# Iterative correction of Hi-C data reveals hallmarks of chromosome organization

Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, Leonid A. Mirny

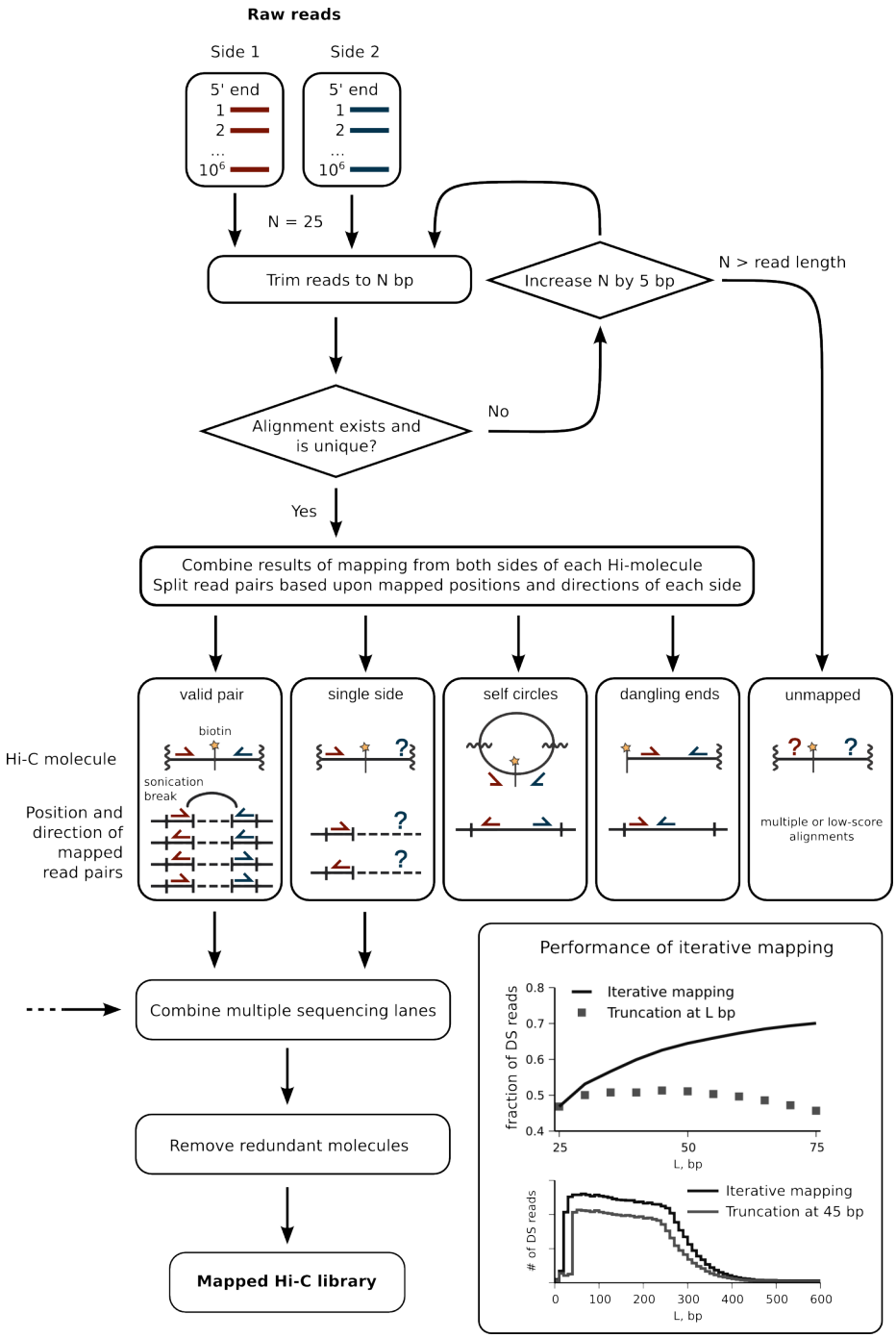
## Supplemental Text and Figures

List of supplemental figures:





Nature Methods: doi:10.1038/nmeth.2148

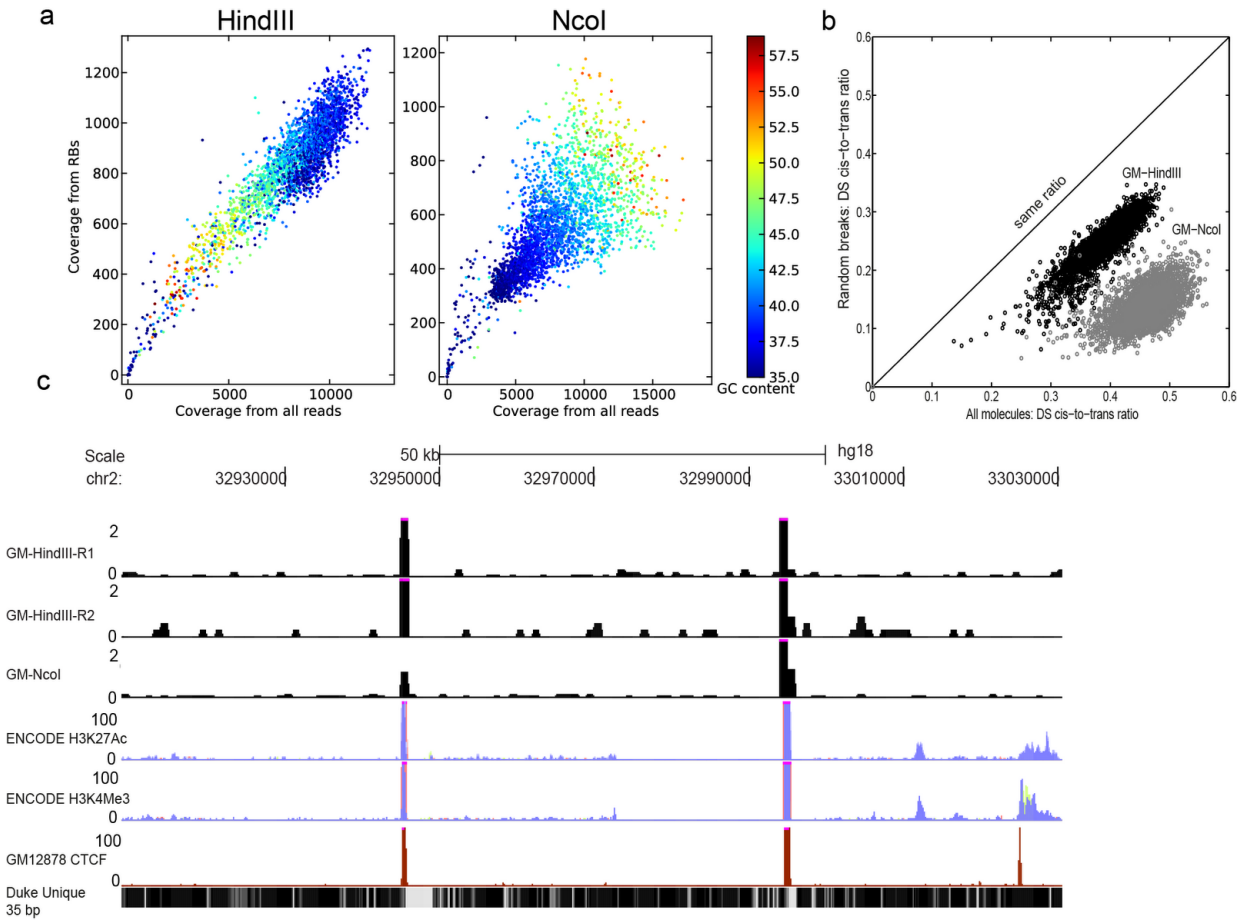



**Supplementary Figure 1. Mapping and filtering pipeline.**

Iterative alignment method processes read-pairs and then allows categorization of Hi-C molecules. As shown in flowchart, read pairs are separately mapped to the genome at increasing truncation lengths and collected only if uniquely aligned. Read pairs are then reassembled and classified based on the relative positions and directions of both sides. In the final step, the data from several sequencing lanes are combined into one data set, and redundant molecules are filtered out. The top plot in the insert shows the fraction of mapped DS reads as a function of the number of steps of iterative mapping, compared with the results of mapping using fixed a truncation length. Iterative mapping gains additional reads at each iteration. Bottom plot shows the distribution of distances between read ends and the nearest upstream restriction cut site. Note that iterative mapping increases the number of reads originated at enzymatic cut sites (0 to 400bp), as well as mapped random breaks (>400bp).



Nature Methods: doi:10.1038/nmeth.2148




**Supplementary Figure 2. Analysis of random breaks.**

Sites that appear as non-restriction site breaks likely arise from a combination of off target enzyme cutting, and open chromatin fragility. Moreover, true enzymatic breaks may appear as non-restriction site breaks due to differences between a studied genome and a reference genome, as well as assembly errors of a reference genome.

