
DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU


HARVARD LIBRARY
Office for Scholarly Communication

Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization

Citation
Imakaev, Maxim, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, and Leonid A Mirny. 2013. "Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization." Nature methods 9 (10): 10.1038/nmeth.2148. doi:10.1038/nmeth.2148. http://dx.doi.org/10.1038/nmeth.2148.

Published Version
doi:10.1038/nmeth.2148

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:11878982

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

NIH Public Access
Author Manuscript
Nat Methods. Author manuscript; available in PMC 2013 November 04.

Published in final edited form as:
Nat Methods. 2012 October ; 9(10): . doi:10.1038/nmeth.2148.

Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization

Maxim Imakaev[1,*], **Geoffrey Fudenberg**[2,*], **Rachel Patton McCord**[3], **Natalia Naumova**[3], **Anton Goloborodko**[1], **Bryan R. Lajoie**[3], **Job Dekker**[3,#], and **Leonid A Mirny**[1,2,4,#]

[1]Department of Physics, MIT, Cambridge, MA

[2]Graduate Program in Biophysics, Harvard University, Cambridge, MA

[3]Program in Systems Biology. Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA

[4]Institute for Medical Engineering and Science, MIT, Cambridge, MA

Abstract

Extracting biologically meaningful information from chromosomal interactions obtained with genome-wide chromosome conformation capture (3C) analyses requires elimination of systematic biases. We present a pipeline that integrates a strategy for mapping of sequencing reads and a data-driven method for iterative correction of biases, yielding genome-wide maps of relative contact probabilities. We validate ICE (Iterative Correction and Eigenvector decomposition) on published Hi-C data, and demonstrate that eigenvector decomposition of the obtained maps provides insights into local chromatin states, global patterns of chromosomal interactions, and the conserved organization of human and mouse chromosomes.

Introduction

Obtaining views of genomic organization and function free from experiment-induced biases remains a major challenge for any genome-scale study. The raw outputs of many genomic technologies are affected both by technical biases, including those from sequencing and mapping[2,3], and biological factors, such as those resulting from intrinsic physical properties of distinct chromatin states[4]. As a result, different regions of the genome appear to have different experimental "visibility", making it difficult to compare their contributions, and potentially leading to false-positives or false-negatives. Recently-developed high-throughput 3C-based methods[5-12] (for example Hi-C) for investigating physical contacts between distal genomic loci have begun to provide key insights into the spatial organization of genomes[7-11,13-16]. However, the raw outputs of 3C-based methods may be influenced by various forms of biases[17].

Here, we present ICE (Iterative Correction and Eigenvector decomposition), a pipeline that includes processing paired sequence reads obtained from genome-wide 3C-based methods[8-11,14] and a method of iterative correction, which eliminates biases and is based on the assumption that all loci should have equal visibility (Fig. 1). Iterative correction

[#]Corresponding authors: Job.Dekker@umassmed.edu; leonid@MIT.edu.
[*]Contributed equally

Author Contributions: IM developed the iterative correction procedure. IM, GF developed data analysis tools. IM, AG, developed and maintain publicly available software. IM, GF, RPM, AG performed data analysis. IM, GF, RPM, NN, AG, BL, JD and LM contributed to conceiving the study and wrote the paper. Authors claim no competing financial interests. Authors are thankful to K. Korolev for many productive discussions.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

leverages the unique pairwise and genome-wide structure of Hi-C data to decompose contact maps into a set of biases and a map of relative contact probabilities between any two genomic loci (Fig. 1b,c), achieving equal visibility across all genomic regions. The obtained corrected interaction maps can then be further decomposed into a set of genome-wide tracks (eigenvectors) describing several levels of higher-order chromatin organization (Fig. 1e). We apply our pipeline to three datasets from a human lymphoblastoid cell line: two datasets generated by Hi-C[8] using either HindIII (Hi-C HindIII) or NcoI (Hi-C NcoI) digestion, and one generated by a Hi-C variant, Tethered Chromosome Capture[10], using HindIII digestion (TCC). We also analyze one HindIII-digested mouse pro-B cell Hi-C dataset[16].

