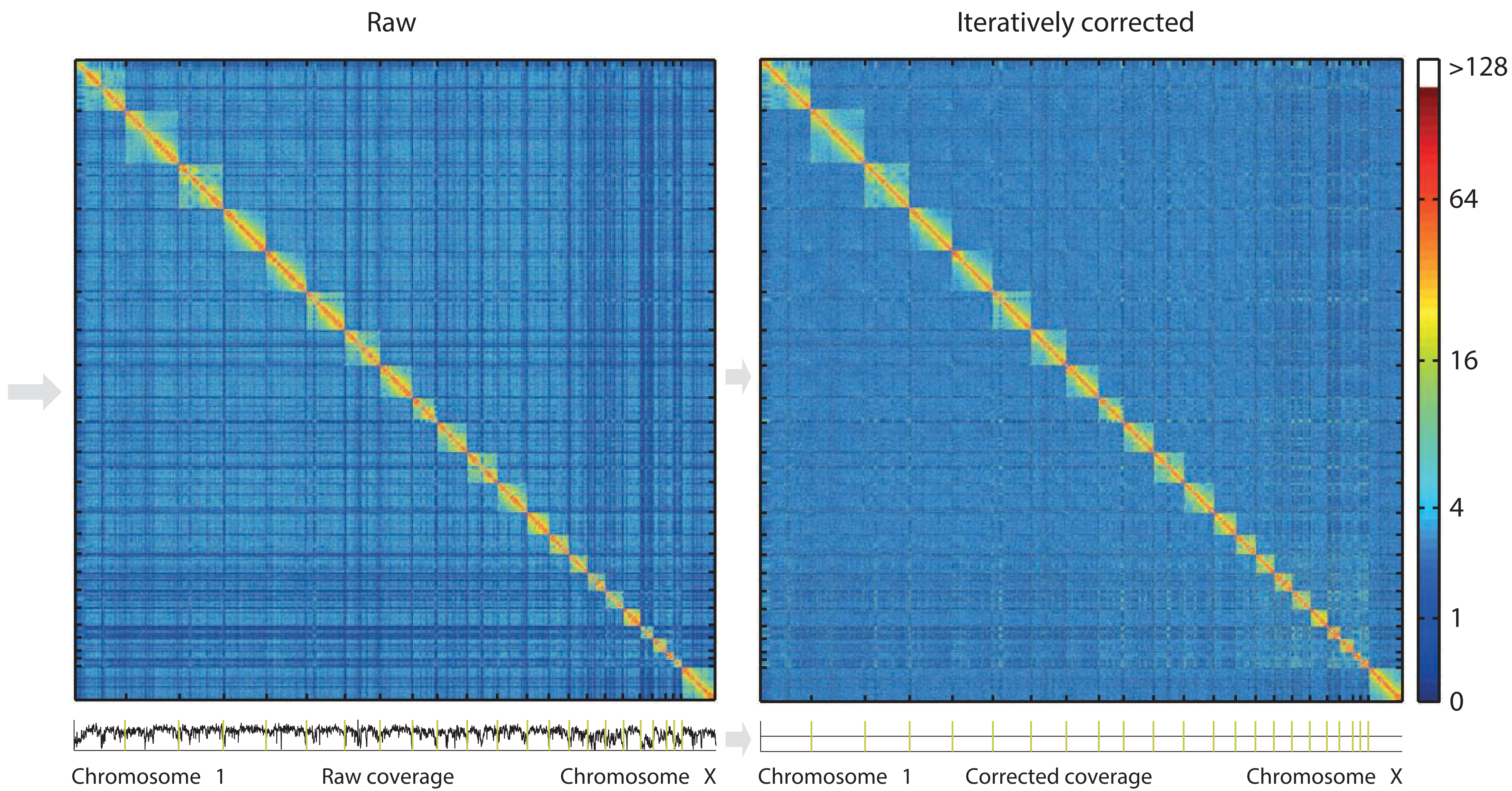
3DAROC16

Normalization & Comparison

Marco Di Stefano, François Serra & Marc A. Marti-Renom
Structural Genomics Group (CNAG-CRG)

