GOTHiC, a simple probabilistic model to resolve complex biases and to identify real interactions in Hi-C data

Borbala Mifsud[1,2†], Inigo Martincorena[1†‡], Elodie Darbo[1], Robert Sugar[1], Stefan Schoenfelder[3], Peter Fraser[3*] and Nicholas M. Luscombe[124*]

Affiliations:

[1] The Francis Crick Institute, London WC2A 3LY, UK.

[2] UCL Genetics Institute, University College London, London WC1E 6BT, UK

[3] Nuclear Dynamics Programme, Babraham Institute, Cambridge CB22 3AT, UK.

[4] Okinawa Institute of Science & Technology, Okinawa 904-0495, Japan.

*Correspondence to: nicholas.luscombe@ucl.ac.uk

and peter.fraser@babraham.ac.uk

† These authors contributed equally.

‡ Present address: Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK.

1. Abstract

Hi-C is one of the main methods for investigating spatial co-localisation of DNA in the nucleus. However, the raw sequencing data obtained from Hi-C experiments suffer from large biases and spurious contacts, making it difficult to identify true interactions. Existing methods use complex models to account for biases and do not provide a significance threshold for detecting interactions. Here we introduce a simple binomial probabilistic model that resolves complex biases and distinguishes between true and false interactions. The model corrects biases of known and unknown origin and yields a p-value for each interaction, providing a reliable threshold based on significance. We demonstrate this experimentally by testing the method against a random ligation dataset. Our method outperforms previous methods and provides a statistical framework for further data analysis, such as comparisons of Hi-C interactions between different conditions. GOTHiC is available as a user-friendly BioConductor package (http://www.bioconductor.org/packages/release/bioc/html/GOTHiC.html).

2. Introduction

Hi-C is a high-throughput technique based on chromosome conformation capture to detect the spatial proximity between pairs of genomic loci[1,2]. It is now routinely used to study the three-dimensional folding of genomes[3-7]. In theory, a sequenced Hi-C read-pair should directly represent an interaction between two loci, with the number of mapped read-pairs corresponding to the frequency of interactions in the sample cell population. However, two challenges must be resolved in order to extract the true signal from Hi-C data.

The first is to identify and resolve systematic biases. Hi-C datasets present many effects common to high-throughput sequencing experiments, for instance amplification biases due to differences in sequence composition across the genome. There are also biases that are specific to Hi-C. For example, variations in the density of restriction sites cause large differences in genomic fragment sizes: longer fragments are more likely to self-ligate, whereas very short fragments are difficult to ligate; both tend to be under-represented in the sequencing library[8]. The complex combination of known and unknown biases cause over- and under-representation of chromosomal regions when a Hi-C dataset is mapped to the reference genome. Thus, the number of observed read-pairs do not directly reflect the frequency of interactions between two genomic loci.

The second challenge is to distinguish between true and false interactions. As depicted in Box 1A, a Hi-C library contains three types of read-pairs. (i) The first represents real interactions in which the ligation reaction occurs between the ends of a pair of crosslinked DNA fragments. (ii) The second corresponds to spurious self-ligations in which the ends of the same DNA fragment are ligated together; as the two ends of read-pairs map to the same DNA fragment, they are easily filtered using a minimum genomic distance between them. (iii) The third represents spurious ligations between two non-crosslinked DNA fragments; read-pairs from these reactions are problematic as they are indistinguishable from those arising through real interactions. The proportions of read-pairs representing real and spurious interactions can vary widely depending on the quality of the sample and library preparations, but it is not unusual to encounter Hi-C datasets in which 20-40% of read-pairs originate from self-ligations.

Two main computational methods have been proposed to deal with biases. The earlier, *hicpipe*, applies a multiplicative model to estimate the probabilities of interactions between two genomic regions as a function of mappability, fragment length and GC content; the numbers of mapped read-pairs are then normalised according to these estimates[8]. A later method, *hiclib*, proposes that the total bias is represented in the sequence coverage as the product of individual biases for each pair of genomic regions. Starting with the assumption that every genomic region should have identical coverages, the method iteratively normalises the original coverage until it becomes uniform along the whole genome[9]. The former method considers known sources of biases, whereas the latter also deals with unknown ones.

Although both methods have been frequently used, they and other subsequent methods exhibit several practical limitations[10,13]. First, the assumptions behind the methods are untested against experimental control data and so their success in eliminating biases is unclear. Second, the issue of distinguishing between true and random interactions remains unresolved. Finally, the software encoding these methods have several dependencies that make them technically demanding to install and operate, and they require 2-16 hours of dedicated server time to process a moderately sized Hi-C dataset.

Here, we introduce a straightforward binomial model that corrects the complex combination of known and unknown biases in Hi-C data (Box 1B). The model calculates accurately the probabilities that the observed number of read-pairs are due to random ligations, and yields a list of statistically significant interactions between pairs of genomic loci. We tested the model using random ligation controls, demonstrating that the method gives high levels of specificity. GOTHiC, the accompanying BioConductor package, is fast, accurate and easy to use.

3. A binomial model for Hi-C data

For a given pair of genomic loci, GOTHiC calculates: (i) the probability of observing a given number of read-pairs between two loci through random ligations; and (ii) the effect size or "strength" of interaction measured as the ratio of observed-over-expected numbers of interactions. GOTHiC assumes that the observed sequence coverage varies as a function of multiple known and unknown biases, including the density of restriction sites, cleavage efficiency, ligation efficiency, amplification and sequencing biases, and mappability. It assumes that the biases affect each end of read-pairs independently; thus the probability of observing a randomly occuring read-pair between two loci is modelled as the product of the relative coverages in the interacting loci. This is a reasonable assumption given our understanding of known biases[8,9]; the advantage of modelling the combined effect of biases is that it incorporates unknown sources and that it is robust against future variants of Hi-C methods.

First, self-ligations and incomplete digestion products are removed by filtering read-pairs mapping to the same fragment and within a specified distance of each other on the genome (default=10kb). Given the relative coverage of two genomic loci, *j* and *h*, the probability of a spurious read-pair linking the two loci can be calculated as:

$$p_{j,h} = 2r_j r_h f_{random}$$

r_j, the relative coverage of a locus, is calculated as:

$$r_j = \frac{reads_j}{2N}$$

where $reads_j$ is the mapped read count for genomic locus *j* and *N* is the total number of read-pairs in the dataset. f_{random} is the fraction of read-pairs in the Hi-C library arising from spurious ligations. Although f_{random} could be estimated experimentally or computationally, in practise this may often be difficult and a conservative upperbound for $p_{j,h}$ can be obtained by excluding this term.