Results

Read alignment and classification

Our pipeline begins with the alignment of read-pairs obtained from genome-wide 3C-based methods to a reference genome. To account for the specific structure of Hi-C ligation products, we align the first portion of each read, truncating the read to a certain length, and then aggregate alignments over increasing truncation lengths (Fig. 1a, Supplementary Fig. 1, and Online Methods). This procedure yields many more double-sided mapped reads than using a fixed truncation length (Fig. 1a). After alignment, the pipeline discards molecular byproducts (Supplementary Figs. 1 and 2, and Online Methods). The remaining read-pairs include: double-sided reads (DS reads), which represent a contact between two mappable portions of the genome, and single-sided reads (SS reads), which often represent a contact between a mappable and an unmappable portion of the genome (Fig. 1a). SS reads make an important contribution to the total coverage in peri-centromeric regions, where decreased intra-chromosomal DS coverage balances a reciprocal increase in the SS coverage (Fig. 1d).

Iterative Correction

The next step removes biases introduced by experimental procedures and by intrinsic properties of the genome, and converts observed Hi-C maps into corrected maps of relative contact probabilities (Fig. 1b,c). We do not assume specific sources of biases and correct collectively for all factors affecting experimental visibility, including DNA sequencing bias or restriction site density. We assume, and demonstrate below, that the bias for detecting contacts between two regions can be represented as the product of the individual biases of these regions. Given this assumption of factorizable biases, the expected contact frequency, E_{ij}, for every pair of regions, (i,j), can be written as: $E_{ij} = B_i B_j T_{ij}$, where B_i and B_j are the biases and T_{ij} is the sought matrix of relative contact probabilities, normalized as $\sum_{i \neq j \pm 1} T_{ij} = 1$. This normalization ensures a uniform coverage profile, equal visibility of each region in an iteratively corrected contact map (Fig. 1c), can reveal specific interactions otherwise buried by visibility-induced biases (Fig. 2a), and allows unbiased comparisons within and between Hi-C datasets. Since an experiment represents a sample from a distribution of possible interactions, the observed interaction frequency is a realization from some distribution with expectation E_{ij}. For a range of distributions, the maximum likelihood solution for biases B_i is obtained by iteratively solving a system of equations (iterative correction), yielding a corrected Hi-C map. We note that this procedure can be extended to include single-sided reads (Supplementary Fig. 3).

We validate our assumption of factorizable biases by analyzing inter-chromosomal biases inferred via a recently proposed computationally-intensive machine learning procedure[17]. This study calculated a matrix of biases, B_{ij}, by explicitly considering restriction fragment level biases associated with fragment length, GC content, and mappability at megabase resolution. We find that B_{ij} can be accurately described as a product of two vectors of biases ($B_{ij} \approx B_i B_j$), explaining 99.99% of the variance (Fig. 2b). Iteratively corrected inter-

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

chromosomal data is highly correlated with previously obtained corrected maps[17] ($r = 0.98$, here and below Spearman correlation, $P < 10e\text{-}10$, Supplementary Fig. 4). Since known biases are factorizable, uncharacterized biases are likely to be factorizable, and would be removed by ICE.

To validate our method, we first compare Hi-C maps obtained using different restriction enzymes (Fig. 2c,d). In raw data, the correlation between Hi-C data generated with different enzymes can be quite low due to enzyme-dependent biases. Corrected maps show an increased between-enzyme correlation of corresponding off-diagonal intra-chromosomal elements (Fig. 2c). Iterative correction also increases between-enzyme correlation for inter-chromosomal maps to the level of correlation between halves of the same dataset (Fig. 2d and Supplementary Fig. 5a). To compare to a previous method[17] we applied the same smoothing technique and obtained a similar between-enzyme correlation $r=0.71$ ($r=0.59$ obtained earlier[17]). Next, we perform cross-validations using 10% or 90% of the read-pairs and obtain biases that are highly correlated ($r=0.98$, $P < 10e\text{-}10$, HindIII), demonstrating that our method does not over-fit (Fig. 2e). We also note that an important property of intra-chromosomal maps, the decay of contact probability with genomic distance, remains unchanged after correction (Fig. 2e).