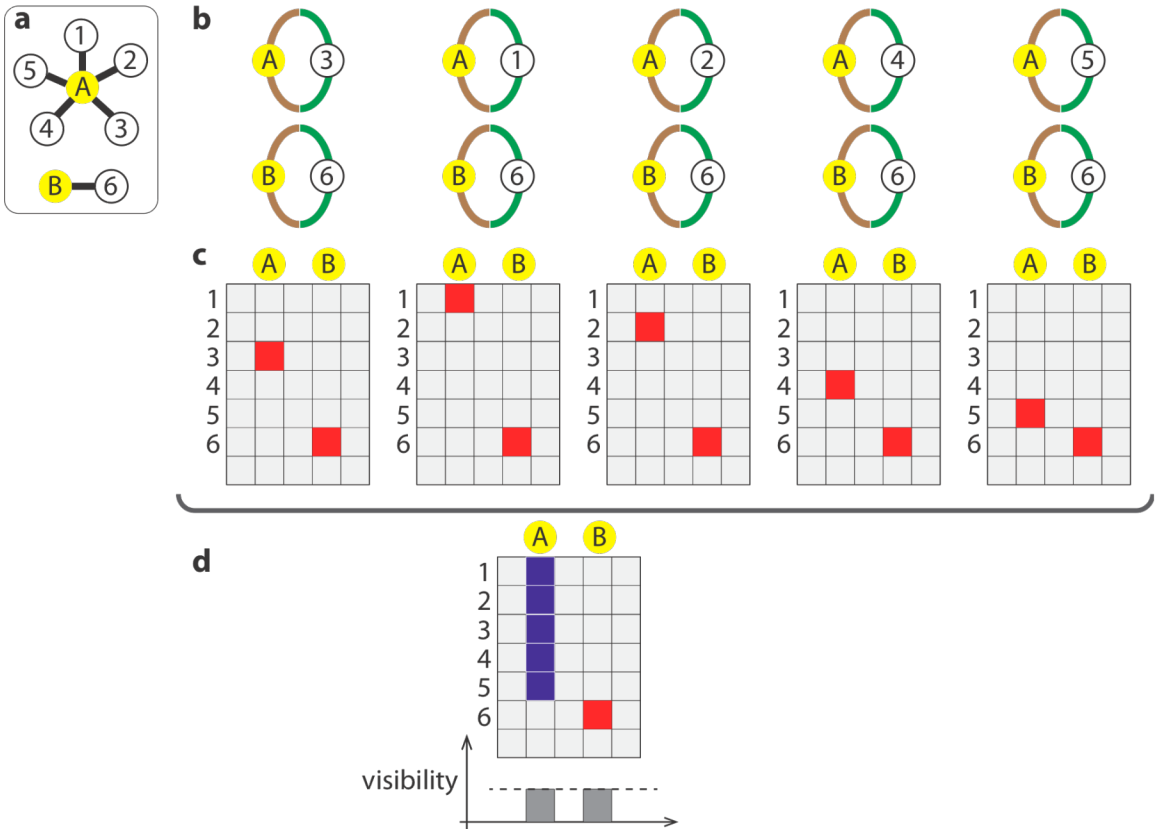
**a.** Correlation of coverage profiles at megabase level between the set of all reads and the subset reads coming from random breaks; colored by GC content. Coverage profiles are highly correlated ($r^{HindIII}$ = .87, p < 1e-10, $r^{NcoI}$ = .76, p < 1e-10); HindIII and NcoI show opposite trends with GC content. These restriction enzyme specific effects can be explained by off-target enzymatic cuts: NcoI has a GC-rich recognition motif and has more off-target cuts in GC rich regions, while HindIII has an AT-rich motif and has more off-target cuts in GC-poor regions. Note that breaks from a biological replicates digested with same enzyme (GM-HindIII-R1 vs GM-HindIII-R2) are more similar than breaks from the same sample digested by different enzymes (GM-NcoI vs GM-HindIII-R1).

**b.** Cis-to-trans ratios of DS reads for random breaks vs. all molecules for HindIII and NcoI. Cis-to-trans ratios are highly correlated between random breaks and all molecules ($r^{HindIII}$ = .91, p < 1e-10, $r^{NcoI}$ = .61, p < 1e-10), but is lower for random breaks, suggesting a correspondingly lower signal-to-noise ratio for random breaks. Similarly, cis-to-trans ratios are consistently lower for NcoI. This may arise due to a strongly decreased ligation efficiency of random breaks vs. restriction cut sites in NcoI vs. HindIII.

**c.** Breaks binned at 1 kb over a 150 kb region of chr2 show that spikes in break frequency that are common between all samples correspond to spikes in other sequencing tracks. These thus likely represent genome assembly errors (unrecognized repeat regions). These spikes lie in close proximity to regions of non-unique sequence (grey patches in “Duke Unique 35 bp” track of UCSC genome browser).



Nature Methods: doi:10.1038/nmeth.2148

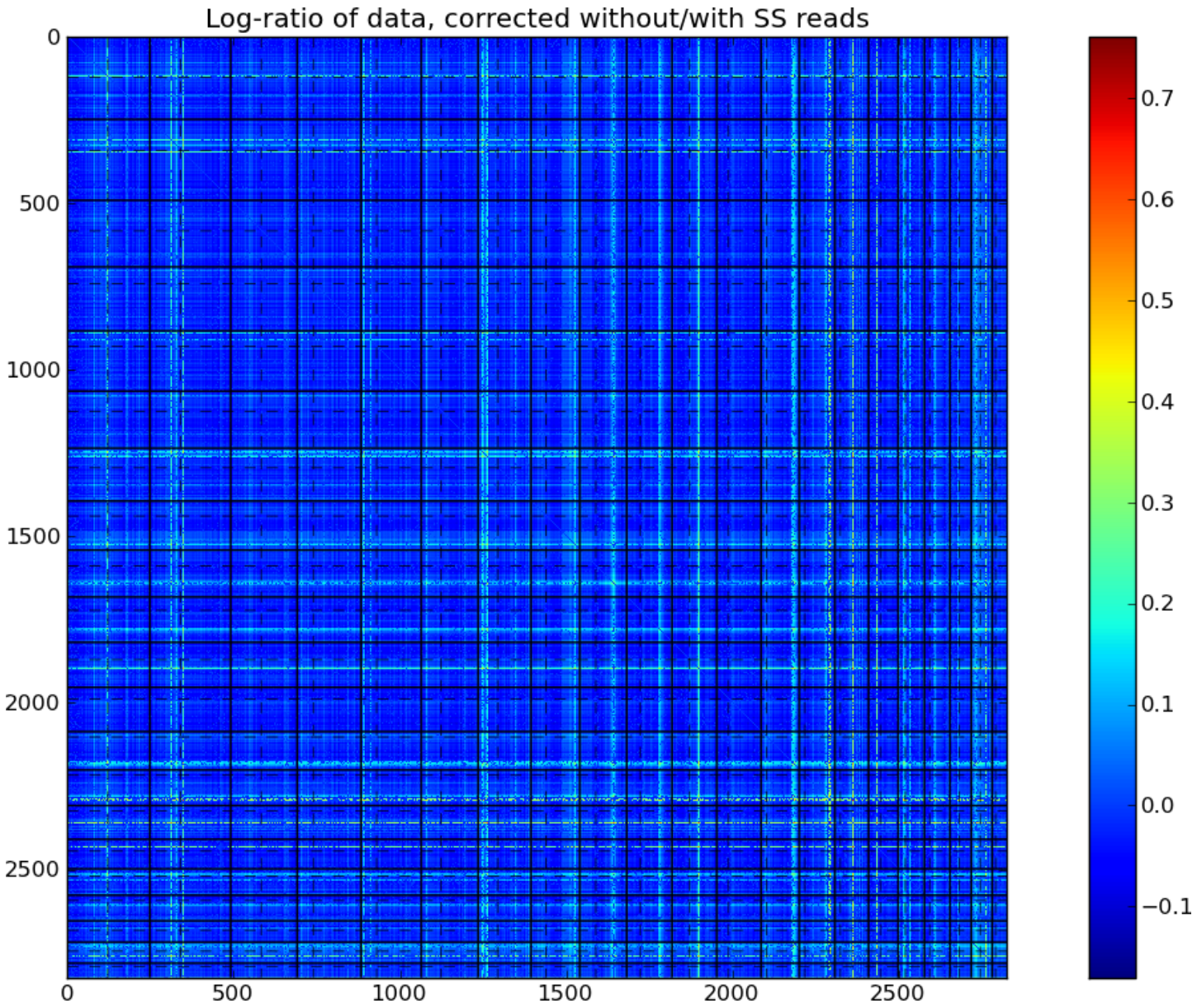



**Supplementary Figure 3. Rationale for equal visibility in Hi-C data**

Some properties of Hi-C data can be better understood by considering a toy example of a small subset of chromosomal interactions (**a**) and a hypothetical ideal Hi-C (**b-d**), which has perfect crosslinking and ligation efficiency, and has no biases or loss of DNA material. Consider two loci A and B (**a**): locus A interacts with five other loci (1,2,..5), and locus B interacts with only one other locus (6). Although crosslinked to all five of its partners, each molecule of locus A will end up being ligated with only one of them (**b**). If we consider hypothetical single cell Hi-C maps (**c**), locus A can have a contact with any one of five of its partners. In contrast, locus B will always appear interacting with locus 6. On the final Hi-C map (**d**) obtained for a population of cells, the number of reads for each contacts of locus A (A-1,A-2,..A-5) will be five times less than the number of reads for the B-6 contact. However, summed over all columns, the visibility of A will be equal to the visibility of B.



Nature Methods: doi:10.1038/nmeth.2148




**Supplementary figure 4. Extension of Iterative Correction to include SS reads.**

Natural log of the ratio of inter-chromosomal human Hi-C (2009) data, iteratively corrected without SS reads to the data, iteratively corrected with SS reads (1MB resolution). Solid lines correspond to chromosome boundaries, and dashed lines to centromeres. Pipeline for iterative correction with SS reads is presented below.

**Iterative correction in presence of SS reads**

Single-sided reads contain important information that could be useful for the analysis of centromeric regions, as can be seen from the Fig 1d. Here we present a version of our pipeline, modified to incorporate SS reads in iterative correction.

Despite strong compensation of cis-reads by SS reads near centromeres, coverage profiles of SS and DS reads overall are highly correlated (Spearman r = .87 for HindIII, r = .92 for NcoI, p <1e-10), suggesting that SS reads are affected by the same factorizable biases as DS reads.



Nature Methods: doi:10.1038/nmeth.2148

The expected frequency of the combination of DS and SS reads for each region can then be written mathematically as:

$$E^{DS}{}_{ij} = B_i B_j T_{ij}^{DS}$$
$$E^{SS}{}_{i} = B_i B_0 T_i^{SS}$$

where the constraint on $T_{ij}$ is expressed as $\sum_j T_{ij}^{DS} + T_i^{SS} = 1$; $B_0$ reflects biases of unmappable regions of the genome. Since the properties of the unmapped sides are unknown, we replace $B_0$ with the average genome-wide bias.

Given the observed SS and DS reads, we obtain an estimate of a true contact probability by solving the system of equations:

$$O^{DS}{}_{ij} = B_i B_j T_{ij}^{DS}$$
$$O^{SS}{}_{i} = B_i B_0 T_i^{SS}$$
$$\sum_{j=(1\ldots N)} T_{ij}^{DS} + T_i^{SS} = 1$$

where $O^{DS}$ is the heatmap of DS reads, $O^{SS}$ is the vector of SS reads, $T^{DS}$ and $T^{SS}$ are the desired corrected heatmap and vector, $B_i$ is the vector of biases to be estimated, $B_0$ is the mean of $B_i$, and N is the number of bins.