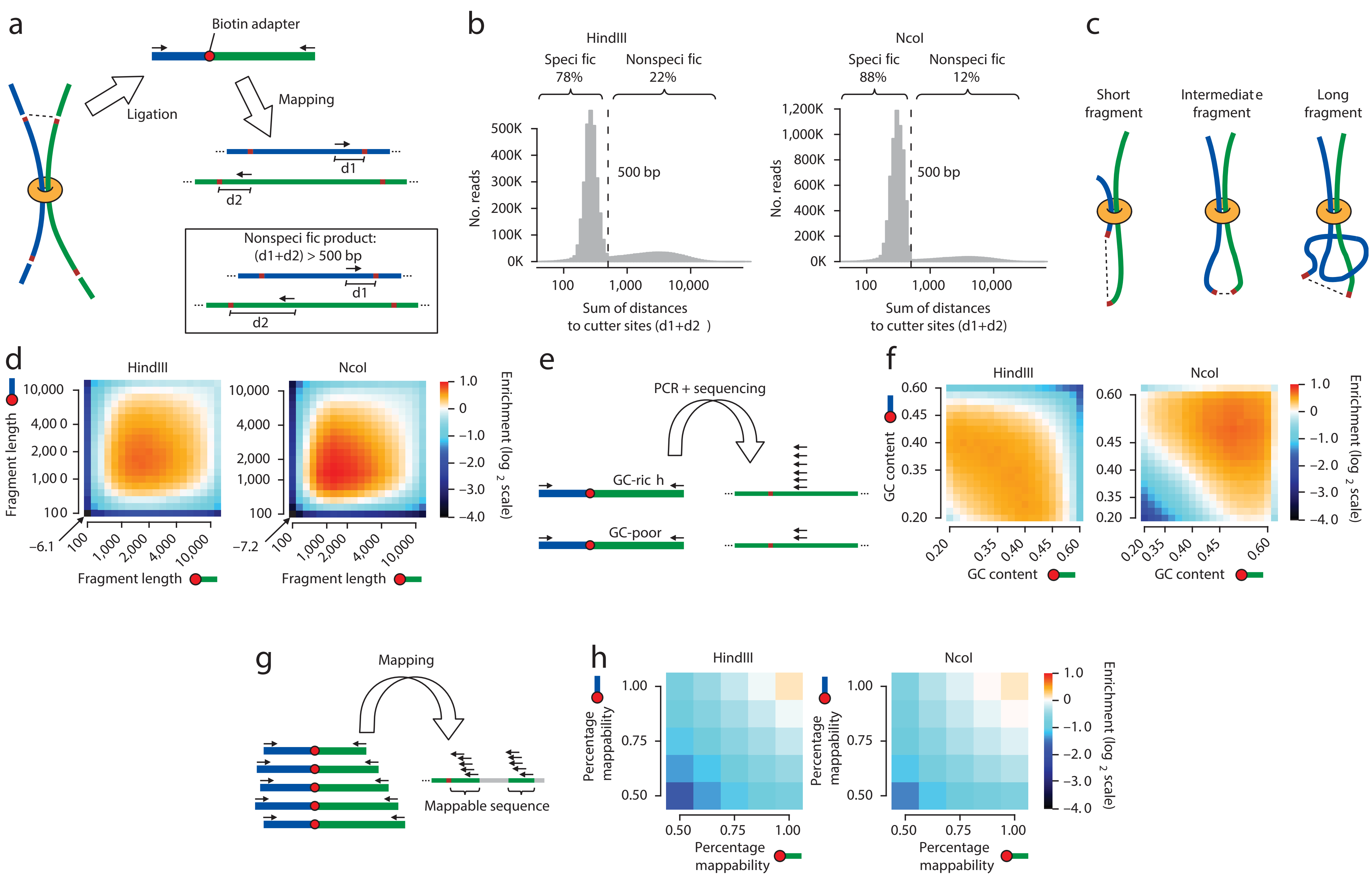
DISCLAIMER

Normalizing HiC data



Normalizing HiC data (a la Tanay)