Previous attempts to correct Hi-C data used a single division by a product of the visibilities of two regions[8,11,17]. Applying this procedure once only partially corrects for non-uniform coverage (Fig. 2c), tends to flip the coverage profile (Supplementary Fig. 5c), and leads to a solution that depends on the initial normalization of the data, thus making results of the correction unpredictable. However, applying this procedure iteratively eliminates all factorizable biases, leads to uniform coverage, and obtains better agreement between datasets (Fig. 2c,d).

Eigenvector analysis reveals patterns of chromosomal organization

The next step in ICE analysis decomposes an iteratively corrected genome-wide map into a series of genomic tracks to reveal the main features of higher-order chromosomal organization (Fig. 3 and Online Methods). Each track k represents interaction preferences (E^k_i) of genomic region i. Independent interaction preference tracks E^k can be found as eigenvectors of the corrected map T_{ij} ($T_{ij} = \sum_k \lambda_k \cdot E^k_i \cdot E^k_j + \text{const}$), where the relative weights of their contributions λ_k are the corresponding eigenvalues. The contribution of each track to the total interaction frequency between a pair of regions in the corrected map T_{ij} is proportional to a product of these preferences ($E^k_i \cdot E^k_j$). Eigenvectors are then sorted (E^1, E^2, E^3…) in descending order by the magnitude of their corresponding eigenvalues. Our decomposition operates directly on corrected Hi-C data, unlike a previous method that makes several additional transformations of the data[8]. Permutation analysis shows that the first 13 eigenvectors are statistically significant ($P < .001$). Moreover, the first three are robust to details of the experiment (Fig. 3e, Supplementary Fig. 6 and Online Methods) and explain 72% of the inter-chromosomal data reconstructed from the first 13 eigenvectors. Thus, we focus on the first three eigenvectors for further analysis of inter-chromosomal interaction preferences.

The leading eigenvector, E^1, provides a genomic track of inter-chromosomal interaction preferences along the genome, and shows correlation with many genomic features (Fig. 3a,b), including GC content ($r = .80$, $P < 1e\text{-}10$), replication timing ($r = .82$, $P < 1e\text{-}10$, GEO GSM500943), DNAse I hypersensitivity ($r = .79$, $P < 1e\text{-}10$, GEO GSE4334) and many histone marks (Supplementary Table 1). The profile of E^1 is similar to chromatin compartments found previously[8], yet E^1 shows higher correlation with many genomic features[18] both along the chromosomes and for average values of whole chromosomes (Fig. 3b, $r=0.95$, $P=4e\text{-}06$, vs $r=-0.31$ for chromatin compartments, Supplementary Fig. 6).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Interaction preferences represented by E^1 connect spatial and functional genomic organization, as regions with high E^1, which are gene-rich and enriched for active chromatin marks, tend to interact more with other similar regions (Fig. 3c). Conversely, gene-poor regions with low E^1 tend to interact more with other gene-poor regions. Despite its tendency to partition active and inactive regions of the genome, E^1 does not show any bimodality (Fig. 3d, left). Neighboring genomic regions display similar interaction preferences as seen from the autocorrelation (Fig. 3d, right) that decays with a characteristic length of about 6Mb. Taken together, these characteristics of E^1 suggest that continuous interaction preferences better capture the complexity of chromatin interaction landscape at megabase resolution than a two-compartment model[8] proposed earlier.