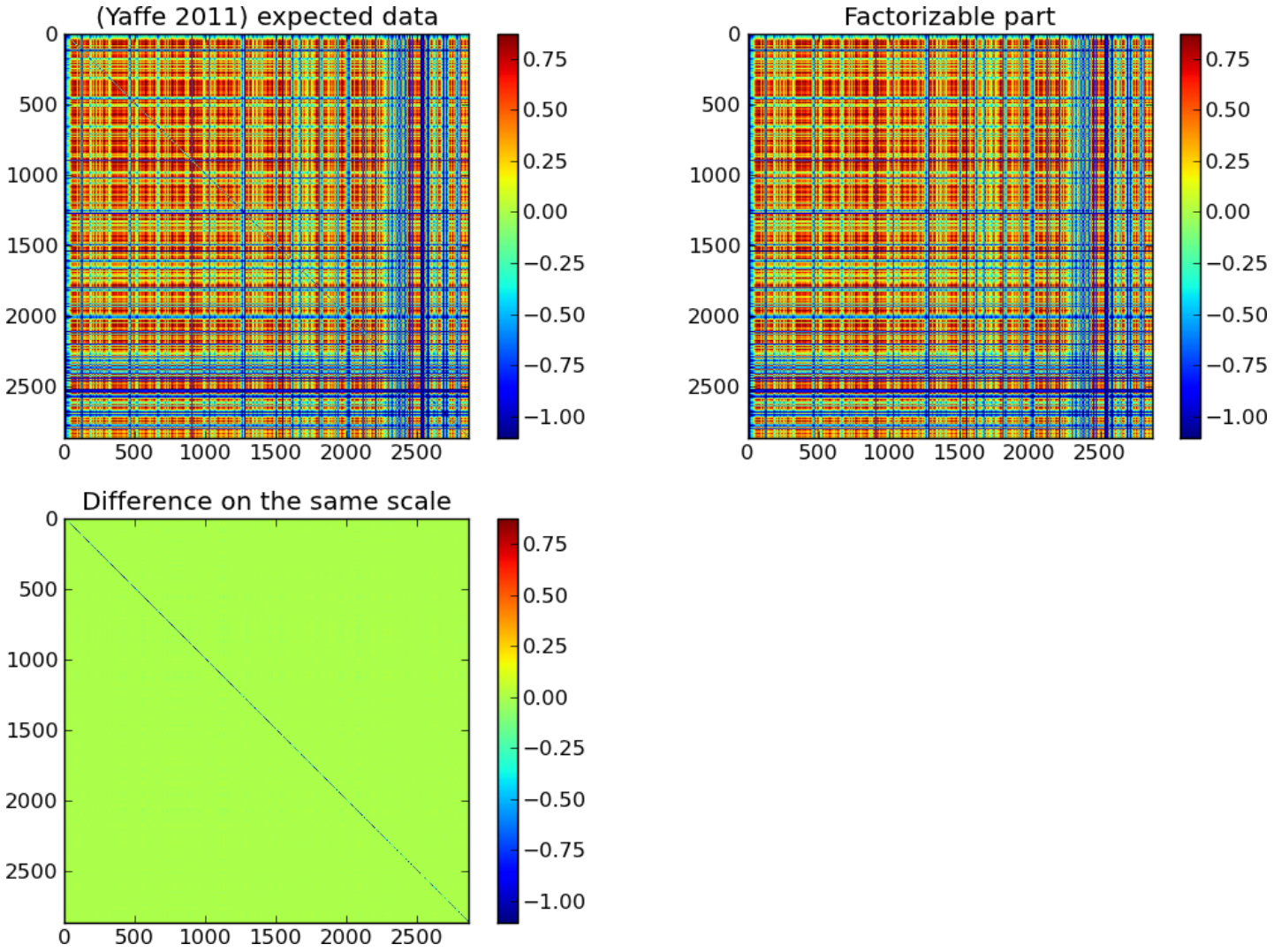
To begin the iterative process, we create a working copy of the matrix $O^{DS}{}_{ij}$, denoted $W_{ij}$. The iterative process gradually changes this matrix change to $T^{DS}{}_{ij}$. Bins with zero coverage do not necessarily need to be removed for the algorithm presented below.

0. Set each element of the vector of total biases $B_i$ to 1
1. Calculate sum of W over all rows, $S^{DS}{}_i = \sum_j W_{ij}$.
2. Calculate a vector of corrected SS reads, $S^{SS}{}_i = \sum_j O_i^{SS} / (B_i B_0)$, where the mean is calculated over regions with non-zero sum of counts.
3. Calculate additional vector of biases $\Delta B_i = S^{DS}{}_i + S^{SS}{}_i$
4. Renormalize $\Delta B_i$ by its mean value over non-zero bins to avoid numerical instabilities.
5. Set zero values of $\Delta B_i$ to one to avoid 0/0 error
6. Divide $W_{ij}$ by $\Delta B_i \Delta B_j$ for all (i,j)
7. Multiply total vector of biases by additional biases: $B_i \rightarrow B_i \cdot \Delta B_i$
8. Repeat steps 1-7 until variance of $\Delta B_i$ becomes negligible. For studied datasets, 50 iterations are enough for this.
9. After these iterative corrections, $W$ describes the true relative contact probabilities, $T_{ij}^{DS}$. Corrected vector of SS reads $T_i^{SS}$ can be calculated as in step 2 using the final vector of biases $B_i$.
10. As a final check, we compare the ratio $\langle T_i^{SS} \rangle / \langle T_{ij}^{DS} \rangle$ with $\langle O_i^{SS} \rangle / \langle O_{ij}^{DS} \rangle$; if these are different by more than an order of magnitude check for errors.

We also note that if SS reads were not used for the correction, they still can be corrected using biases inferred from DS reads only.



Nature Methods: doi:10.1038/nmeth.2148




**Supplementary figure 5. Factorizability of matrices of biases computed via fragment-level approach.**

(*Top left*): the heatmap of the natural log of genome-wide biases ($B_{ij}$) from (Yaffe et al., 2011[2]) for Hi-C HindIII dataset at 1MB resolution. To better visualize the difference between two matrices, expected maps were truncated at 5 and 95 percentiles, and then mean-centered.

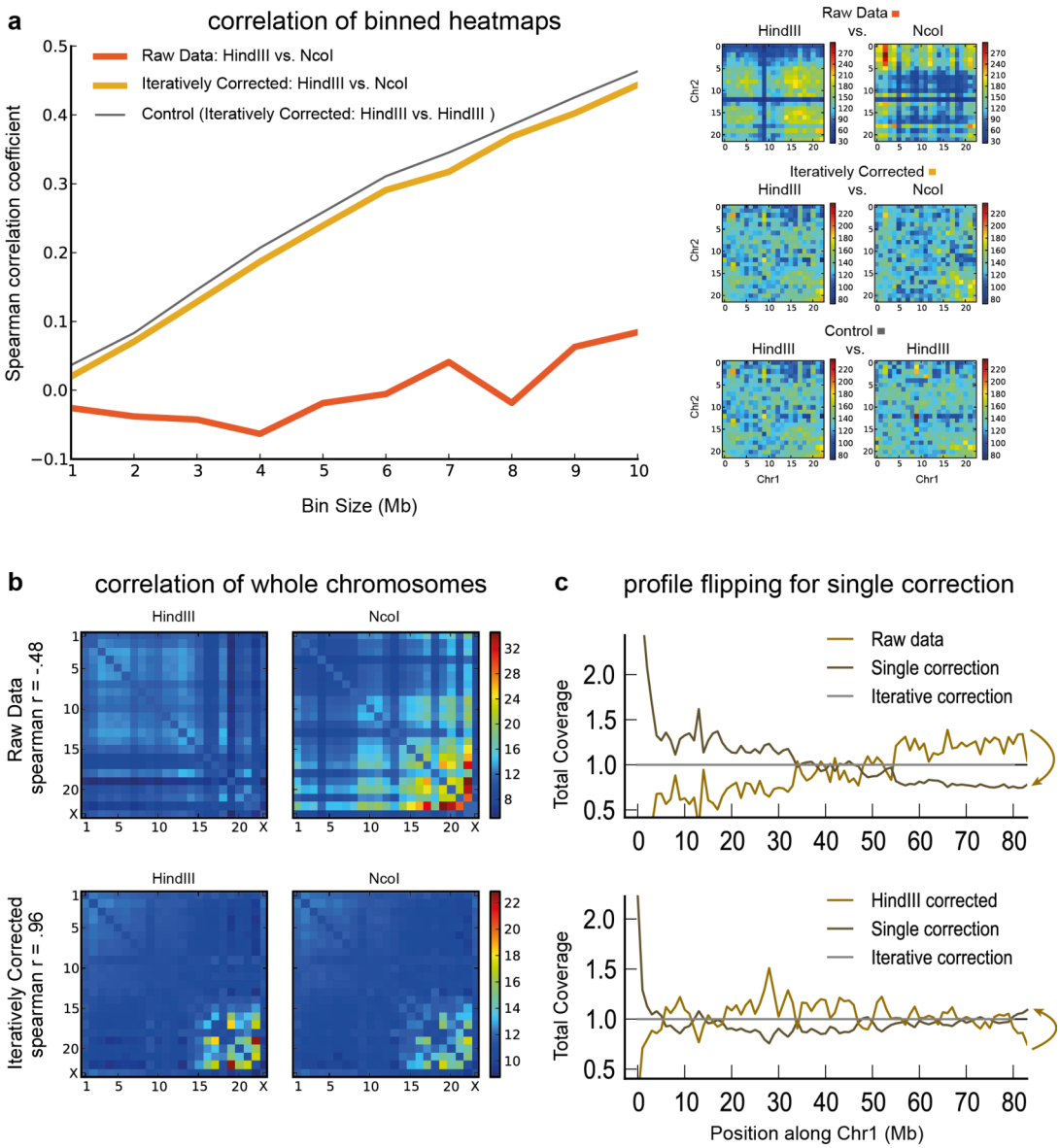
The map of biases was then factorized using iterative correction; this approximates $B_{ij}$ as a product of two bias vectors $B_iB_j$.

(*Top right*): the factorizable part of the genome-wide map of biases, which is the product of two bias vectors ($B_iB_j$)

(*Bottom right*) Algebraic difference between two maps, plotted on the same scale. The diagonal is highlighted due to the absence of expected maps for the diagonal bins. Two above matrices correlate with Spearman r=0.99995, so the difference between them is negligible.