Given the probability of a read linking the two loci, the probability of observing *n* or more read-pairs between them by chance in a dataset of *N* reads, is given by the binomial cumulative density:

$$pval_{j,h} = P(x \geq n_{j,h}) = 1 - \sum_{i=0}^{n_{j,h}-1} \binom{N}{i} (p_{j,h})^i (1 - p_{j,h})^{N-i}$$

This yields a p-value for each interaction as a function of the coverage of both loci and the total number of reads in the experiment. Using the Benjamini-Hochberg multiple-testing correction (with L*(L-1)/2 tests, where L is the number of loci investigated), we

obtain a q-value that can be used directly to identify statistically significant interactions at a pre-defined false discovery rate.

The log of observed-over-expected ratio (R) can be used as a measure of effect size or as a normalised measure of interaction frequency.

$$R_{j,h} = log_2 \frac{n_{j,h}}{p_{j,h}N}$$

4. How well does GOTHiC perform?

4.1 Assessing performance using a Hi-C dataset and random ligation control

To assess performance, we applied GOTHiC to two datasets generated from the same mouse fetal liver cell sample: (i) one produced using the standard Hi-C protocol and (ii) another containing only randomly ligated read-pairs. The latter was produced by reversing the cross-links before the ligation step and it is analogous to an "input" control that is commonly used for background correction in ChIP-seq studies. As expected in a random control, 93-95% of read-pairs occur between loci on different chromosomes, in contrast to 20-40% of read-pairs in Hi-C datasets.

Read coverage is highly variable across the genome (Figure 1A): it correlates well with previously reported effects of GC content, mappability and restriction-site density, though not all variation is captured by these factors. The raw contact maps in Figure 1B emphasise how variations in sequence coverage affect the interpretation of unnormalised Hi-C data, in which regions of higher coverage ostensibly show stronger interactions and vice versa. Strikingly, the trend is apparent even in the random ligation control (blue arrow, right panel), which does not contain any true interactions. The high correlation in coverages between the real and random datasets (Pearson's r=0.99) indicates that virtually all of the variation in coverage observed in a Hi-C sample is explained by experimental biases.

The processed contact maps in Figure 1C show how effectively GOTHiC deals with these biases, as the patterns influenced by underlying variations in coverage are removed (left panel). GOTHiC also identifies statistically significant interactions with high specificity (red squares, left panel). There is good separation in log (observed/expected) values between "true" and "false" interactions (Figure 1D, top), which is also reflected in the distribution of p-values (middle panel). GOTHiC identified ~90,000 statistically significant interactions in the Hi-C dataset (FDR <5%). In contrast, GOTHiC calls almost no interactions in the random ligation experiment (Figure 1C, right panel). This dataset confirms the specificity of the binomial model and the accuracy of FDR estimates, as violations of the underlying assumptions should lead to a large number of false positives. In fact, GOTHiC calls just 22 false positive interactions in the random ligation dataset from more than 3 million tests; this means that the p-values accurately reflect the probability of observing a given number of reads between any two loci as a result of experimental biases.

In addition to calling statistically significant interactions, GOTHiC removes much of the underlying bias. Figure 1E demonstrates that the detection of significant interactions as well as the general ranking of interactions by their q-value is largely independent of coverage, as the proportion of significant interactions is stable across different coverage bins, and in each coverage bin the proportion of interactions falling into the different quartiles is near a quarter.

Alternatively to the q-value, the log-ratio, *R*, between the observed number of reads and the expected number of reads (log observed/expected) may be used as a normalised measure of interaction frequency. This value is similar to the log fold-change measure in differential expression analyses, and it would tend to show a high variance in regions of low coverage due to the low expected values, and the integer read counts, similarly to log fold-change of lowly expressed genes. However, the *R* value can be used for a dual cut-off to identify significant interactions above a desired effect size (as in volcano plots).

The output from GOTHiC can also be used to flag poor quality Hi-C libraries. We have observed that inadequate dilution or cross-linking can yield libraries with a high fraction of spurious read-pairs (*i.e.*, self-ligations and ligations between non-crosslinked fragments). As shown in the control dataset (Figure 1D), this will lead to more uniform distribution of p-values, as expected by chance, and GOTHiC will successfully control the false discovery rate, yielding a small number of significant interactions.

4.2 Reproducibility between replicates using different restriction enzymes

It has been shown that treating the same biological sample with different restriction enzymes can cause large differences in coverage[3]. To evaluate the performance of GOTHiC in these conditions, we applied it to previously published Hi-C datasets produced using HindIII and NcoI on a human lymphoblastoid cell line. These enzymes target distinct restriction motifs that are distributed differently along the genome; this results in different fragment densities, GC contents and mappability biases. Figure 2A highlights the remarkable impact on the coverage profiles and the raw contact maps (left and right panels, yellow highlighted boxes).

Despite these strong biases, GOTHiC outputs very consistent contact maps and statistically significant interactions (Figure 2B). Loci with very different numbers of read-pairs in the raw data are identified as interacting at similar significance levels after processing (Figure 2A and 2B, highlighted regions). We find 92,892 and 103,117 significant interactions in HindIII and NcoI experiments respectively, of which 80,500 overlap (Figure 2C), and the interaction rankings obtained from the two experiments show high correlation (Spearman's r=0.79) (Figure 2D).

5. Comparison with existing methods

Finally, in order to benchmark GOTHiC's performance, we applied the two main published methods, *hicpipe* and *hiclib*, to the mouse fetal liver and human lymphoblastoid Hi-C datasets (Figure 3).

As previously observed from the contact maps, the number of reads between two loci is strongly affected by the coverage of these loci (Figure 3C, boxplots in the top panel). Although the normalised interaction strength values from *hicpipe* and *hiclib* do not appear to show obvious biases in the contact maps (Figure 3A,B, S1A,B), more detailed assessment reveals that the outputs from both methods continue to suffer from coverage-dependent biases (Figure 3C, middle and bottom panels). The interaction strength measures are inversely correlated with coverage, suggesting overcorrection of the raw data – in other words, interactions in the 1st and 2nd quartiles for strength are enriched in the low coverage bins. In contrast for GOTHiC, both the significant interactions and the interactions ranked by q-values are are much less affected by coverage (Figure 1E, Figure S1C b).

