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

Maxim Imakaev¹,⁵, Geoffrey Fudenberg²,⁵, Rachel Patton McCord³, Natalia Naumova³, Anton Goloborodko¹, Bryan R Lajoie³, Job Dekker³ & Leonid A Mirny¹,²,⁴

Extracting biologically meaningful information from chromosomal interactions obtained with genome-wide chromosome conformation capture (3C) analyses requires the elimination of systematic biases. We present a computational pipeline that integrates a strategy to map sequencing reads with a data-driven method for iterative correction of biases, yielding genome-wide maps of relative contact probabilities. We validate this ICE (iterative correction and eigenvector decomposition) technique on published data obtained by the high-throughput 3C method Hi-C, and we 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.

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, and by biological factors, such as those resulting from intrinsic physical properties of distinct chromatin states. 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 methods⁴–¹¹ (such as Hi-C) for investigating physical contacts between distal genomic loci have begun to provide key insights into the spatial organization of genomes⁴–¹⁵. However, the raw outputs of 3C-based methods may be influenced by various forms of biases.

Here, we present ICE, a pipeline that includes the processing of paired sequence reads obtained from genome-wide 3C-based experiments⁷–⁹,¹³,¹⁵ and a method of iterative correction that eliminates biases and is based on the assumption that all loci should have equal visibility (Fig. 1). Iterative correction 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 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 data sets from a human lymphoblastoid cell line: two data sets generated by Hi-C⁷ using either HindIII (Hi-C HindIII) or Ncol (Hi-C Ncol) digestion and one generated by a Hi-C variant, tethered chromosome capture (TCC)⁹, using HindIII digestion. We also analyze one HindIII-digested mouse pro–B cell Hi-C data set¹⁵.

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 does (Fig. 1a). After alignment, the pipeline discards molecular byproducts (Supplementary Figs. 1 and 2 and Online Methods). The remaining read pairs include double-sided (DS) reads, which represent a contact between two mappable portions of the genome, and single-sided (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 pericentromeric regions, where decreased intrachromosomal 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 it 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

¹Department of Physics, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA. ²Graduate Program in Biophysics, Harvard University, Cambridge, Massachusetts, USA. ³Program in Systems Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. ⁴Institute for Medical Engineering and Science, MIT, Cambridge, Massachusetts, USA. ⁵These authors contributed equally to this work. Correspondence should be addressed to J.D. (Job.Dekker@umassmed.edu) or L.A.M. (leonid@MIT.edu).

RECEIVED 14 MAY; ACCEPTED 4 AUGUST; PUBLISHED ONLINE 2 SEPTEMBER 2012; DOI:10.1038/NMETH.2148
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 (DS) or single-sided (SS) reads are retained. Blue 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 (all-by-all) Hi-C maps binned at 1-Mb 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 interchromosomal 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 ($E_i$ and $E_j$) obtained by ICE (at 1-Mb 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.

To validate our method, we first compared 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 because of enzyme-dependent biases. Corrected maps show an increased between-enzyme correlation of corresponding off-diagonal interchromosomal elements (Fig. 2c). Iterative correction also increases between-enzyme correlation for interchromosomal maps to the level of correlation between halves of the same data set (Fig. 2d and Supplementary Fig. 6a). To compare to a previous method16, we applied the same smoothing technique and obtained a similar between-enzyme correlation $r = 0.71$ ($r = 0.59$ was obtained earlier16). Next, we performed cross-validations using 10% or 90% of the read pairs and obtain biases that are highly correlated ($r = 0.98$, $P < 10^{-10}$, HindIII), demonstrating that our method does not over-fit (Fig. 2e). We also note that an important property of intrachromosomal 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$^{7,10,16}$. Applying this procedure once only partially corrects for nonuniform coverage (Fig. 2c), tends to flip the coverage profile (Supplementary Fig. 6c) 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 data sets (Fig. 2c,d).

Eigenvector analysis 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 ($E_i^k$) of genomic region $i$. Independent interaction preference tracks $E_i^k$ can be found as eigenvectors of the corrected
map $T_j$ ($T_j = \sum \lambda_k E_k^j E_k^j + \text{constant}$), where the relative weights of their contributions $\lambda_k$ are the corresponding eigenvalues. The contribution of each track to the total interaction frequency between a pair of regions in the corrected map $T_j$ is proportional to a product of these preferences ($E_k^j E_k^j$). Eigenvectors are sorted ($E_1^j, E_2^j, E_3^j, \ldots$) 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.\textsuperscript{2} Permutation analysis shows that the first 13 eigenvectors are statistically significant ($P < 0.001$). Moreover, the first three are robust between human data sets (Fig. 3e, Supplementary Fig. 7 and Online Methods) and explain 72% of the interchromosomal data reconstructed from the first 13 eigenvectors. Thus, we focus on the first three eigenvectors for further analysis of interchromosomal interaction preferences.

The leading eigenvector, $E_1^j$, provides a genomic track of interchromosomal interaction preferences along the genome, and it shows correlation with many genomic features (Fig. 3a,b), including GC content ($r = 0.80, P < 10^{-15}$), replication timing ($r = 0.82, P < 10^{-15}$, GEO GSM500943), DNase I hypersensitivity ($r = 0.79, P < 10^{-15}$, GEO GSE4334) and many histone marks (Supplementary Table 1). The profile of $E_1^j$ is similar to chromatin compartments found previously,\textsuperscript{7} yet $E_1^j$ shows higher correlation with many genomic features\textsuperscript{17} both along the chromosomes and for average values of whole chromosomes (Fig. 3b; $r = 0.95, P = 4 \times 10^{-6}$, versus $r = -0.31$ for chromatin compartments reported previously, Supplementary Fig. 7).