Nature Methods: doi:10.1038/nmeth.2148







**Supplementary figure 6. Iterative Correction improves between-enzyme correlation for inter-chromosomal data**

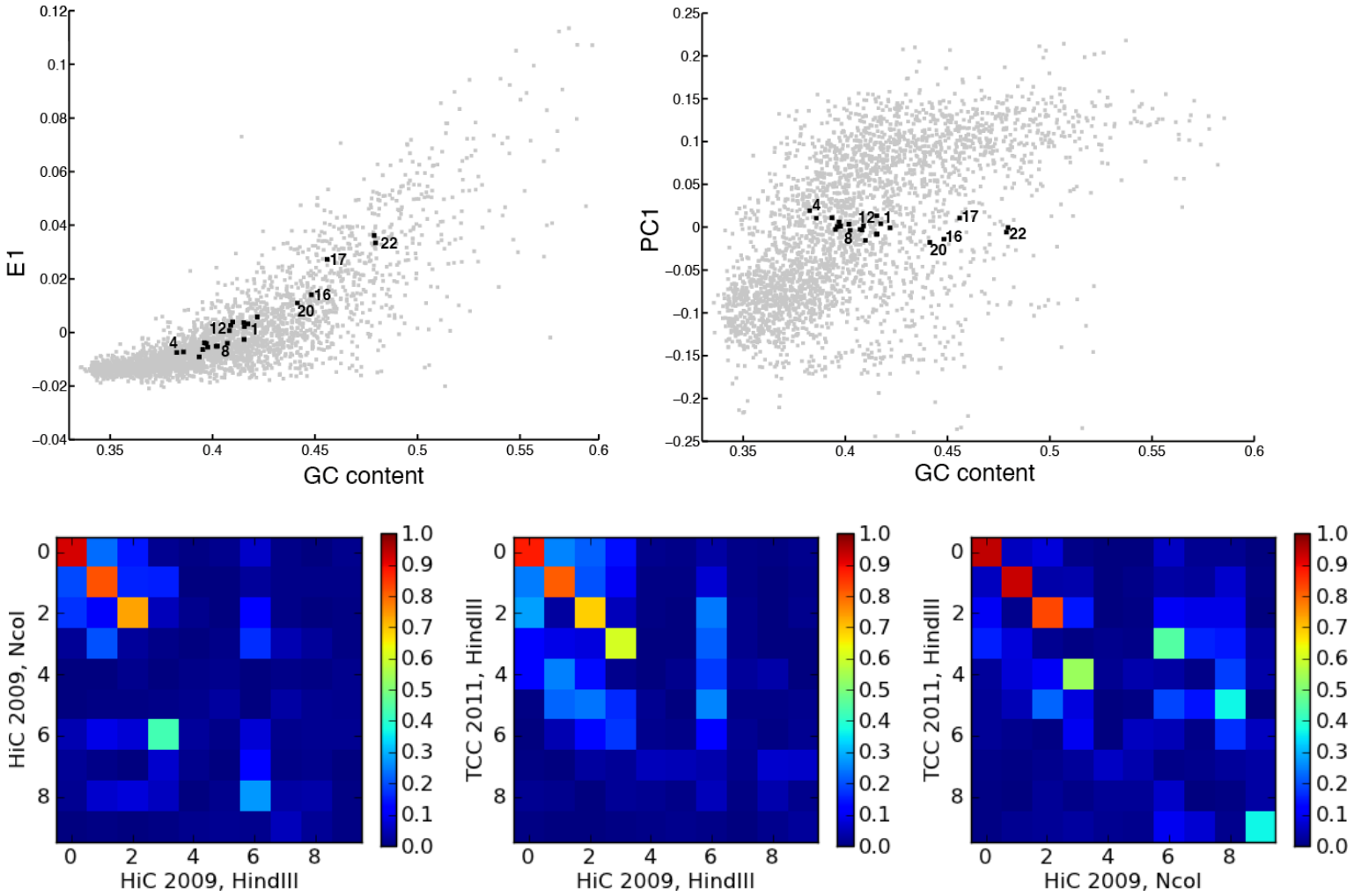
**a.** (left) Correlation of contact frequencies between Hi-C datasets as a function of bin size. The correlation of inter-chromosomal data for HindIII vs. NcoI greatly increases after iterative correction. To take into account sampling noise due to coverage, we perform a split-sample test. We split HindIII dataset in half, and downsample the NcoI dataset to the coverage of the split HindIII dataset. We then correlate the split HindIII with equal-coverage (downsampled) NcoI before (orange line) and after iterative correction (yellow line). As a control, we compare this with the correlation between the two split halves of the HindIII dataset (grey line). We note that the correlation between HindIII and NcoI gets very close to the maximum possible correlation due to sampling noise, as determined by the split sample control. A similar result was obtained with split NcoI compared with downsampled HindIII. (right) Representative maps before and after iterative correction (for Chr1 vs. Chr2).

**b.** Whole-chromosome average contact frequencies obtained for HindIII and NcoI datasets . Maps for two enzymes are much more correlated after iterative correction (r=0.96) than before (r=0.48). Chromosomal averages and iterative correction were performed on megabase resolution heatmaps.

**c.** Top panel shows coverage profile after single correction vs. iterative correction for raw chromosome 1 inter-chromosomal data (Human Hi-C, binned to 1MB); bottom panel shows the same after pre-correction by HindIII restriction fragment density. Both panels show coverage along fragment of chromosome 1 (0-80 megabases). Single correction leads to flipping of the coverage profile, and depends on the particular choice of initial corrections. Iterative correction leads to a flat coverage profile irrespective of initial correction, yielding exactly the same corrected matrix in both cases.



Nature Methods: doi:10.1038/nmeth.2148




**Supplementary Figure 7. Consistency of leading eigenvectors.**

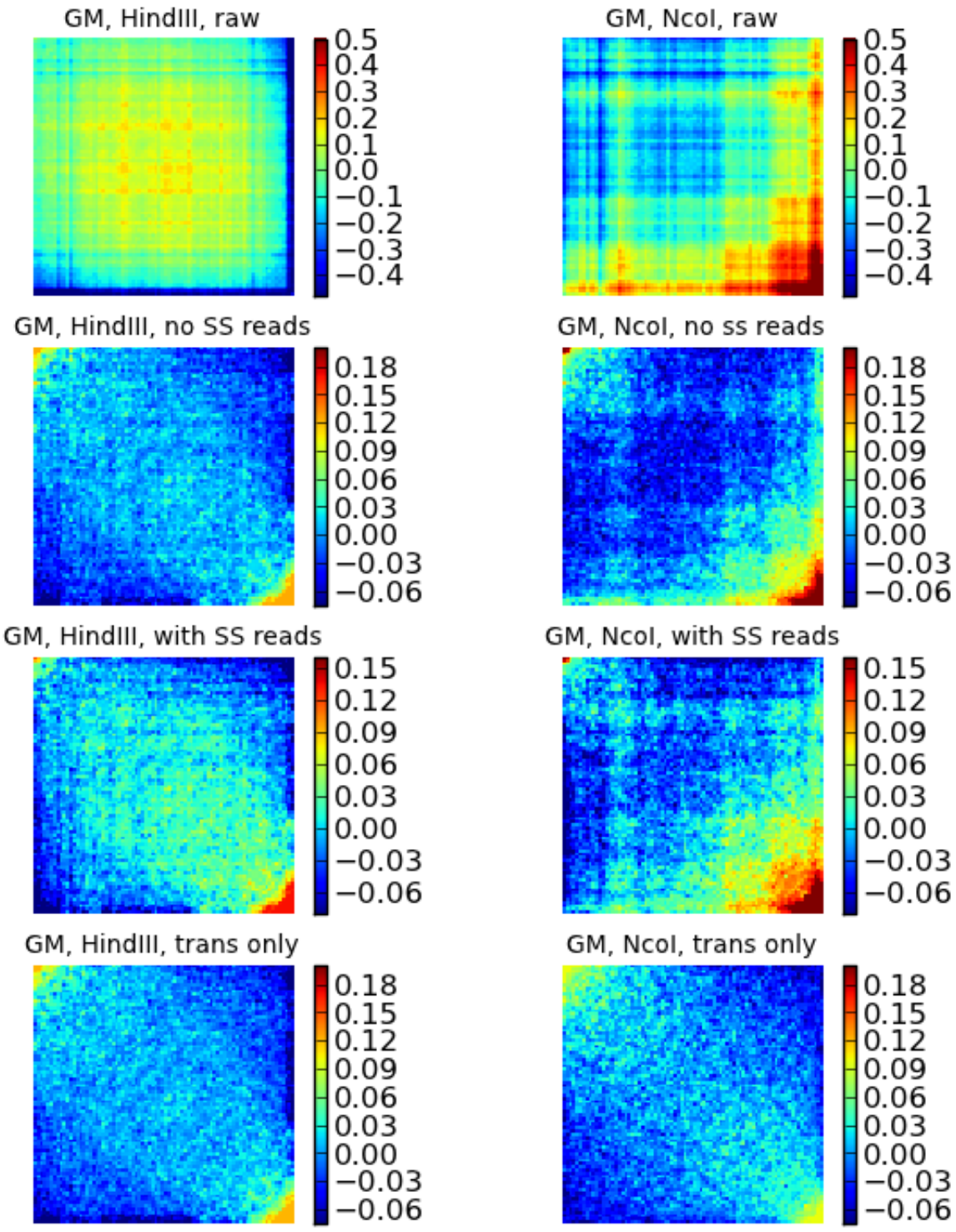
(*top left*) Scatter plot showing correlation between GC content (at 1 Mb resolution) and $E^1$, as in Fig 3a. Black dots represent mean chromosomal value of $E^1$ and mean chromosomal value of GC content, and show strong correlation (r = 0.95, p = 4e-06). Some chromosomes are indicated by numbers.

(*top right)* Similar scatter plot showing correlation between GC content and previously published principal components which were used to describe the plaid structure of intra-chromosomal interactions (i.e. first component for all chromosomes except 4 and 5, where second component was found to better-characterize the plaid structure[1]). Since these principal components were obtained from intra-chromosomal data, and for each chromosome separately, this does not facilitate genome-wide comparisons. A strong correlation of $E^1$ with functional information and GC content was found for each chromosome. However, as illustrated by GC content, this does not hold for chromosomal average values (r = -0.31, p = 0.15).

(*bottom*) Absolute value of Pearson correlation coefficient between different eigenvectors obtained from different datasets. 0 corresponds to the first eigenvector. Note that Pearson correlation between different eigenvectors from the same dataset is strictly zero. First three eigenvectors show significantly increased correlation between all datasets, confirming their biological significance. Fourth eigenvector also correlates strongly between the two highest-quality datasets.