Finally, we examined the overlap in interaction scores between the three methods (Figure 3D, Figure S1F). Interactions identified as significant by GOTHiC tend to be highly ranked by *hiclib* and *hicpipe*, indicating good agreement. Moreover, using the number of significant interactions returned by the binomial test of GOTHiC as a cut-off to select the top-ranked interactions returned by the other methods, revealed a very high overlap between all three methods (Figure S2). Thus, GOTHiC is at least as successful as existing methods in removing biases, but also provides significance values and a statistical framework for further analyses.

6. Discussion

Sequencing libraries produced by Hi-C experiments are noisy because of technical artifacts (self-ligations and random ligations) and complex biases caused by the intrinsic characteristics of the genome sequence (GC content, unequal distribution of restriction enzyme sites, uniqueness and mappability of the sequences). Here, we proposed a simple solution to analyze Hi-C data using a simple binomial test, which successfully removes artifacts and sequencing biases to detect true genomic interactions even in the noisiest Hi-C datasets.

GOTHiC's approach is simpler than existing methods, which require the identification and separate modeling of individual biases[8] or an iterative correction of biases[9,10]. It yields similar rankings to previous methods, with comparable or even slightly improved bias removal and reproducibility between replicates. Most importantly, unlike any othe method, GOTHiC calculates p-values that allows the identification of true genomic interactions and the removal of artefactual interactions with a well-controlled false discovery rate.

GOTHiC is implemented as an R package, which requires a mapped read file as input and returns a list of significant interactions. This implementation can analyze a whole-genome Hi-C dataset of 30 million uniquely mapped reads at 1Mb resolution in ~2 hours using a single core machine with ~200Mb memory, and can be several fold faster if run with the parallel option on more cores.

The sensitivity of the method could be further improved by estimating the fraction of inter-molecular ligations (f_{random}). Our use of an upperbound (f_{random}=1) provides a conservative estimate, ensures high specificity and should be preferred unless accurate information on the noise fraction across the genome is available.

Finally, we envisage that the simple probabilistic framework introduced here could be further expanded to other applications in Hi-C, such as combining replicates, or identifying interaction changes between conditions. Significance levels and observed/exptected ratios obtained from GOTHiC can be used as the basis for algorithms predicting the 3D structure of genomes[11], those finding topologically associated domains[12] or for those that estimate the confidence for an interaction as a function of the genomic distance separating two interacting regions[13].

7. References

1. Dekker, J. Capturing Chromosome Conformation. *Science* **295,** 1306–1311 (2002).
2. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes & Development* **26,** 11–24 (2012).
3. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions

Reveals Folding Principles of the Human Genome. *Science* **326,** 289–293 (2009).
4. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).
5. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503,** 290–294
6. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–1680
7. Chandra, T. *et al.* Global Reorganization of the Nuclear Landscape in Senescent Cells. *CellReports* **10,** 471–483 (2015).
8. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics* **43,** 1059–1065 (2011).
9. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Meth* **9,** 999–1003 (2012).
10. Li, W., Gong, K., Li, Q., Alber, F. & Zhou, X. J. Hi-Corrector: a fast, scalable andmemory-efficient package for normalizinglarge-scale Hi-C data. *Bioinformatics* 1–3 (2014). doi:10.1093/bioinformatics/btu747/-/DC1
11. Baù, D. *et al.* The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* **18,** 107–114 (2010).
12. Levy-Leduc, C., Delattre, M., Mary-Huard, T. & Robin, S. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30,** i386–i392 (2014).
13. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Research* **24,** 999–1011 (2014).

8. Figures

Box 1. Schematic overview of the binomial model
(**A**) After crosslinking and digesting the chromatin, the DNA is ligated resulting in three types of ligation products. In order to detect real interactions, we first filter out self-ligations. With the remaining paired-reads, we then calculate the relative coverage across the genome in order to estimate the random interaction probability. (**B**) We finally apply the binomial test to distinguish between random and real interactions.

Figure 1. GOTHiC applied to mouse fetal liver Hi-C experiments.
(A) From the top, distributions of the relative coverage, the GC content percentage, the mappability score and the number of fragments per 1Mb (y-axis) across mouse Chromosome 10 (x-axis in Mb) (GC content and mappability scores are as in[8]).
(B-C) Contact maps of mouse Chromosome 10 containing raw read counts (interactions with at least 3 reads) and binomial significances respectively resulting from classic Hi-C experiment (left panel) and random ligation experiment (right panel) in fetal liver cells. The intensity of the signal is summarized by the gradient above each contact map. Significant interactions are colored with a red gradient in **C**. Arrows pinpoint a region of high coverage and its impact on the observed number of interactions (**B**, right panel). The coverage is represented at the left side of each contact map.
(D) The top panel represents the distribution of observed/expected log ratio of significant (red) and non-significant (blue) interactions in the fetal liver cell sample. Middle and bottom panels represent the distribution of binomial p-values in the fetal liver cell and random samples respectively.
(E) Influence of the relative coverage on the distribution of interaction significance. GOTHiC interaction ranking in the Hi-C (upper panel) and random ligation (lower panel) samples. The ranked lists were divided into quartiles, the first quartiles correspond to the top ranked interactions. Significant interactions are shown in red.

Figure 2. GOTHiC applied to human lymphoblastoid Hi-C experiments.
(A-B) Contact maps of human Chromosome 3 containing raw read counts (interactions with at least 3 reads) and binomial significances respectively resulting from HindIII Hi-C experiment (left panel) and NcoI Hi-C experiment (right panel). The intensity of the signal is summarized by the gradient above each contact map. Significant interactions are colored with a red gradient in **B**. The coverage is represented at the left side of each contact map. **(C)** Venn diagram representing the overlap between significant interactions detected in HindIII (orange percentage) and NcoI (blue percentage) samples.
(D) Correlation between the HindIII (x-axis)/NcoI (y-axis) common significant interactions (80,448 interactions) according to their rank. Spearman's correlations are indicated above the plot.