Yaffe, E., & Tanay, A. (2011). Nature Genetics, 43(11), 1059–1065

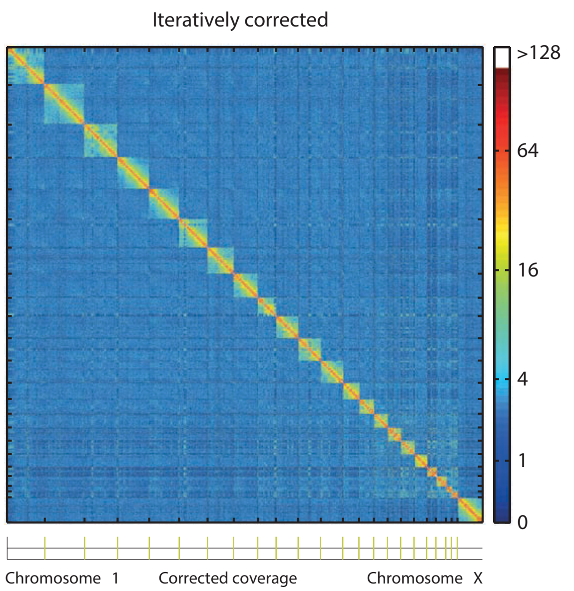


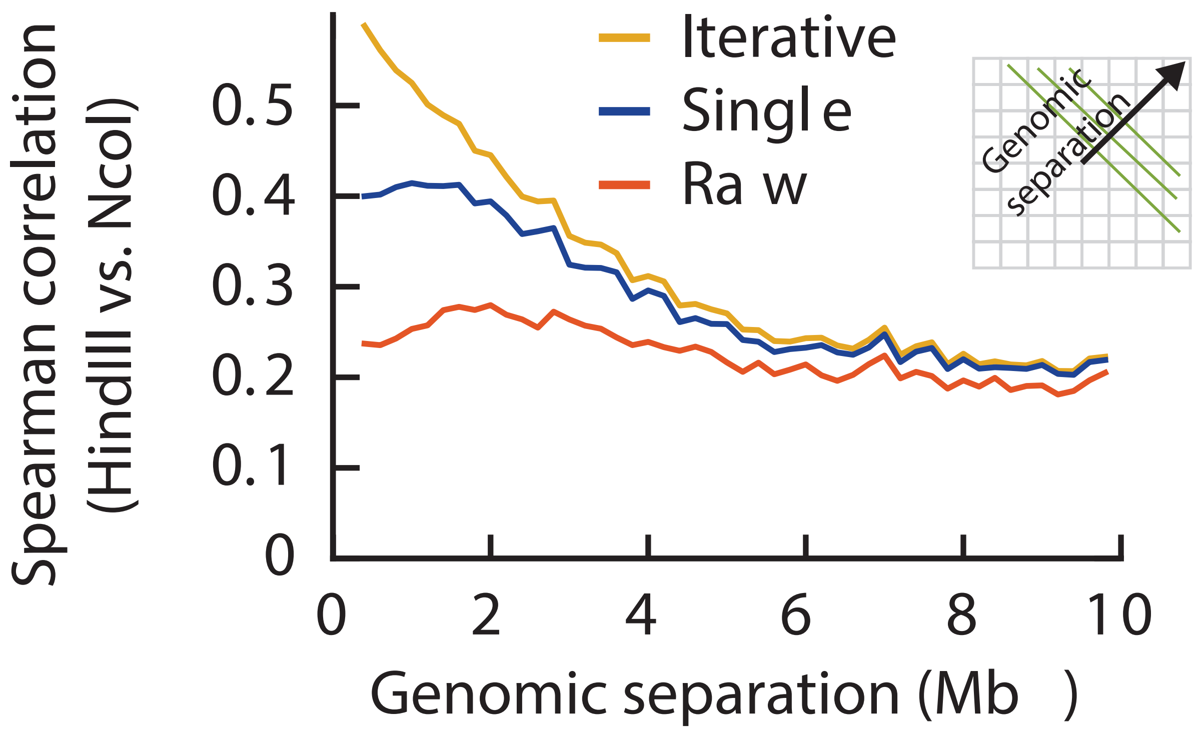
Normalizing HiC data (a la Mirny)

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., et al. (2012). Nature Methods, 9(10), 999–1003.