Nature Methods: doi:10.1038/nmeth.2148

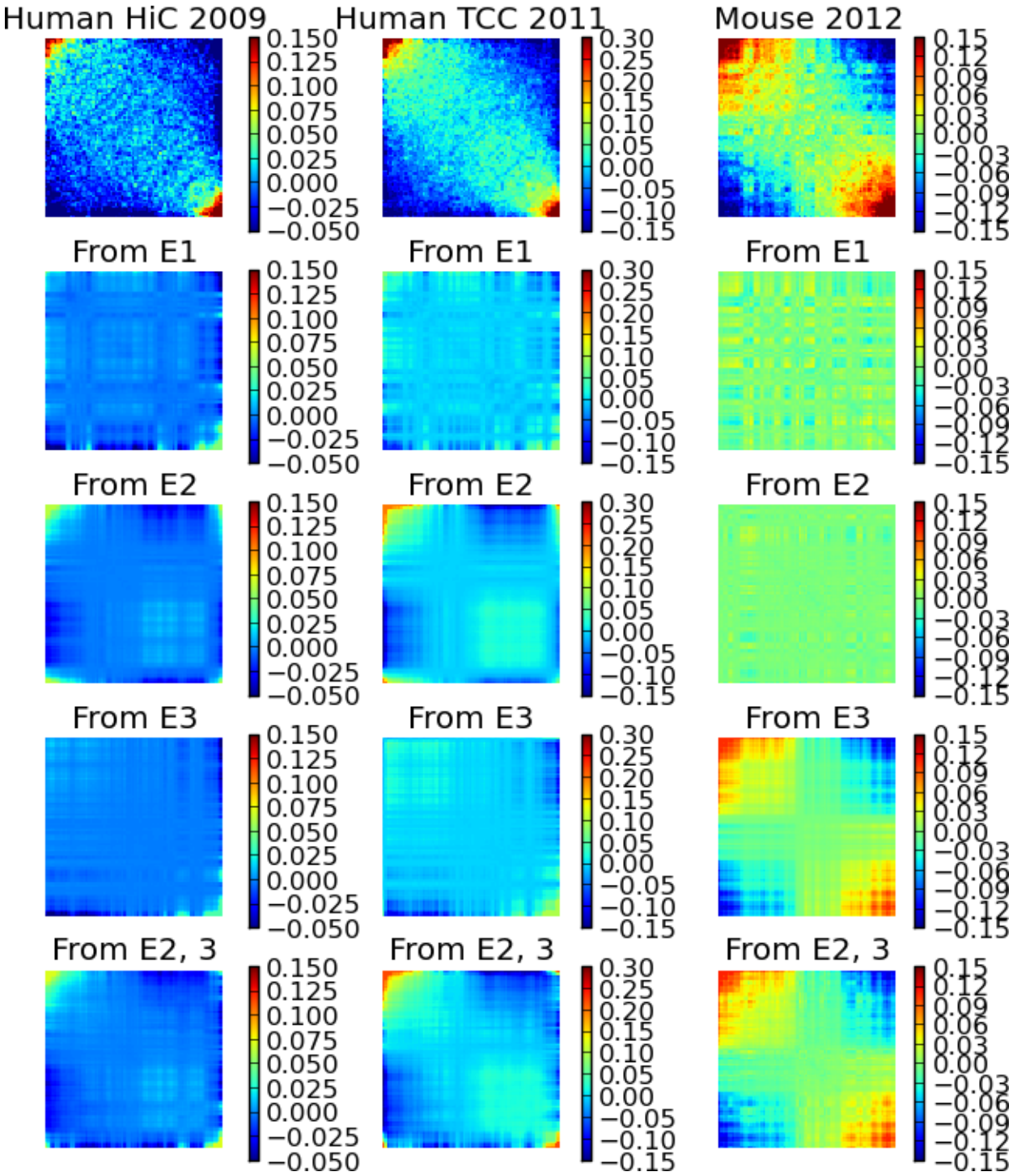



**Supplementary figure 8. Average inter-arm contact maps.**

Average inter-chromosomal inter-arm maps show log-enrichment of contact probability for interaction between two chromosomal arms, similarly to Fig. 4b. Axes for each map are the relative location along a chromosomal arm, from centromere (top left) to telomere (bottom right). Here, average inter-arm maps were obtained from the raw data (*top*), from iteratively corrected inter- and intra-chromosomal data without SS reads (*middle top*), with SS reads (*middle bottom*), and iteratively corrected trans reads only (*bottom*). Average trans maps were calculated at 1Mb resolution, using the method presented in Online Methods.



Nature Methods: doi:10.1038/nmeth.2148

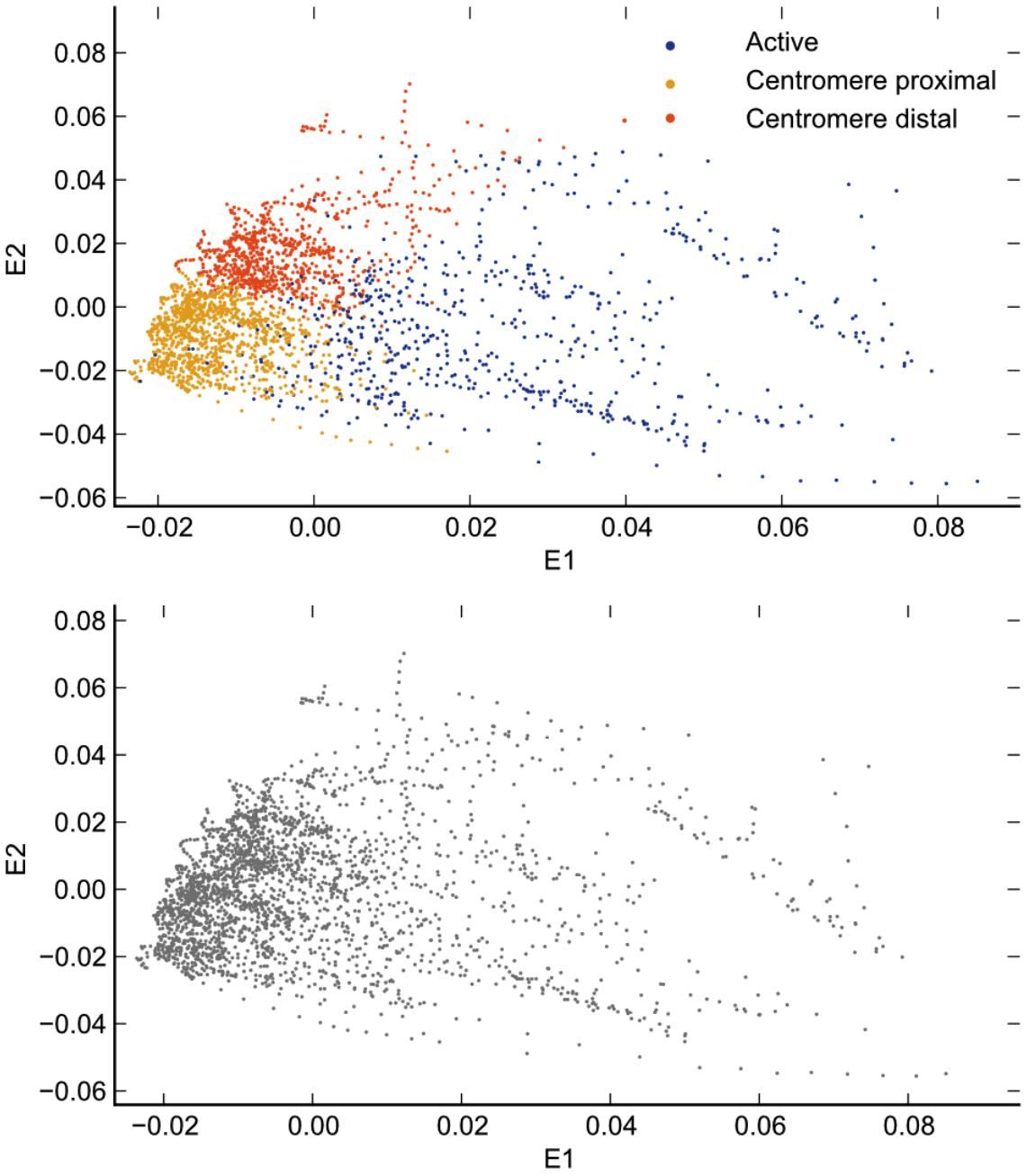



**Supplementary figure 9. Inter-arm maps reconstructed from the first three eigenvectors.**

Inter-arm maps as in Fig. 4b, were calculated from iteratively corrected data (top row) and from projections of the iteratively corrected data on eigenvectors E1, E2 and E3. Color schemes are consistent within each dataset. $E^1$ makes a very limited contribution to the average inter-arm map for both Human and Mouse datasets; a combination of $E^2$ and $E^3$ eigenvectors resembles the original map. In acrocentric mouse chromosomes, $E^3$ is responsible for both centromere and telomere attraction, while the role of $E^2$ is unknown.