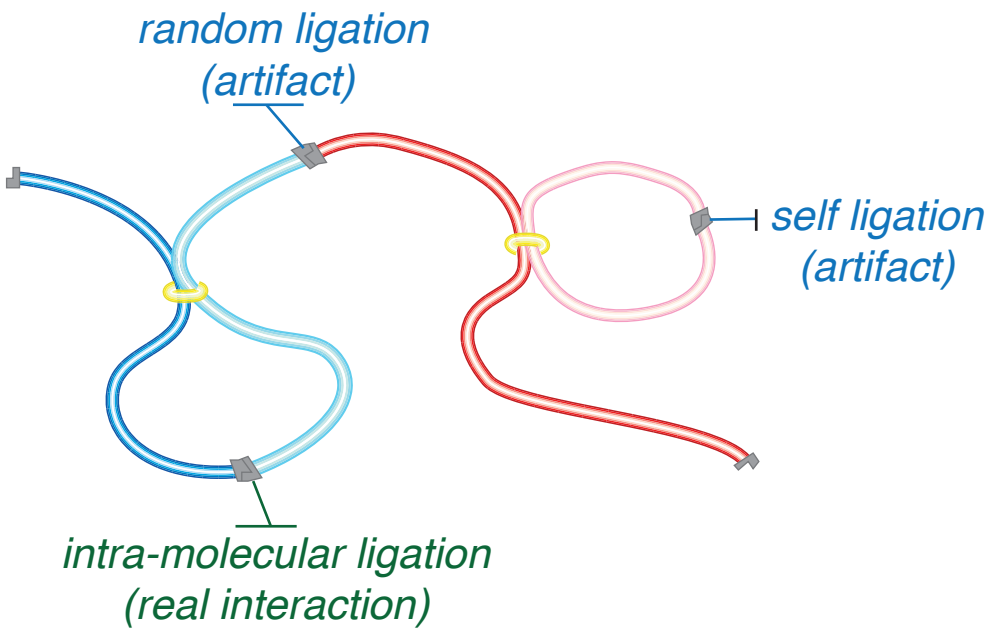
Figure 3. Comparison of mouse the fetal liver Hi-C data after processing by hiclib, hicpipe and GOTHiC.
(A-B) Contact maps of mouse Chromosome 10 containing relative probability computed by hiclib and observed/expected log ratio obtained with hicpipe respectively resulting from classic Hi-C experiment (left panel) and random ligation experiment (right panel) in fetal liver. The intensity of the signal is summarized by the gradient above each contact map.
(C) Influence of the relative coverage on the distribution of number of observed interactions (top panel), hiclib and hicpipe interaction ranking (middle and bottom panels), in the HiC (left) and random ligation (right) samples. The ranked lists were divided into quartiles, the first quartiles correspond to the top ranked interactions. The

distribution of the number of reads per interaction is represented in the top panel with green box plots (corresponding y-axis is placed on the right of the plot).
(D) Correspondence between binomial significant interactions (88292) and hiclib and hicpipe ranking. Blue bar corresponds to non-significant interactions from GOTHiC, red bar to significant ones. The green gradients represent the ranking of the interaction resulting from hiclib (left) and hicpipe (right) processing. Red bars indicate the significant interactions detected with GOTHiC.

Supplementary Figure S1. Comparison of human lymphoblastoid Hi-C data after processing by hiclib, hicpipe and the GOTHiC.
(A-B) Contact maps of human Chromosome 3 containing relative probability computed by hiclib and observed/expected log ratio obtained with hicpipe respectively resulting from HindIII experiment (left panel) and NcoI experiment (right panel). The intensity of the signal is summarized by the gradient above each contact map.
(C) Influence of the relative coverage on the distribution of (a) number of observed interactions, (b) GOTHiC, (c) hiclib and (d) hicpipe interaction ranking in the HindIII (left) and NcoI (right) samples. The ranked lists were divided into quartiles, the first quartiles correspond to the top ranked interactions. The distribution of the number of reads per interaction is represented in the top panel with green box plots (corresponding y-axis is placed on the right of the plot). 92,897 and 103,114 interactions were called significant using GOTHiC in the HindIII and NcoI samples respectively. In order to compare with the predictions of **(D)** hiclib and **(E)** hicpipe, we selected the 92,897 and 103,114 top ranked interactions of these methods and first computed the overlap (top) and correlation (bottom) between the two samples.
(F) Correspondence between binomial significant interactions and hiclib and hicpipe ranking. The green-to-blue gradients represent the ranking of the interaction resulting from hiclib (left) and hicpipe (right) processing. Red bars indicate the significant interactions detected by GOTHiC in both HindIII and NcoI experiments. Orange bars indicate the significant interactions detected only in the HindIII experiment and blue bars indicate the significant interactions detected only in the NcoI experiment.

Supplementary Figure S2: Overlap of top-ranked interactions from hiclib and hicpipe with significant interactions from GOTHiC.
GOTHiC identified 88,292 significant interactions in the mouse fetal liver cell Hi-C dataset. **(A)** Venn diagram showing the overlap between the significant interactions identified by GOTHiC and the top 88,292 interactions from the hiclib and hicpipe outputs. **(B)** There were 80,448 significant interactions detected by GOTHiC that overlapped between the HindIII and NcoI experiments in the human lymphoblastoid cell line. The Venn diagram shows the overlap of between the GOTHiC, hiclib and hicpipe outputs.