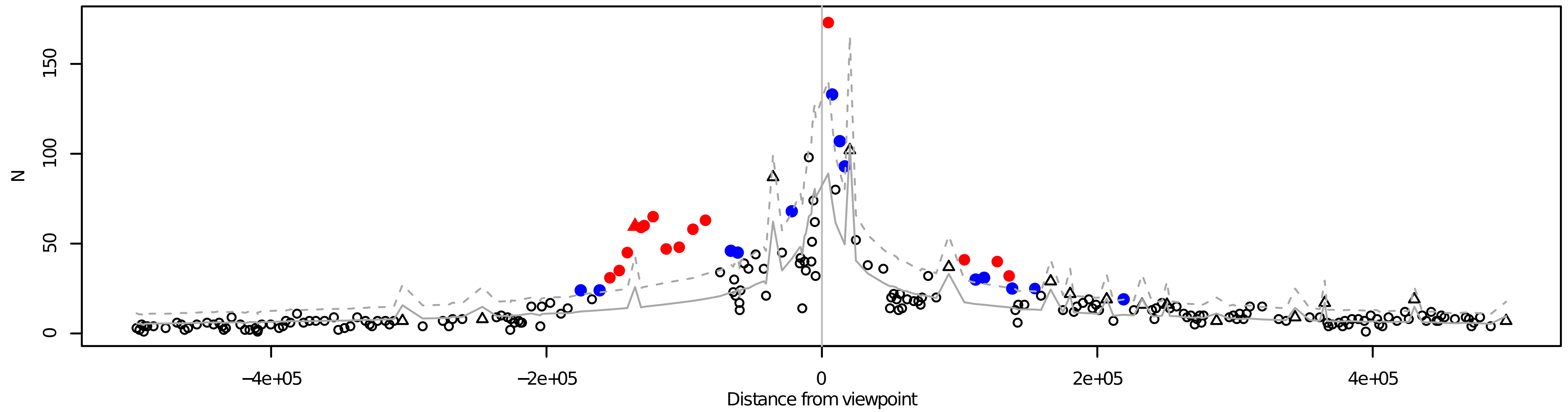
$$O_{ij} = B_i B_j T_{ij}$$

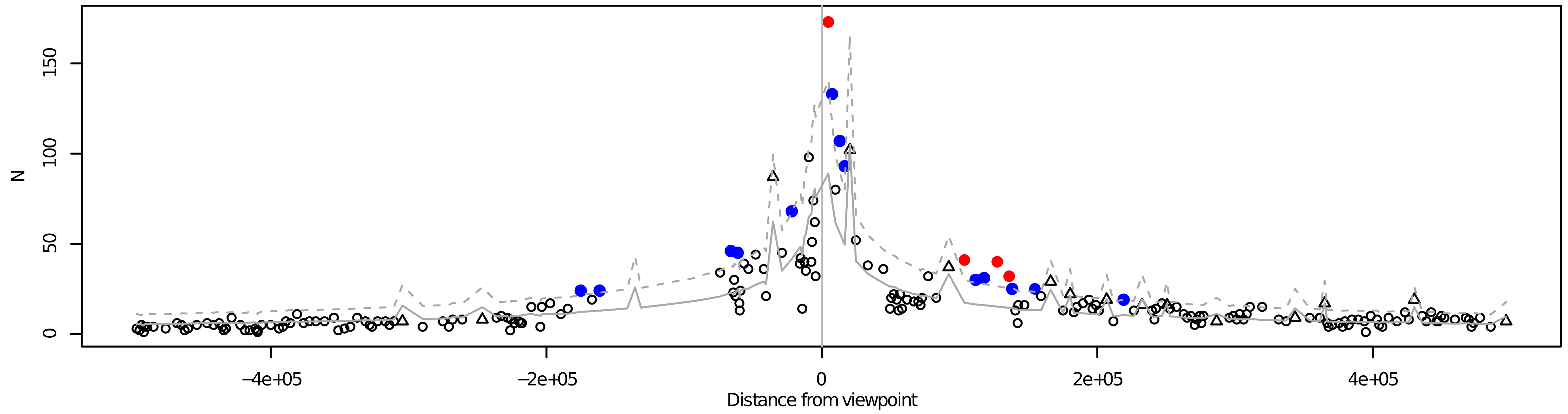
$$\Sigma_{i=1,|i-j|>1}^{N} T_{ij} = 1$$



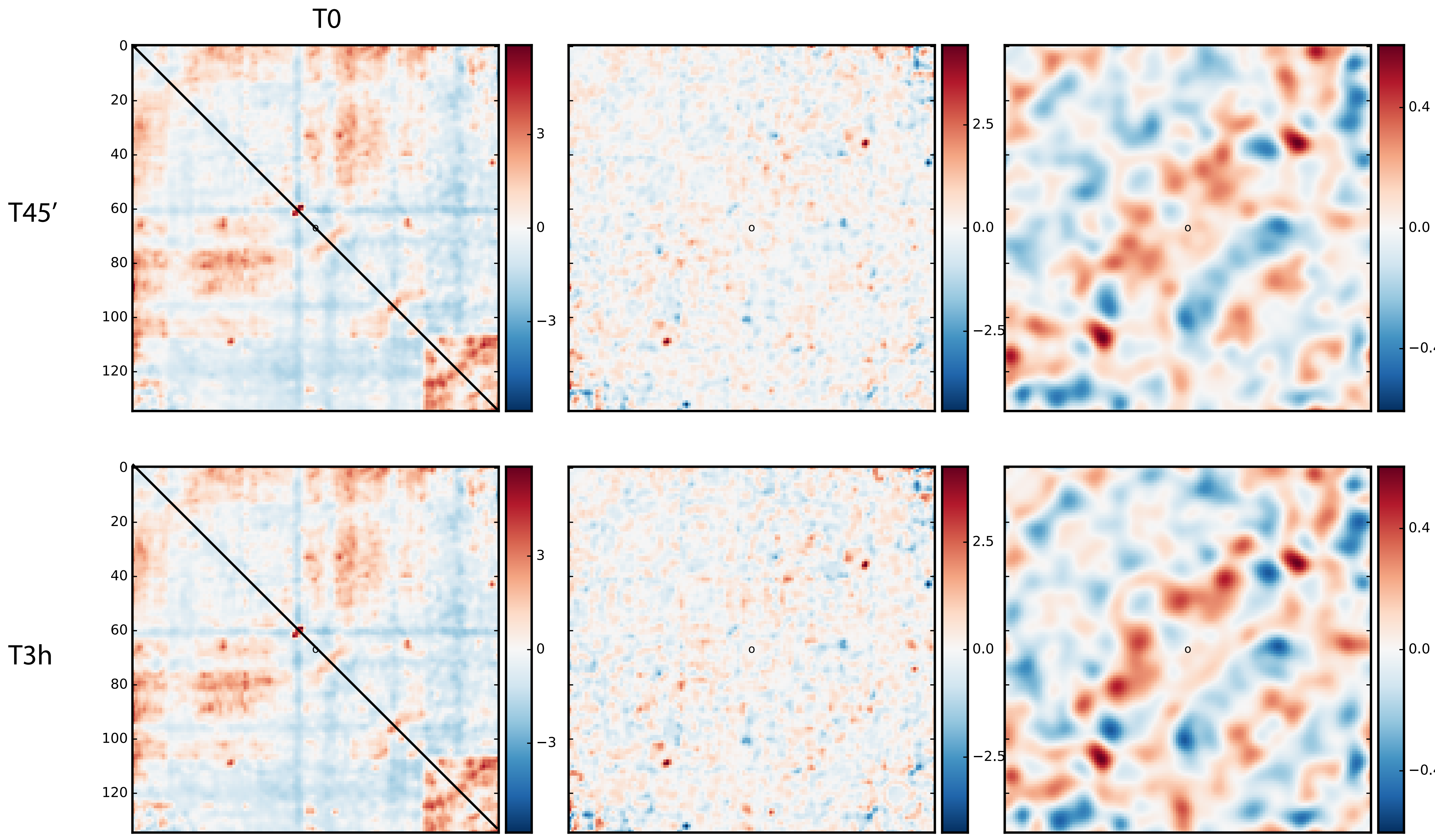


Comparing HiC data





Z-score differences (DekkerLab)



Comparing HiC data (GOTHIC)

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., et al. (2015). *Nature Genetics*, 1–12.

Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C

Borbala Mifsud[1,2,10], Filipe Tavares-Cadete[1,9], Alice N Young[3,10], Robert Sugar[1], Stefan Schoenfelder[3], Lauren Ferreira[3], Steven W Wingett[4], Simon Andrews[4], William Grey[5], Philip A Ewels[3], Bram Herman[6], Scott Happe[6], Andy Higgs[6], Emily LeProust[6,9], George A Follows[7], Peter Fraser[3], Nicholas M Luscombe[1,2,8] & Cameron S Osborne[3,5]

Transcriptional control in large genomes often requires looping interactions between distal DNA elements, such as enhancers and target promoters. Current chromosome conformation capture techniques do not offer sufficiently high resolution to interrogate these regulatory interactions on a genomic scale. Here we use Capture Hi-C (CHi-C), an adapted genome conformation assay, to examine the long-range interactions of almost 22,000 promoters in 2 human blood cell types. We identify over 1.6 million shared and cell type–restricted interactions spanning hundreds of kilobases between promoters and distal loci. Transcriptionally active genes contact enhancer-like elements, whereas transcriptionally inactive genes interact with previously uncharacterized elements marked by repressive features that may act as long-range silencers. Finally, we show that interacting loci are enriched for disease-associated SNPs, suggesting how distal mutations may disrupt the regulation of relevant genes. This study provides new insights and accessible tools to dissect the regulatory interactions that underlie normal and aberrant gene regulation.

Genome organization influences transcriptional regulation by facilitating interactions between gene promoters and distal regulatory elements. Many contacts have been identified using chromosome conformation capture methodologies[1–3]. For example, the ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) method has been used to map long-range interactions extending over hundreds of kilobases; however, these studies have only interrogated the subset of interactions involving highly transcriptionally active genes, whereas long-range interactions for weakly expressed and transcriptionally inactive genes remain unknown. Although the 5C (chromatin conformation capture carbon copy) method is not restricted by the nature of interactions, thus far, it has only been applied to a few small genomic regions. The Hi-C method simultaneously captures all genomic interactions, which provides a population-average snapshot of the genome conformation within a single experiment[4]; yet, owing to the enormous complexity of Hi-C libraries, it is costly to sequence to sufficient depth to provide enough spatial resolution to interrogate specific contacts between gene promoters and distal regulatory elements[5,6]. To circumvent these issues, we have used solution hybridization selection, originally developed for exon sequencing[7]—and recently used to capture the interactions of a few hundred promoters from 3C libraries[8]—to enrich Hi-C libraries for genome-wide, long-range contacts of both active and inactive promoters.

RESULTS

A genome-wide, long-range interaction capture assay

We prepared three HindIII-digested Hi-C libraries from GM12878 cells, a human Epstein-Barr virus (EBV)-transformed lymphoblastoid cell line that has been comprehensively assayed in the Encyclopedia of DNA Elements (ENCODE) Project, and two libraries from *ex vivo* CD34$^+$ hematopoietic progenitor cells. One Hi-C library from each cell type was sequenced to examine the di-tag (paired-end read) interaction distribution and depth of read coverage (**Supplementary Table 1**). As anticipated, we observed a higher density of di-tag interaction reads between restriction fragments in *cis* as compared with fragments in *trans*, with the highest density occurring between fragments separated by less than 20 kb (**Supplementary Fig. 1a**,**b**). We also observed demarcation of the genome into distinct contiguous, highly intraconnected topologically associated domains (TADs)[5] (**Supplementary Fig. 1c** and **Supplementary Table 2**). The distribution of read coverage was typical for a Hi-C experiment. In our initial comparison, we downsampled all data sets to 45 million unique sequencing reads. Each restriction fragment was represented by an average of 143 and 139 reads in the GM12878 and CD34$^+$ libraries, respectively (**Supplementary Fig. 1d**). We processed the reads using binomial statistics to identify ligation fragments that were significantly enriched ($q < 0.05$). This approach recognizes ligation products between