Furthermore, we find evidence for the evolutionary conservation of genome-wide chromosome organization by comparing E^1 for human and mouse datasets. E^1 has high correlation ($r = 0.81$, $P < 1e\text{-}10$) in syntenic regions[19] of human and mouse genomes at the megabase level (Fig. 4a). Moreover, the conservation of E^1 cannot be explained by a confounding effect of similar GC content profiles as demonstrated by a GC-content stratified permutation test (Fig. 4a, Online Methods).

We then study the interaction preference tracks, E^2 and E^3, which constitute the greatest contributions to the corrected map after E^1. Both E^2 and E^3 vary with position along chromosomal arms (Figs. 1e and 3f), with increased magnitude near centromeres for E^2 and telomeres for E^3 (Supplementary Fig. 8) This pattern of interaction is prominent on average inter-arm maps, which reveal an enrichment of centromere-centromere and telomere-telomere contacts (Fig. 4b). Average inter-arm maps constructed from projections of the data on E^2 and/or E^3, but not E^1, show a similar pattern of contact enrichment, directly confirming that arm-level organization is largely captured by E^2 and E^3 (Supplementary Fig. 8). This pattern is consistent with co-localization of centromeres and a similar co-localization of telomeres, as described in imaging studies[20,21]. We observe a consistent pattern of contact enrichment for all studied human and mouse datasets, despite the acrocentric structure of mouse chromosomes (Fig. 4b). For the mouse dataset, both centromere-centromere and telomere-telomere enrichment are captured by E^3 (Supplementary Fig. 8), while E^2 refines the signal. The consistent pattern of average inter-arm maps suggests that interactions between chromosomal arms are among the most prominent features of higher-order chromatin organization in the human and mouse genomes[20,21]

Multiple attempts have been made to identify distinct chromatin types based on Hi-C data[8,17]. We compare the E^1 and E^2 representation of inter-chromosomal interactions to a model of three chromatin types identified earlier by k-means clustering[17] (Fig. 3g). We find that the suggested clusters do not show evident separation and the suggested division into three chromatin types is ambiguous[22] (Supplementary Fig. 9). [22]We also note that E^1 captures variation in epigenomic tracks much better than the three chromatin types (Supplementary Fig. 10).

Discussion

By requiring equal visibility of genomic loci, the iterative correction in ICE yields a matrix of relative contact probabilities. This approach preserves and highlights specific contacts, simultaneously ensuring that high-frequency contacts cannot be explained solely by elevated visibilities of participating loci (Fig. 2a). Iterative correction can be used to reveal relative contact probabilities of contact maps for individual chromosomes or for the genome-wide inter-chromosomal contact map. Most importantly, it allows unbiased comparison of Hi-C data within and between datasets, cell types and organisms.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

We note that our data-driven method is specific to techniques that yield a pairwise and genome-wide matrix of contacts; while other 3C-based methods that do not yield all-by-all interaction maps have similar systematic biases (4C[5,6], 5C[7]), they must be corrected using an alternate approach[17]. We also note that iterative correction operates on binned data, and thus does not correct Hi-C data at resolutions below a chosen bin size (here, 200 kb and 1 Mb). However, with sufficient sequencing depth, iterative correction can be performed at increasingly high resolution, potentially up to that of a single restriction fragment (see Supplementary Text).

Our analysis of inter-chromosomal Hi-C data suggests that at megabase resolution, 3D genomic organization depends upon at least two continuous features: one that relates to genomic sequence and local epigenetic chromatin states, and a second related to position along the chromosome arm. The first feature further suggests that interphase chromatin folding may be encoded by a combination of the genomic sequence itself and local chromatin activity. In combination, these two features constitute the experimentally robust signal in recent datasets. Moreover, the prominence of these features is remarkably consistent between human and mouse genomes. Taken together, our analysis implicates these features as general principles of mammalian interphase inter-chromosomal organization.

ICE is available at http://bitbucket.org/mirnylab/hiclib

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The work of IM, GF, AG and LM was supported by the National Cancer Institute Physical Sciences-Oncology Center at MIT (U54CA143874). This work was supported by NIH Grants HG003143 (to JD), F32GM100617 (to RPM), and a W.M. Keck Foundation Distinguished Young Scholar in Medical Research Award (To JD).