A



B

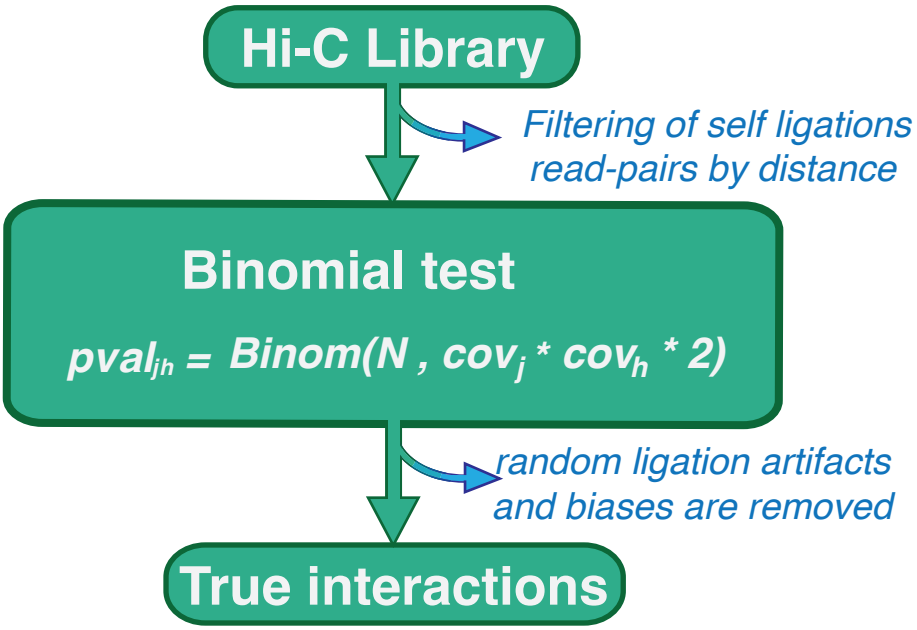
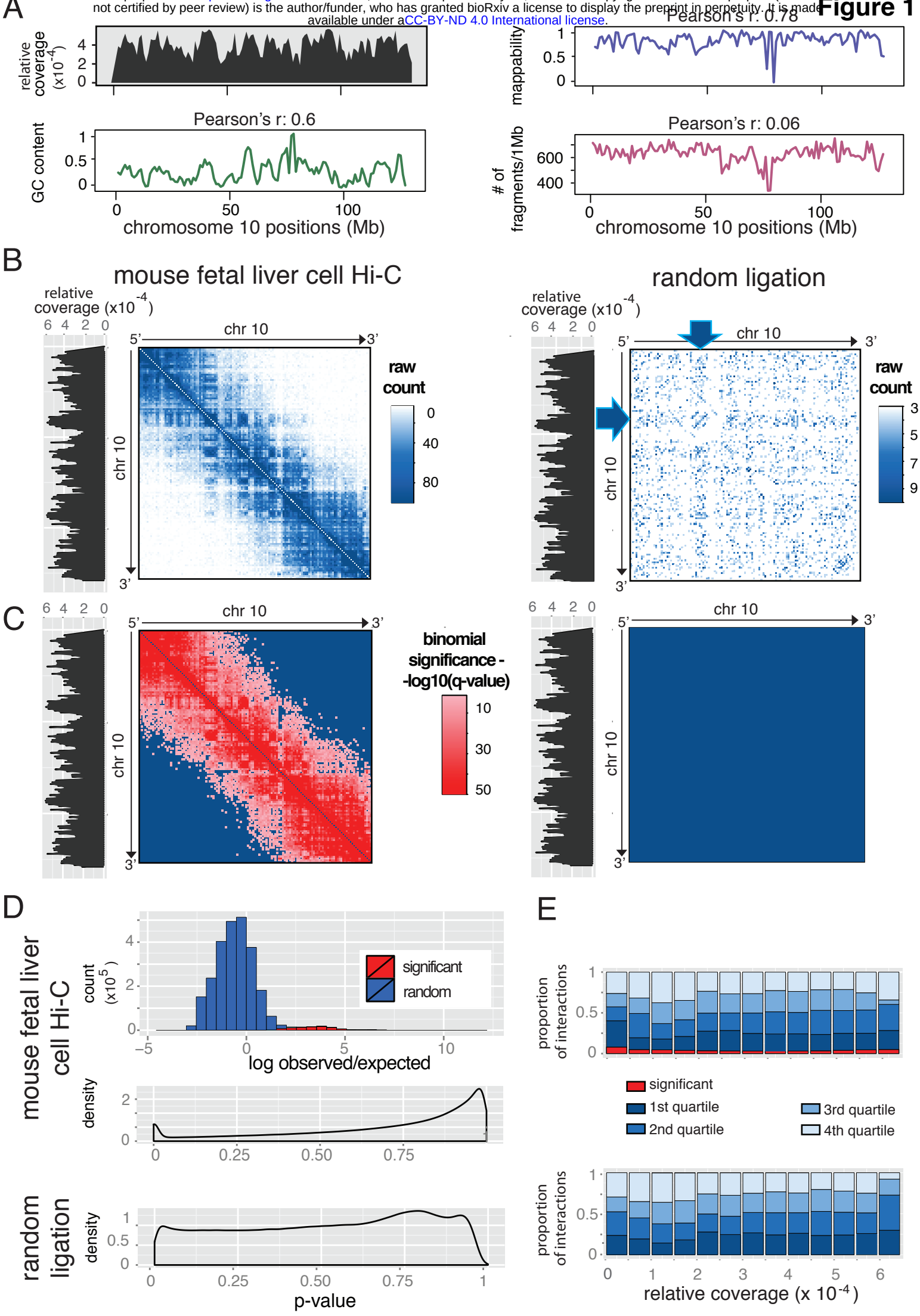