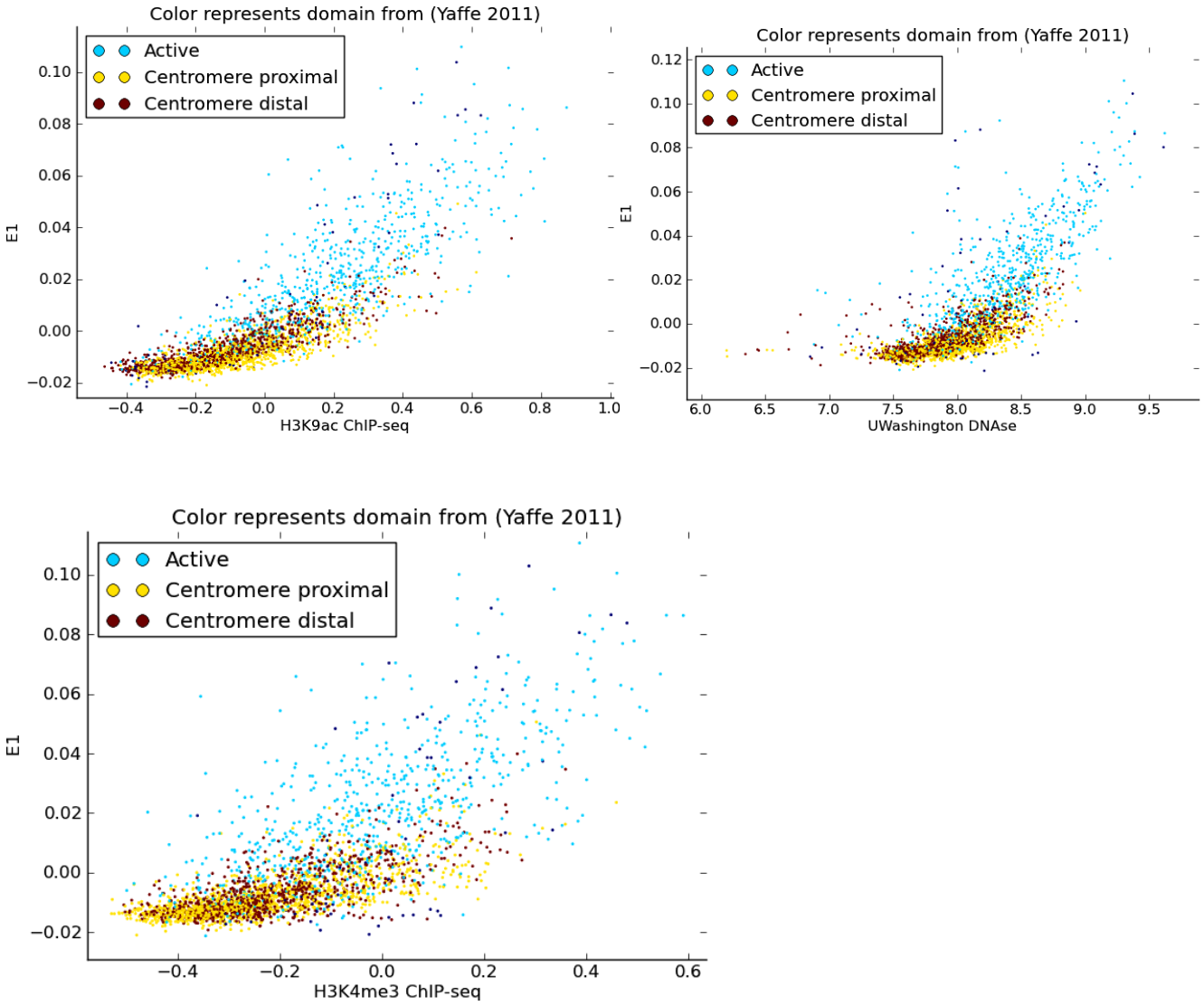
Nature Methods: doi:10.1038/nmeth.2148




**Supplementary Figure 10: Previously suggested chromatin clusters shown in the space of leading eigenvectors.** One-megabase data from Yaffe et al., 2011[2] projected on the space of leading eigenvectors obtained for the same (smoothed) data. (top) Colored by chromatin types (clustered) identified by Yaffe and Tanay, 2011[2] using k-means (k=3) clustering. (bottom) The same data, not colored. Notice the lack of clear clusters and rather ambiguous cluster assignment by k-means.



Nature Methods: doi:10.1038/nmeth.2148



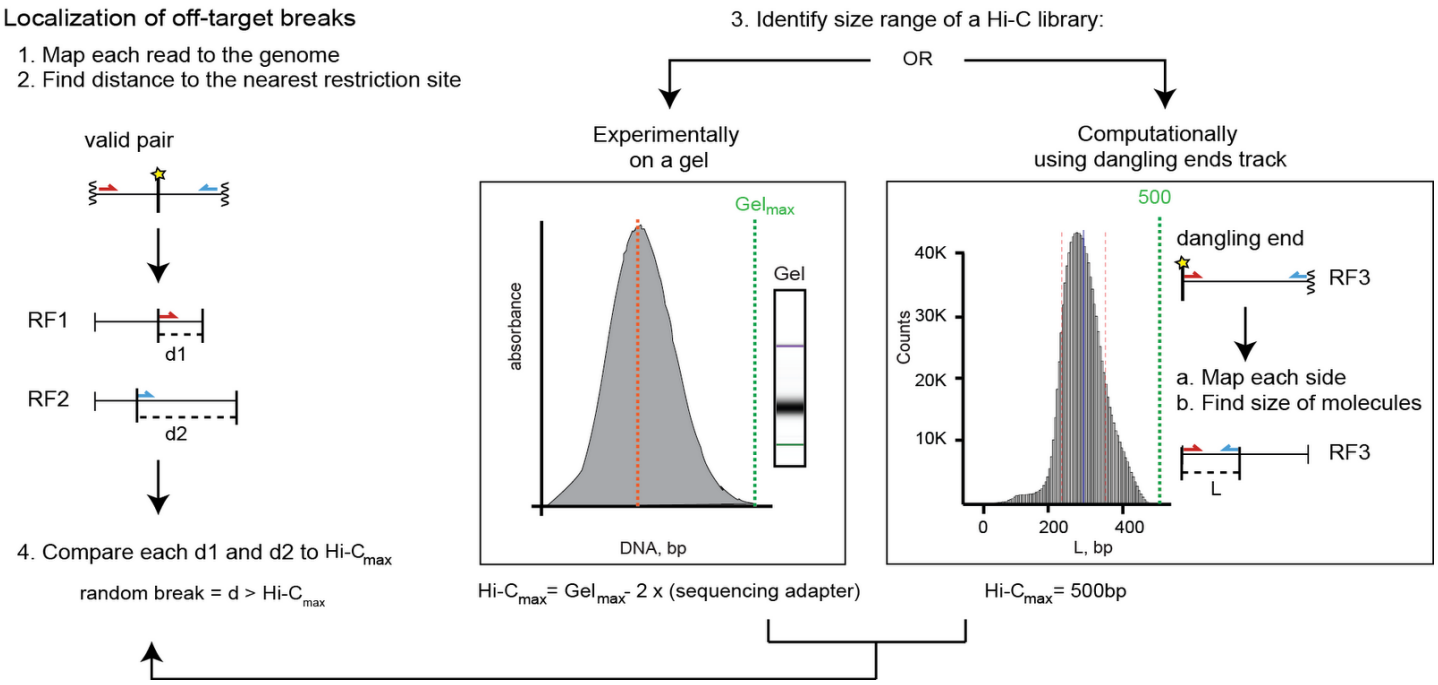

**Supplementary Figure 11: The first eigenvector correlates with functional genomic features better than previously identified chromatin clusters.**

The first eigenvector *E1* is plotted against H3k9ac (*top left*), DNAse I (*top right*) and H3k4me3 (*bottom*) for 1Mb regions and colored by chromatin cluster types identified by Yaffe and Tanay, 2011[2].

This analysis shows that *E1* is a better predictor of genomic features than previously identified clusters. Correlation of $E^1$ with each feature holds not only genome-wide, but also within a cloud of each color; this allows *E1* to explain a greater fraction of the genome-wide variance of these histone marks, as compared with mean values for k-means clusters (69% vs. 41% for H3K9ac, 55% vs. 39%, and 53% vs. 42% for H3K4me3).



Nature Methods: doi:10.1038/nmeth.2148

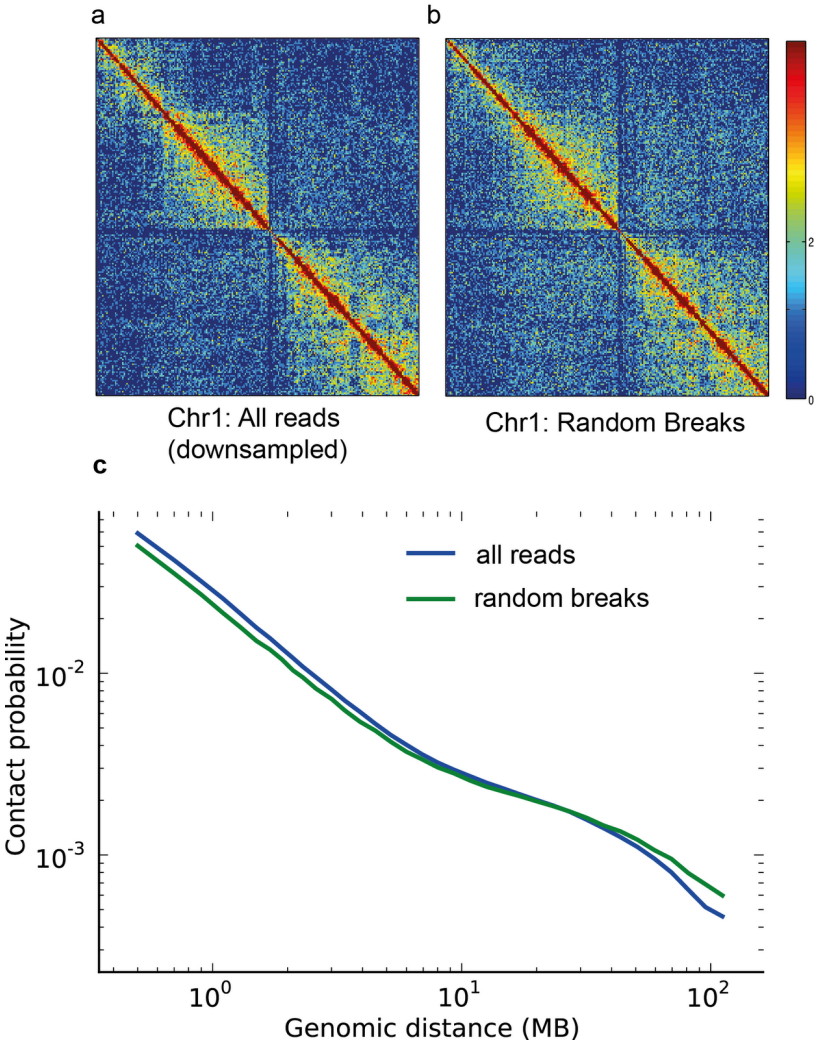



**Supplementary Figure 12: Determination of random breaks assignment in Hi-C libraries.**

Random breaks are identified individually for each side of the read pair by a comparison with the maximum linear length of molecules in a Hi-C library. This maximum size can be measured experimentally on a gel (after accounting for adapter length) or estimated computationally by plotting the size distribution of dangling end molecules. Random breaks are reads which are further from their closest possible restriction cut site than this maximum size. Gel analysis was not publicly available for data sets analyzed here, and the computational method was used instead.