References

1. Chhabra SR, Butland G, Elias DA, et al. Appl Environ Microbiol. 2011; 77:7595–7604. [PubMed: 21908633]
2. Cheung MS, Down TA, Latorre I, et al. Nucleic acids research. 2011; 39:e103. [PubMed: 21646344]
3. Quail MA, Kozarewa I, Smith F, et al. Nature methods. 2008; 5:1005–1010. [PubMed: 19034268]
4. Teytelman L, Ozaydin B, Zill O, et al. PloS one. 2009; 4:e6700. [PubMed: 19693276]
5. Simonis M, Klous P, Splinter E, et al. Nature genetics. 2006; 38:1348–1354. [PubMed: 17033623]
6. Zhao Z, Tavoosidana G, Sjolinder M, et al. Nature genetics. 2006; 38:1341–1347. [PubMed: 17033624]
7. Dostie J, Richmond TA, Arnaout RA, et al. Genome research. 2006; 16:1299–1309. [PubMed: 16954542]
8. Lieberman-Aiden E, van Berkum NL, Williams L, et al. Science. 2009; 326:289–293. [PubMed: 19815776]
9. Duan Z, Andronescu M, Schutz K, et al. Nature. 2010; 465:363–367. [PubMed: 20436457]
10. Kalhor R, Tjong H, Jayathilaka N, et al. Nature biotechnology. 2011; 30:90–98.
11. Sexton T, Yaffe E, Kenigsberg E, et al. Cell. 2012; 148:458–472. [PubMed: 22265598]
12. Dekker J, Rippe K, Dekker M, et al. Science. 2002; 295:1306–1311. [PubMed: 11847345]
13. van Steensel B, Dekker J. Nature biotechnology. 2010; 28:1089–1095.
14. Dixon JR, Selvaraj S, Yue F, et al. Nature. 2012; 485:376–380. [PubMed: 22495300]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

15. Nora EP, Lajoie BR, Schulz EG, et al. Nature. 2012; 485:381–385. [PubMed: 22495304]
16. Zhang Y, McCord RP, Ho YJ, et al. Cell. 2012; 148:908–921. [PubMed: 22341456]
17. Yaffe E, Tanay A. Nature genetics. 2011; 43:1059–1065. [PubMed: 22001755]
18. Birney E, Stamatoyannopoulos JA, Dutta A, et al. Nature. 2007; 447:799–816. [PubMed: 17571346]
19. Kent WJ, Baertsch R, Hinrichs A, et al. Proceedings of the National Academy of Sciences of the United States of America. 2003; 100:11484–11489. [PubMed: 14500911]
20. Weierich C, Brero A, Stein S, et al. Chromosome Research. 2003; 11:485–502. [PubMed: 12971724]
21. Alcobia I, Quina AS, Neves H, et al. Experimental cell research. 2003; 290:358–369. [PubMed: 14567993]
22. Ding C, He XF. Advances in Knowledge Discovery and Data Mining, Proceedings. 2004; 3056:414–418.
23. Langmead B, Trapnell C, Pop M, et al. Genome biology. 2009; 10:R25. [PubMed: 19261174]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