[1]The Francis Crick Institute, London, UK. [2]UCL Genetics Institute, University College London, London, UK. [3]Nuclear Dynamics Programme, Babraham Institute, Cambridge, UK. [4]Bioinformatics Group, Babraham Institute, Cambridge, UK. [5]Department of Medical and Molecular Genetics, King's College London School of Medicine, London, UK. [6]Diagnostics and Genomics Division, Agilent Technologies, Santa Clara, California, USA. [7]Department of Haematology, Cambridge University Hospitals National Health Service (NHS) Foundation Trust, Cambridge, UK. [8]Okinawa Institute of Science and Technology, Okinawa, Japan. [9]Present addresses: Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan (F.T.-C.) and Twist Bioscience, San Francisco, California, USA (E.L.). [10]These authors contributed equally to this work. Correspondence should be addressed to C.S.O. (cameron.osborne@kcl.ac.uk) or N.M.L. (nicholas.luscombe@ucl.ac.uk).

Received 5 December 2014; accepted 2 April 2015; published online 4 May 2015; doi:10.1038/ng.3286

Comparing HiC data (CHICAGO)

Cairns, J., Freire-Pritchett, P., Wingett, S. W., Várnai, C., Dimond, A., Plagnol, V., et al. (2016). *Genome Biology*, 1–17.

Cairns *et al. Genome Biology* (2016) 17:127
DOI 10.1186/s13059-016-0992-2

Genome Biology

METHOD Open Access

CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data

Jonathan Cairns[1†], Paula Freire-Pritchett[1†], Steven W. Wingett[1,2], Csilla Várnai[1], Andrew Dimond[1], Vincent Plagnol[3], Daniel Zerbino[4], Stefan Schoenfelder[1], Biola-Maria Javierre[1], Cameron Osborne[5], Peter Fraser[1] and Mikhail Spivakov[1*]

Abstract

Capture Hi-C (CHi-C) is a method for profiling chromosomal interactions involving targeted regions of interest, such as gene promoters, globally and at high resolution. Signal detection in CHi-C data involves a number of statistical challenges that are not observed when using other Hi-C-like techniques. We present a background model and algorithms for normalisation and multiple testing that are specifically adapted to CHi-C experiments. We implement these procedures in CHiCAGO (http://regulatorygenomicsgroup.org/chicago), an open-source package for robust interaction detection in CHi-C. We validate CHiCAGO by showing that promoter-interacting regions detected with this method are enriched for regulatory features and disease-associated SNPs.

Keywords: Gene regulation, Nuclear organisation, Promoter-enhancer interactions, Capture Hi-C, Convolution background model, *P* value weighting

Background

Chromosome conformation capture (3C) technology has revolutionised the analysis of nuclear organisation, leading to important insights into gene regulation [1]. While the original 3C protocol tested interactions between a single pair of candidate regions ("one vs one"), subsequent efforts focused on increasing the throughput of this technology (4C, "one vs all"; 5C, "many vs many"), culminating in the development of Hi-C, a method that interrogated the whole nuclear interactome ("all vs all") [1, 2]. The extremely large number of possible pairwise interactions in Hi-C samples, however, imposes limitations on the realistically achievable sequencing depth at individual interactions, leading to reduced sensitivity. The recently developed Capture Hi-C (CHi-C) technology uses sequence capture to enrich Hi-C material for multiple genomic regions of interest (hereafter referred to as "baits"), making it possible to profile the global interaction profiles of many thousands of regions globally ("many vs all") and at a high resolution (Fig. 1) [3–7].