Nature Methods: doi:10.1038/nmeth.2148




**Supplementary Figure 13: Properties of random breaks.**

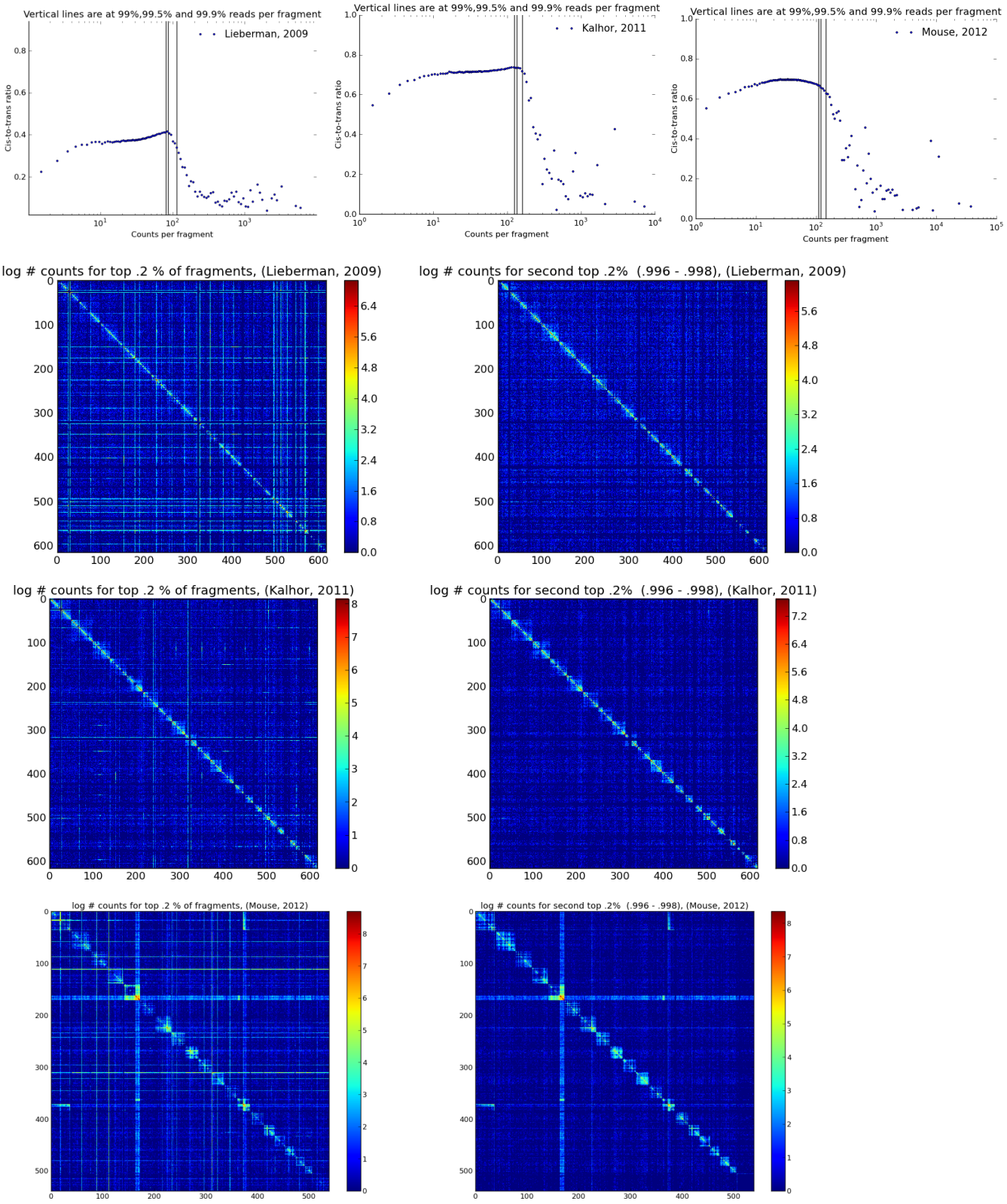
**a:** Raw heatmap of all chromosome 1 double-sided reads for Hi-C HindIII data, down-sampled using binomial sampling to match the number of reads in the random break heatmap.

**b:** Heatmap of random break double-sided reads; for random breaks, at least one side was further from the nearest restriction site than the maximum size of molecules in the library, 500bp.

**c:** Scaling of the contact probability with genomic separation. This scaling for random breaks are very similar to the scaling for all reads. Scalings were calculated as described in Online Methods.



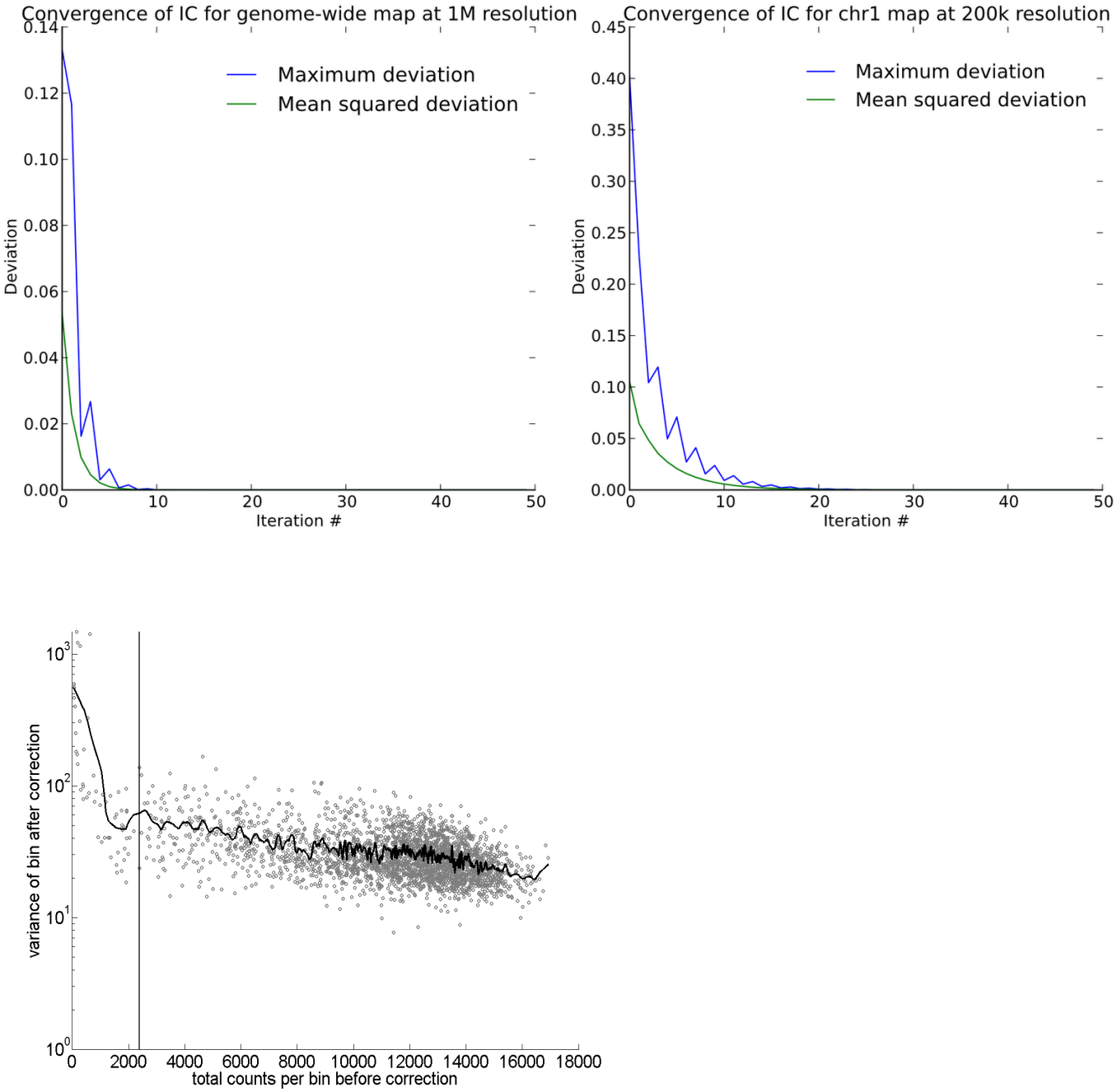
Nature Methods: doi:10.1038/nmeth.2148




**Supplementary Figure 14: Properties and filtering of high-count fragments.** (*top*) Cis/trans ratio plotted as a function of # reads per fragment. It decreases significantly for fragments with the largest # of counts, suggesting that these counts are erroneous due to mis-mapping to repetitive genomic sequences located on many chromosomes (potentially from poor genome assembly), fragile sites, or random inter-molecular ligation in solution. (*bottom*) Heatmaps reconstructed from interactions of top 0.2% of fragments show uniform stripes (random ligations), short stripes in trans (probably, mis-mapping), and generally produce noisier heatmaps than heatmaps similarly obtained from the second top 0.2% of fragments.



Nature Methods: doi:10.1038/nmeth.2148



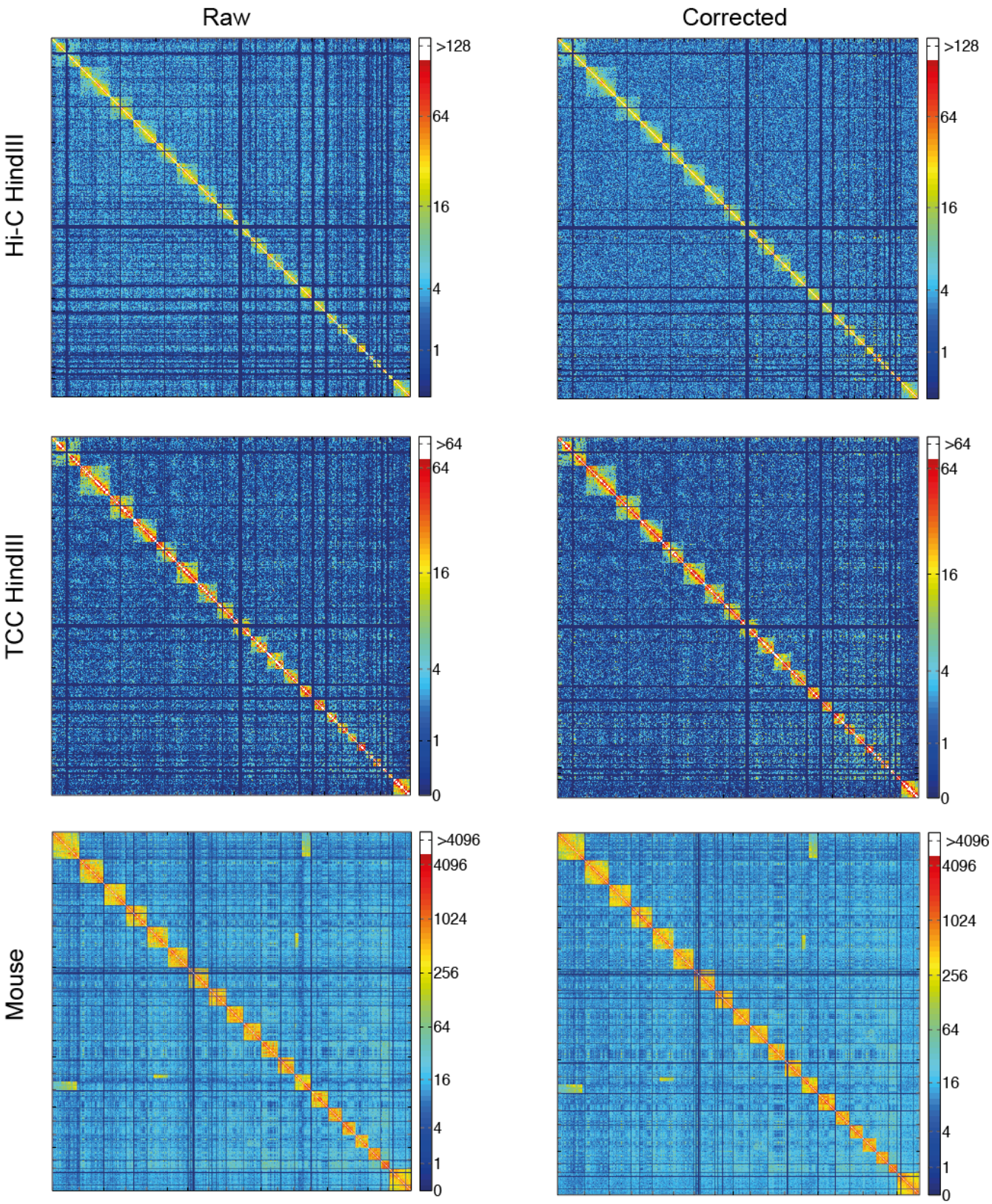

**Supplementary figure 15: Properties of iterative correction.**

(*top*) Convergence of maximum relative deviation and sqrt(mean square relative deviation) of biases during iterative correction, as compared to the final biases obtained with 100 iterations. 25 iterations were always enough for IC to converge. Throughout the paper, 50 iterations were used, twice the worst encountered case (intra-chromosomal map for chromosome 1). Shown for Hi-C HindIII data.