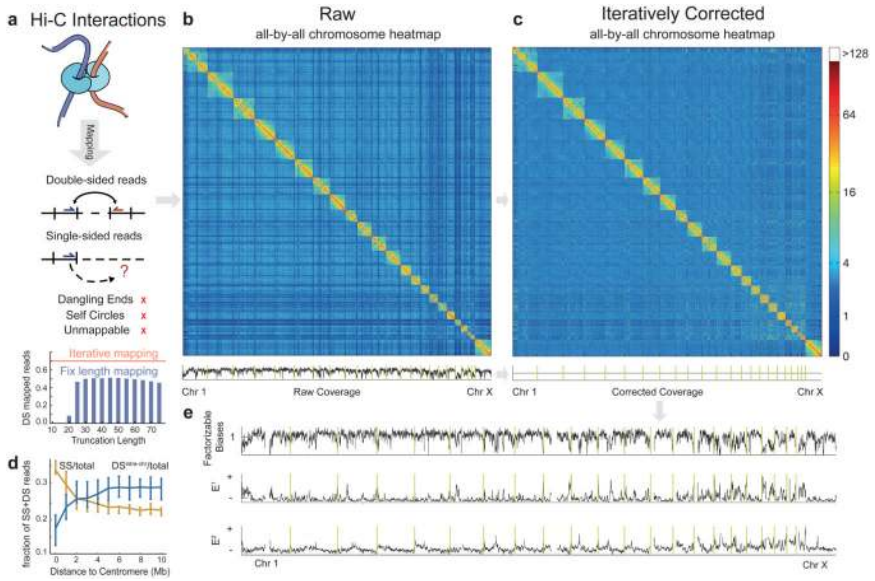

Figure 1. Pipeline for mapping, filtering, and iterative correction of Hi-C reads

(**a**) Interacting chromatin regions are sequenced and reads are mapped to the genome using iterative mapping. Only the depicted double-sided reads (DS), or single-sided reads (SS) are retained. Bars show the fraction of DS reads mapped by truncation to fixed length, red line shows result of iterative mapping.(**b, c**) Raw and iteratively corrected whole-genome Hi-C maps binned at 1Mb resolution (filtered-out megabases are not shown). Coverage profile is the sum of each column in the map. Vertical yellow lines show chromosome boundaries. Note that after iterative correction the coverage profile is uniform. (**d**) Fractions of SS and DS intra-chromosomal reads as a function of centromeric distance, plotted at 1 Mb resolution for distances up to 10 Mb from each centromere; lines represent mean values and vertical bars represent 25th and 75th percentiles](**e**) Factorizable biases and eigenvectors (E^1 and E^2) obtained by ICE (at 1Mb resolution). Regions that do not pass filters (see Online Methods) or contain no mapped reads are shown as gaps. Vertical yellow lines show boundaries of chromosomes.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