CHi-C data possess statistical properties that set them apart from other 3C/4C/Hi-C-like methods. First, in contrast to traditional Hi-C or 5C, baits in CHi-C comprise a subset of restriction fragments, while any fragment in the genome can be detected on the "other end" of an interaction. This asymmetry of CHi-C interaction matrices is not accounted for by the normalisation procedures developed for traditional Hi-C and 5C [8–10]. Secondly, CHi-C baits, but not other ends, have a further source of bias associated with uneven capture efficiency. In addition, the need for detecting interactions globally and at a single-fragment resolution creates specific multiple testing challenges that are less pronounced with binned Hi-C data or the more focused 4C and 5C assays, which involve fewer interaction tests. Finally, CHi-C designs such as Promoter CHi-C and HiCap [3–5, 11] involve large numbers (many thousands) of spatially dispersed baits. This presents the opportunity to increase the robustness of signal detection by sharing information across baits. Such sharing is impossible in the analysis of 4C data that focuses on only a single bait and is of limited use in 4C-seq containing a small number of baits [12–14].

These distinct features of CHi-C data have prompted us to develop a bespoke statistical model and a

* Correspondence: mikhail.spivakov@babraham.ac.uk
†Equal contributors
[1]Nuclear Dynamics Programme, Babraham Institute, Cambridge, UK
Full list of author information is available at the end of the article

Comparing HiC data (diffHiC)

Lun, A. T. L., & Smyth, G. K. (2015). *BMC Bioinformatics*, 1–11.

Lun and Smyth *BMC Bioinformatics* (2015) 16:258
DOI 10.1186/s12859-015-0683-0

BMC Bioinformatics

SOFTWARE Open Access

CrossMark

diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data

Aaron T.L. Lun[1,2] and Gordon K. Smyth[1,3*]

Abstract

Background: Chromatin conformation capture with high-throughput sequencing (Hi-C) is a technique that measures the *in vivo* intensity of interactions between all pairs of loci in the genome. Most conventional analyses of Hi-C data focus on the detection of statistically significant interactions. However, an alternative strategy involves identifying significant changes in the interaction intensity (i.e., differential interactions) between two or more biological conditions. This is more statistically rigorous and may provide more biologically relevant results.

Results: Here, we present the diffHic software package for the detection of differential interactions from Hi-C data. diffHic provides methods for read pair alignment and processing, counting into bin pairs, filtering out low-abundance events and normalization of trended or CNV-driven biases. It uses the statistical framework of the edgeR package to model biological variability and to test for significant differences between conditions. Several options for the visualization of results are also included. The use of diffHic is demonstrated with real Hi-C data sets. Performance against existing methods is also evaluated with simulated data.

Conclusions: On real data, diffHic is able to successfully detect interactions with significant differences in intensity between biological conditions. It also compares favourably to existing software tools on simulated data sets. These results suggest that diffHic is a viable approach for differential analyses of Hi-C data.

Keywords: Hi-C, Genomic interaction, Differential analysis

Background

Chromatin conformation capture with high-throughput sequencing (Hi-C) is a technique that is widely used to study global chromatin organization *in vivo* [1]. Briefly, samples of nuclear DNA are cross-linked and digested with a restriction enzyme to release chromatin complexes into solution (Fig. 1). Each complex may contain multiple restriction fragments, corresponding to an interaction between the associated genomic loci. After some processing, proximity ligation is performed between the ends of the restriction fragments. This favours ligation between restriction fragments in the same complex. The ligated DNA is sheared and purified for high-throughput paired-end sequencing. Each sequencing fragment represents a ligation product, such that each read in the pair originates from a different genomic locus. The intensity of an interaction between a pair of genomic loci can be quantified as the number of read pairs with one read mapped to each locus. The output from the Hi-C procedure spans the genome-by-genome "interaction space" whereby all pairwise interactions between loci can potentially be detected. As such, careful analysis is required to draw meaningful biological conclusions from this type of data.

Most analyses of Hi-C data have focused on identifying "significant" interactions from a single sample [2, 3]. This is challenging because non-specific ligation and apparent interactions can arise from a variety of uninteresting technical causes and rigorous analysis requires a precise quantitative understanding of these artifacts. Identifying biologically interesting interactions from a single sample requires elaborate modeling of the background signal in Hi-C experiments in order to correct for systematic biases due to GC content, mappability and fragment length [3]. Such modeling inevitably involves assumptions and approximations. Furthermore, the interaction space

*Correspondence: smyth@wehi.edu.au
[1]The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, Melbourne, Australia
[3]Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Melbourne, Australia
Full list of author information is available at the end of the article

cnag CRG