Figure 1



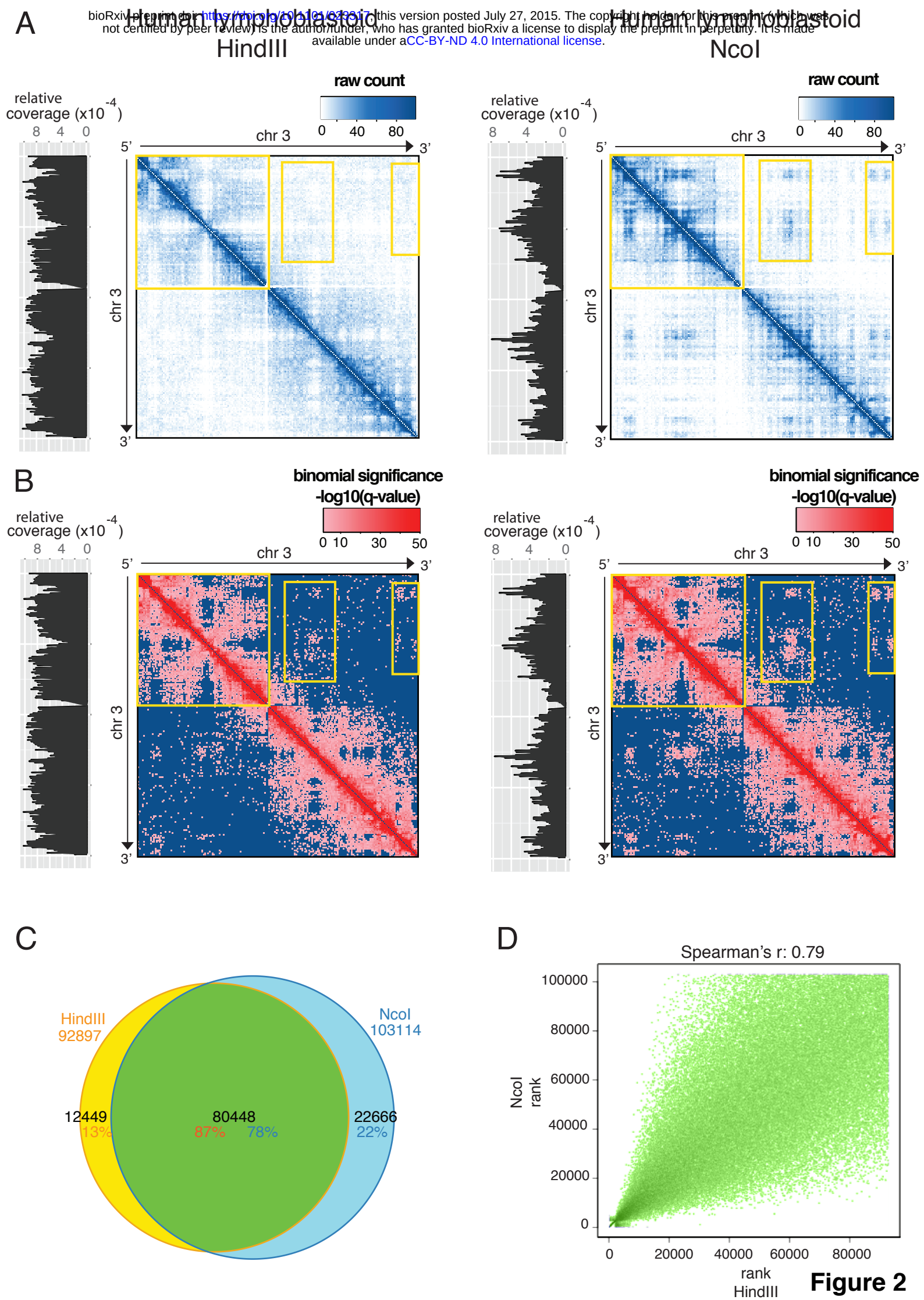
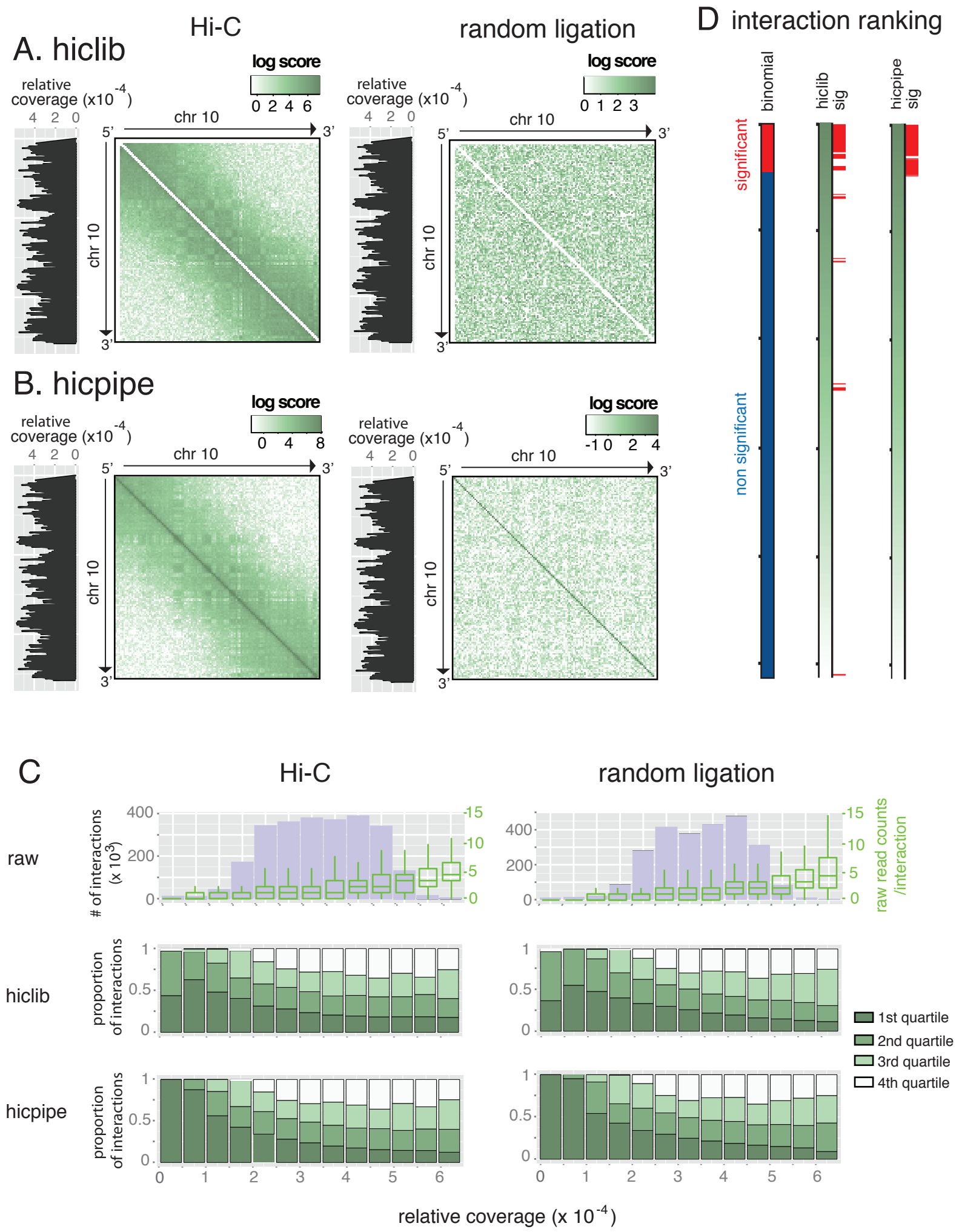