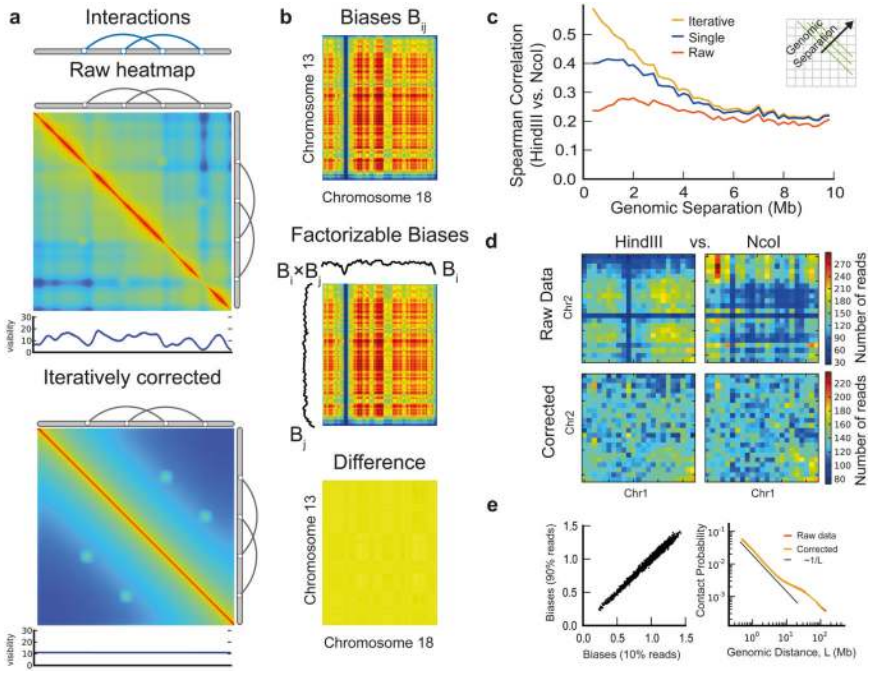

Figure 2. Iterative correction of Hi-C data

(**a**) Illustration of iterative correction using simulated data. (Top) two specific interactions (shown by arches) within a chromosome, (middle) its simulated Hi-C heatmap and a vector of random experimental visibility. Notice that visibility-induced noise obscures specific interactions. (bottom) Iteratively corrected map of the chromosome, where visibility is equalized, revealing two specific interactions as bright spots on the heatmap. (**b**) Matrix of biases computed by Yaffe and Tanay[17] at 1Mb resolution (top) can be approximated by a product of bias vectors $B_i \times B_j$ (middle), yielding an essentially identical matrix of biases (r =0.99), with their algebraic difference shown at the bottom in the same colorscheme (also Supplementary Fig. 4). (**c**) Comparison of intra-chromosomal Hi-C maps obtained using HindIII and NcoI enzymes (200kb resolution). The correlation is computed between off-diagonal regions of the map and plotted as a function of distance from the main diagonal, that is, the genomic separation, as shown in the inset. Analysis was performed on raw data (red), single corrected (blue) and iteratively corrected (yellow). (**d**) Inter-chromosomal heatmaps (chr1 vs. chr2, coarse-grained to 10MB, contact frequencies shown by color for HindIII and NcoI before (top row) and after correction (bottom raw) (also see Supplementary Fig. 5). (**e**) (left) Cross-validation for biases inferred from 10% vs. 90% of the reads. (right) Scaling of intra-chromosomal contact probability with genomic distance, L for Hi-C HindIII[8] data, at 200 kb resolution, before (red) and after correction (yellow). Black line shows 1/L scaling reported previously[8].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