(*bottom*) Variance of bins with low # counts increases significantly after iterative correction. To avoid this, we remove 2% of bins with lowest # of counts, where 2% cutoff is indicated by the vertical line.



Nature Methods: doi:10.1038/nmeth.2148




**Supplementary figure 16: Comparative performance of iterative correction across datasets.**

As in Fig1b,c, but showing filtered bins (and bins with no counts) as rows and columns of all zeros. Hi-C HindIII from Lieberman-Aiden et al., 2009[1], TCC from Kalhor et al., 2011[3], and Mouse from Zhang and McCord et al., 2012[4]. Note the clearly visible translocations for mouse chromosome 13, which appear as bright rectangles in the inter-chromosomal map.



Nature Methods: doi:10.1038/nmeth.2148

**Supplementary table 1: Spearman correlations of Human and Mouse eigenvectors with genomic and epigenomic features.**

Raw genomic tracks and matching control tracks were averaged over 5kb windows spawning the entire genome. To obtain enrichment of a particular 1MB bin, track-to-control log-ratios were averaged over all 5kb regions within the bin, ignoring the regions with zero # of track or control counts. The latter was done to ensure proper averaging over bins, containing significant amount of non-mappable regions. DNAse tracks had no control, and log(# counts) was averaged over all 5kb windows. For human, ENCODE[5] genomic tracks were downloaded from UCSC: http://hgdownload.cse.ucsc.edu/goldenpath/hg18/encodeDCC/wgEncodeBroadChipSeq/. For mouse, C2C12 tracks were downloaded from http://hgdownload.cse.ucsc.edu/goldenpath/mm9/encodeDCC/wgEncodeCaltechHist/, and B-cell DNAse was downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/mm9/encodeDCC/wgEncodeUwDnase/.

| **Human** | GC | EigLieberman2009 | *E1* HiC | *E1* TCC |
|---|---|---|---|---|
| H3K27me3 | 0.2383 | 0.1606 | 0.1249 | 0.1953 |
| H3K4me3 | 0.6609 | 0.6328 | 0.7584 | 0.7454 |
| H3K4me1 | 0.7397 | 0.7703 | 0.8460 | 0.8980 |
| H3K36me3 | 0.5784 | 0.6329 | 0.7206 | 0.7415 |
| H3K4me2 | 0.7386 | 0.7597 | 0.8504 | 0.8933 |
| H4K20me1 | 0.8347 | 0.6363 | 0.7221 | 0.7752 |
| H3K9ac | 0.7677 | 0.7766 | 0.8684 | 0.9226 |
| H3K27ac | 0.6436 | 0.7282 | 0.8055 | 0.8516 |
| DNAse | 0.7238 | 0.7333 | 0.7899 | 0.8429 |
| CTCF | 0.7306 | 0.6477 | 0.7711 | 0.7877 |
| GC | 1 | 0.6300 | 0.8016 | 0.7926 |
| EigLieberman2009 | 0.6300 | 1 | 0.7849 | 0.8367 |
| *E1* HiC | 0.8016 | 0.7849 | 1 | 0.9283 |
| *E1* TCC | 0.7926 | 0.8367 | 0.9283 | 1 |



Nature Methods: doi:10.1038/nmeth.2148

| **Mouse** | GC | *E1* |
|---|---|---|
| H3k36me3 | 0.728 | 0.734 |
| H3ac | 0.783 | 0.719 |
| H3k27me3 | 0.468 | 0.244 |
| H3k73me3 | 0.744 | 0.702 |
| H3k79me2 | 0.810 | 0.751 |
| H3k04me3 | 0.915 | 0.758 |
| UW_DNAse | 0.849 | 0.773 |
| GC | 1 | 0.709 |



Nature Methods: doi:10.1038/nmeth.2148

## Supplementary Note

*Maximum Likelihood Estimate (MLE) of relative contact probability*

Here we show that the equation used for iterative correction of Hi-C data is the MLE of the relative contact probability. As in the main text, $O_{ij}$ is the observed count of reads for a pair of loci *i* and *j*, $T_{ij}$ is the sought matrix of "true" relative contact probabilities that is normalized for equal visibility $\Sigma_i T_{ij}=1$, and $B_i$ is a vector of biases. We assume that the observed count $O_{ij}$ is a realization of some underlying probability distribution with pdf given by *f(.)* and the expected value given by $T_{ij}B_iB_j$, that is

$$O_{ij} \leftarrow f\left(O_{ij};T_{ij}B_iB_j\right).$$

Given the distribution *f(.),* the matrix $T_{ij}$ and the vector $B_i$ one can calculate the likelihood *L* of the observation as

$$L = \prod_{ij} f(O_{ij};T_{ij}B_iB_j),$$

and then find $B_i$ and $T_{ij}$ that maximize the likelihood.

First consider the case of the Poisson distribution: $f(O;E) = E^O/O!\exp(-E)$, then the log-likelihood, *LL,* of the data is

$$LL = \sum_{ij}\left[O_{ij}\log\left(T_{ij}B_iB_j\right) - T_{ij}B_iB_j - \log\left(O_{ij}!\right)\right]$$

Differentiating with respect to $T_{lm}$ and with respect to $B_m$ and setting derivatives to zero gives

$$\partial LL/\partial T_{lm} = O_{lm}/T_{lm} - B_lB_m = 0; \qquad T_{lm} = O_{lm}/B_lB_m$$

$$\partial LL/\partial B_m = \sum_i\left[O_{im}/B_m - T_{im}B_i\right] = 0; \quad \sum_i\left[O_{im}/B_m\,B_i - T_{im}\right] = 0$$

This system is redundant as the second equation is satisfied for any *m* if the first one is satisfied. Taking the first equation together with the normalization of $T_{ij}$ yields

$$\boxed{\sum_i O_{ij}/B_iB_j = 1},$$

which is solved by iterations to find $B_i$ and then find $T_{ij} = O_{ij}/B_iB_j$.

It is easy to see that other distributions (e.g. exponential) give the same result. In fact, the same result is obtained for a broad class of probability distribution functions as long as $\partial f(O;E)/\partial E = 0$ is satisfied by $O = E$, which is equivalent to the consistency of estimation of *E* for a sample $\{O^{(k)}\}$ of size *n*: $\lim_{n\to\infty}\sum_{k=1}^{n} O^{(k)}/n = E$.

In summary, we demonstrated that for a broad class of distributions that generate observed counts from expected, the matrix of relative contact probabilities can be found by iterative correction explained in the paper and the Online Methods.

*Hardware and Software Requirements*

The library of Hi-C tools we present here is written in Python, and has been tested only on linux/unix operating systems. The library consists of three parts: i) Iterative Mapping, ii) fragment-level filtering and analyses, iii) binned data analysis, including Iterative Correction.

Iterative mapping is performed using bowtie2, and requires at least 8GB RAM (16GB preferred). Fragment-based analysis requires approximately 20-30 bytes per read, dependent on the filters used. For example, a 500 million read Hi-C dataset will require no more than 16GB RAM. Current implementation of binned data analysis allows for working with the entire human genome at 200kb resolution using 8GB RAM, or 100kb using 32GB.

The time-liming step of our method is mapping, and is currently the only part of the code which utilizes multi-threading. Using bowtie2, it takes about 2 days to map 500 million reads on a single high-performance CPU. Downstream analysis tools are much quicker. Fragment-based filtering of a dataset can be performed in about 1 hour for 500 million raw reads, and is limited by HDD performance. The binned data analysis can be performed in minutes (seconds, depending upon resolution).



Nature Methods: doi:10.1038/nmeth.2148

*Bin Size for Iterative Correction*
Selecting an optimal bin size is very important for obtaining accurate corrected Hi-C maps. The main factor that determines accuracy of corrected Hi-C maps is a minimum number of reads in a bin before correction; if this number is $N_{min}$, then maximum relative error would be $1/\mathrm{sqrt}(N_{min})$. This value depends on resolution, library complexity, inclusion of cis reads, and on the number of low-visibility bins excluded from the analysis. Our software keeps track of this number, and issues a warning if this number is less than 100 (maximum relative error of 10%). However, for all analyzed datasets the threshold of 10% was never reached, even for lowest-read-number dataset at 200kb resolution with 2% of low-count bins removed.

*Computational requirements and resolution*
There are two challenges for extending iterative correction to higher resolutions: the first relates to Hi-C library coverage and complexity, and the second relates to computational complexity. Ultimately, we believe that neither will pose an insurmountable hurdle. In terms of computational resources, our current implementation allows analysis of the Human genome data at 50-200 kb resolution with modern desktop machines (64GB - 8GB RAM respectfully). However, larger matrices are not conceptually an issue for extending our method to a higher resolution, or even single-fragment resolution, as sparse matrix logic can be used; in this case the complexity of the algorithm is linear in the number of reads.

## References


1. E. Lieberman-Aiden, N. L. van Berkum, L. Williams et al., *Science* **326** (5950), 289 (2009).
2. E. Yaffe and A. Tanay, *Nature genetics* **43** (11), 1059 (2011).
3. R. Kalhor, H. Tjong, N. Jayathilaka et al., *Nature biotechnology* **30** (1), 90 (2011).
4. Y. Zhang, R. P. McCord, Y. J. Ho et al., *Cell* **148** (5), 908 (2012).
5. E. Birney, J. A. Stamatoyannopoulos, A. Dutta et al., *Nature* **447** (7146), 799 (2007).




Nature Methods: doi:10.1038/nmeth.2148