Figure 2

Figure 3

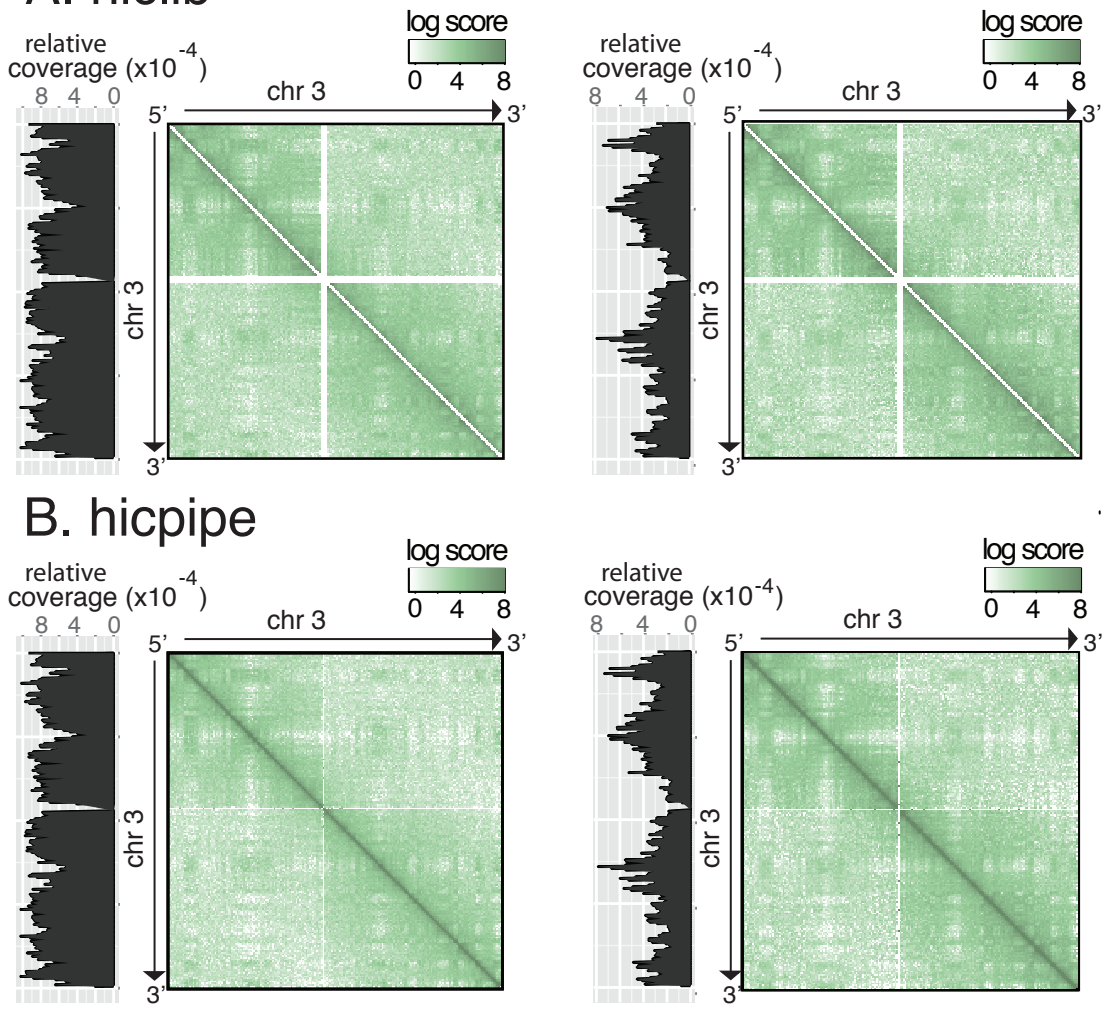


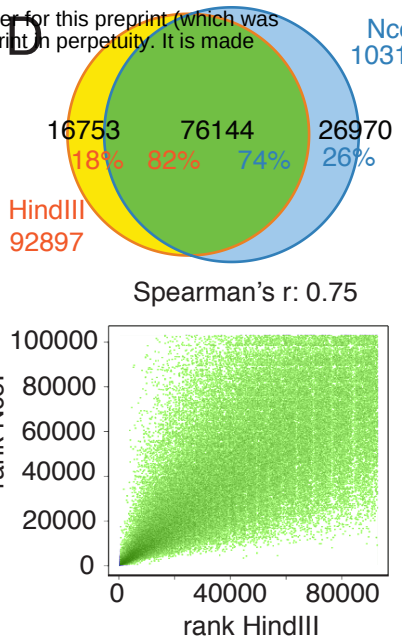
Human lymphoblastoid

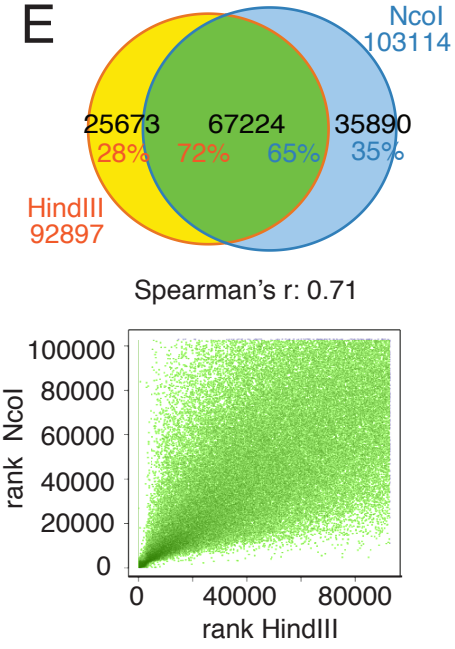
HindIII NcoI