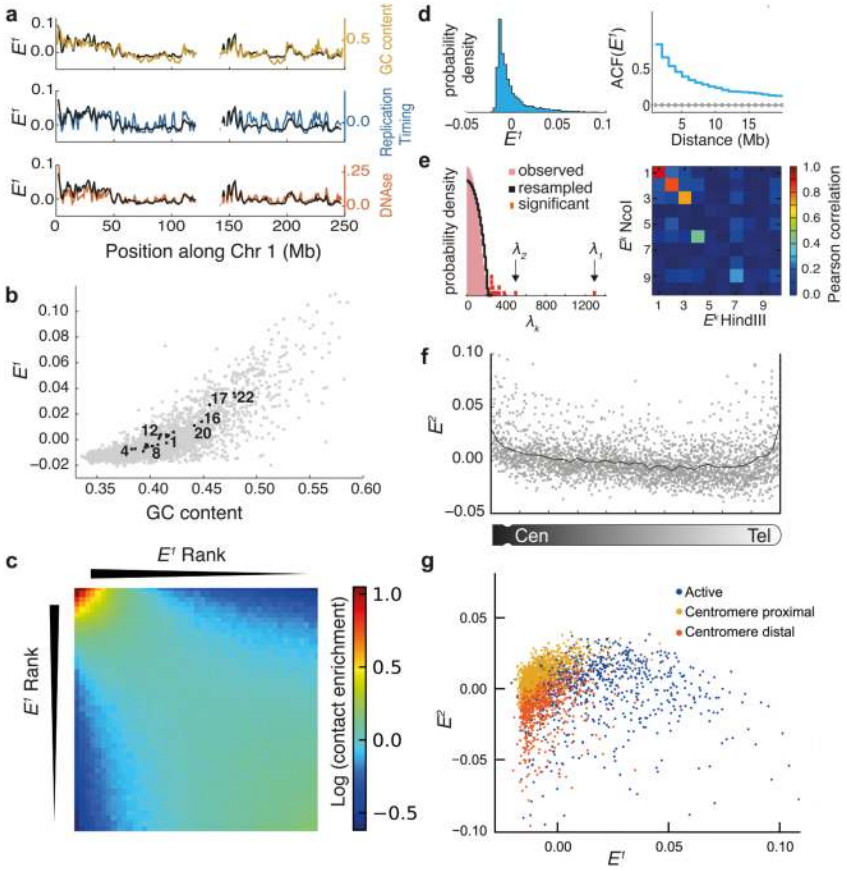

Figure 3. Eigenvector decomposition of iteratively corrected Hi-C data reveals genome-wide features of chromosome organization

(**a**) Profiles of E^1 and genomic features along chr1 (1Mb resolution), E^1 from Hi-C HindIII data[8](**b**) Scatter plot of E^1 vs GC content. Gray dots show GC content and E^1 of individual 1Mb regions. Black squares show mean chromosomal values of E^1 and mean GC content. Several chromosomes are indicated by numbers. (**c**) Heatmap of inter-chromosomal contacts between pairs of genomic regions as a function of their E^1 values; heatmap shows natural log of contact enrichment (see Online Methods). Notice the tendency of regions with similar values of E^1 to interact with each other. (**d**) (Left) Distribution of E^1 values. (Right) Autocorrelation of E^1 (blue) compared to 1000 shuffled E^1 (gray line shows mean, errorbars show standard deviation). (**e**) (Left) Distribution of observed eigenvalues (λ_k) and the distribution of eigenvalues for randomly re-sampled data (see Online Methods). Thirteen significant eigenvalues are shown in red. (Right) Matrix of Pearson correlation coefficients of leading eigenvectors obtained for NcoI and HindIII Hi-C data, revealing robustness of top three eigenvectors. (**f**) Variation of E^2 along chromosomal arms, with higher values near centromeres and telomeres. Grey points show values for individual genomic regions, black line shows the mean. (**g**) Genome-wide inter-chromosomal interactions mapped onto E^1 and E^2 space at 1Mb resolution. Regions are colored according to previously proposed[17] chromatin types[17]. Notice no [17]evident separation into distinct clusters. E^1 and E^2 calculated for Hi-C HindIII dataset[8].

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

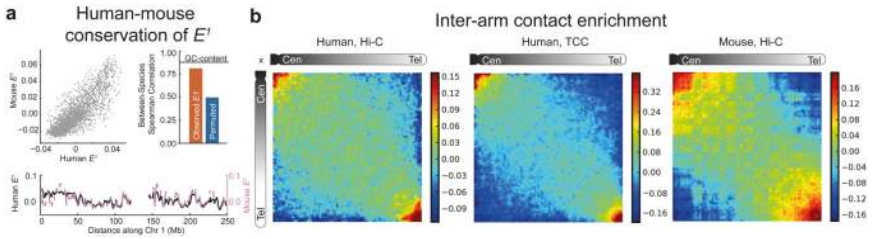

Figure 4. Cross-dataset and cross-species comparisons reveals evolutionary conserved genome-wide chromosome organization

(**a**) (top left) Scatter plot of E^1 for human vs. mouse in syntenic regions; (top right) comparison of observed between-species correlation of E^1 (r=.81, P < 1e-10) with GC-content stratified permuted data (r=.50, P < 1e-10); (bottom) human vs. syntenic mouse E^1 along human chr1; gaps in the mouse profile reflect regions of human chr1 without a corresponding syntenic region in mouse. Human E^1 is for TCC HindIII[10] data, mouse E^1 was calculated for mouse Hi-C[16] data. (**b**) Heatmaps of iteratively-corrected inter-chromosomal contact probability averaged over all chromosomal arm pairs; heatmaps show the natural log of the contact enrichment, re-scaled and re-binned to 80×80 map (see Online Methods). The data are for human lymphoblastoid Hi-C HindIII[8], human lymphoblastoid TCC[10], and mouse pro-B cell Hi-C[16. 16]