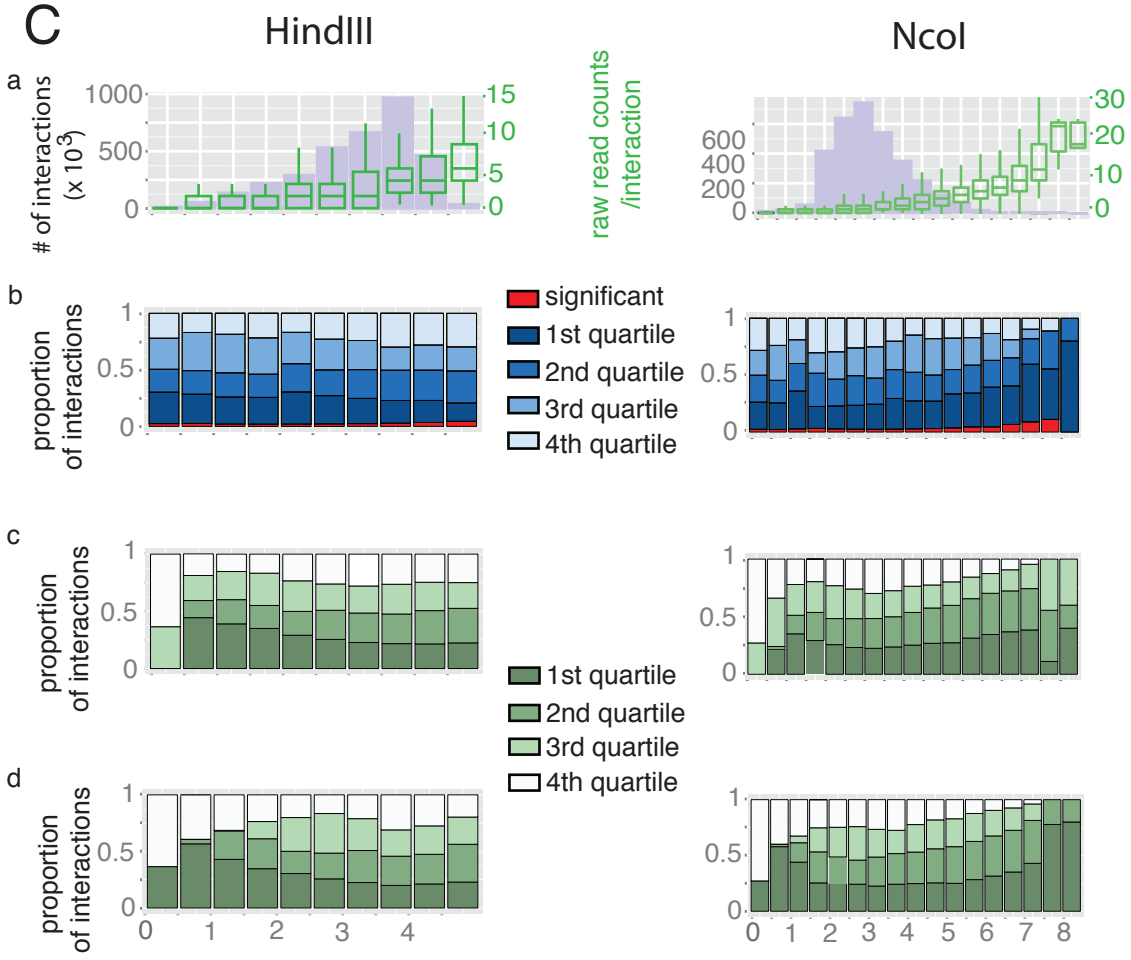
A. hiclib

B. hicpipe









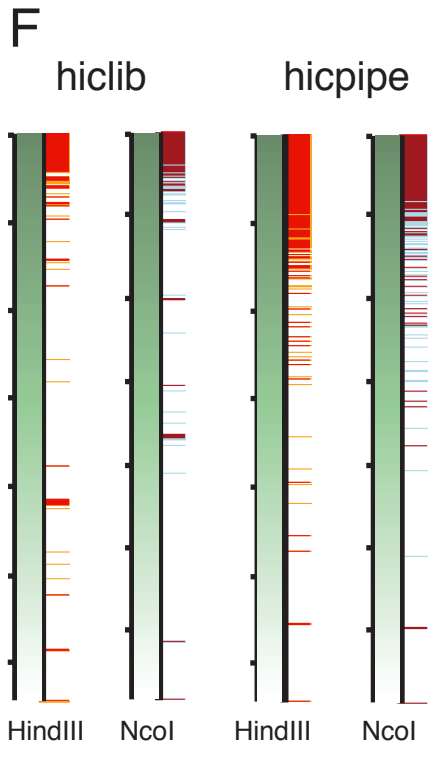


Figure S1

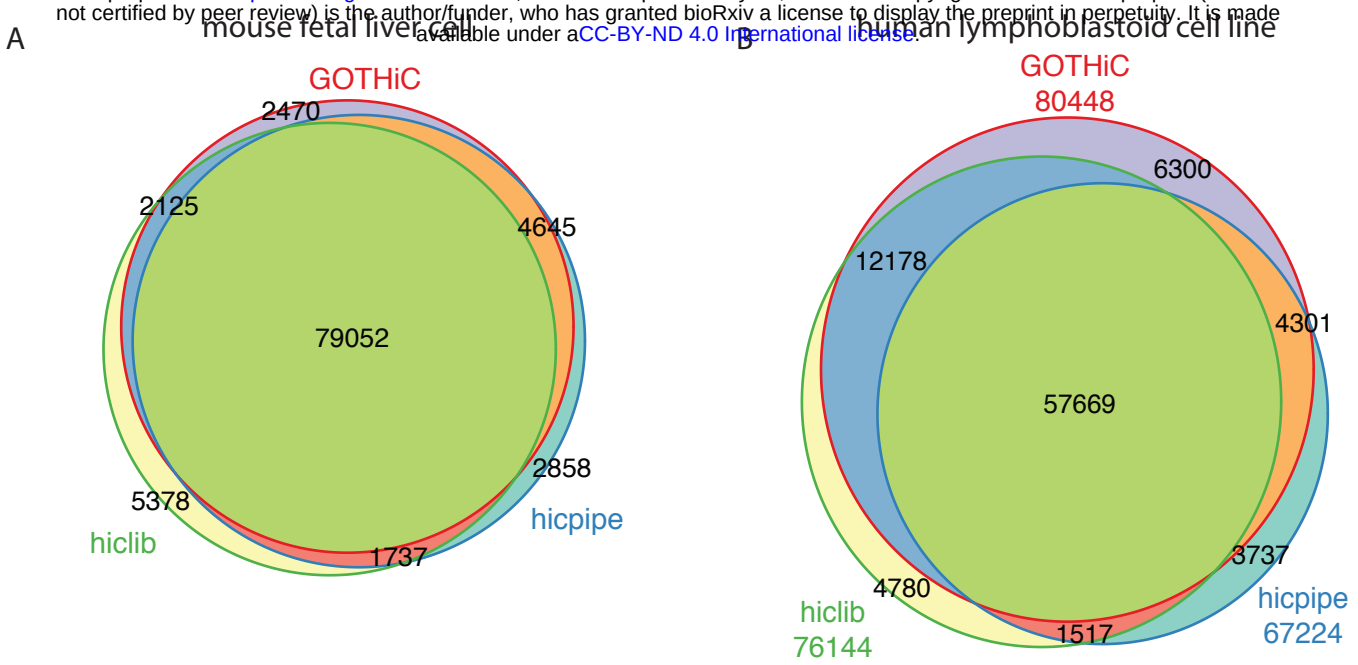


Figure S2