Interaction preferences represented by $E_1^j$ connect spatial and functional genomic organization, as regions with high $E_1^j$, 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 lower $E_1^j$ tend to interact more with other gene-poor regions. Despite its tendency to partition active and inactive regions of the genome, $E_1^j$ does not show any bimodality (Fig. 3d, top). Neighboring genomic regions display similar interaction preferences, as seen from the autocorrelation (Fig. 3d, bottom), which decays with a characteristic length of about 6 Mb. Taken together, these characteristics of $E_1^j$ suggest that continuous interaction preferences better capture the complexity of chromatin interaction landscape at megabase resolution than does a two-compartment model\textsuperscript{7} proposed earlier.

Furthermore, we found evidence for the evolutionary conservation of genome-wide chromosome organization by comparing $E_1^j$ for human and mouse data sets. $E_1^j$ has high correlation ($r = 0.81, P < 10^{-10}$) in syntenic regions\textsuperscript{18} of human and mouse genomes at the megabase level (Fig. 4a). Moreover, the conservation of $E_1^j$ 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 studied the interaction preference tracks, $E_2^j$ and $E_3^j$, which constitute the greatest contributions to the corrected map after $E_1^j$. Both $E_2^j$ and $E_3^j$ vary with position along chromosomal arms (Figs. 1e and 3f), with increased magnitude mostly near centromeres for $E_2^j$ and near telomeres for $E_3^j$. This pattern of interaction is prominent on average inter-arm maps, which reveal an enrichment of centromere-centromere and telomere-telomere contacts (Fig. 4b and Supplementary Fig. 8). Average inter-arm maps constructed from projections of the data on $E_2^j$ and/or $E_3^j$, but not $E_1^j$, show a similar pattern of contact enrichment, directly confirming that arm-level organization is largely captured by $E_2^j$ and $E_3^j$ (Supplementary Fig. 9). This pattern is
consistent with colocalization of centromeres and a similar colocalization of telomeres, as described in imaging studies\textsuperscript{19,20}. We observed a consistent pattern of contact enrichment for all studied human and mouse data sets, despite the acrocentric structure of mouse chromosomes (Fig. 4b). For the mouse data set, both centromere-centromere and telomere-telomere enrichment are captured by $E^3$ (Supplementary Fig. 9); $E^2$ possibly 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\textsuperscript{19,20}.

Multiple attempts have been made to identify distinct chromatin types based on Hi-C data\textsuperscript{7,16}. We compared the $E^1$ and $E^2$ representation of interchromosomal interactions to a model of
three chromatin types identified earlier by \( k \)-means clustering\(^{16} \) (Fig. 3g). We found that the suggested clusters do not show evident separation and the suggested division into three chromatin types is ambiguous\(^{21} \) (Supplementary Fig. 10). We also note that \( E^1 \) captures variation in epigenomic tracks much better than the three chromatin types (Supplementary Fig. 11).

**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 maps of relative contact probabilities for individual chromosomes or genome-wide. Most importantly, it allows an unbiased comparison of Hi-C data within and between data sets, cell types and organisms.

We note that our data-driven method is specific to techniques that yield a pairwise and genome-wide matrix of contacts; although other 3C-based methods that do not yield all-by-all interaction maps (4C\(^{4,5} \), 5C\(^{6} \)) have similar systematic biases, they must be corrected using an alternate approach\(^{16} \). We also note that iterative correction operates on binned data, and thus it 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 Note).

Our analysis of interchromosomal Hi-C data suggests that at megabase resolution, three-dimensional 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 data sets. 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 interchromosomal organization.

**METHODS**

Methods and any associated references are available in the online version of the paper.

The ICE software is available at https://bitbucket.org/mirnaylab/hiclib.

Note: Supplementary information is available in the online version of the paper.

**ACKNOWLEDGMENTS**

The authors are grateful to K. Korolev for many productive discussions. The work of M.I., G.F., A.G. and L.M. were supported by the US National Cancer Institute Physical Sciences–Oncology Center at MIT (U54CA143874). This work was supported by US National Institutes of Health grants HG003143 (to J.D.) and F32GM100617 (to R.P.M.) and by a W.M. Keck Foundation Distinguished Young Scholar in Medical Research Award (to J.D.)

**AUTHOR CONTRIBUTIONS**

M.I. developed the iterative correction procedure. M.I. and G.F. developed data analysis tools. M.I. and A.G. developed and maintain publicly available software. M.I., G.F., R.P.M. and A.G. performed data analysis. M.I., G.F., R.P.M., N.N., A.G., B.R.L., J.D. and L.A.M. contributed to conceiving the study and wrote the paper.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2148. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Cheung, M.S., Down, T.A., Latorre, I. & Ahringer, J. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res.* 39, e103 (2011).
2. Quail, M.A. et al. A large genome center’s improvements to the Illumina sequencing system. *Nat. Methods* 5, 1005–1010 (2008).
3. Teytelman, L. et al. Impact of chromatin structures on DNA processing for genomic analyses. *PLoS ONE* 4, e6700 (2009).
4. Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* 38, 1346–1354 (2006).
5. Zhao, Z. et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* 38, 1341–1347 (2006).
6. Dostie, J. et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* 16, 1299–1309 (2006).
7. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293 (2009).
8. Duan, Z. et al. A three-dimensional model of the yeast genome. *Nature* 465, 363–367 (2010).
9. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, N. & Chen, L. Genome architecture revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* 30, 90–98 (2012).
10. Sexton, T. et al. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* 148, 458–472 (2012).
11. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* 295, 1106–1111 (2002).
12. van Steensel, B. & Dekker, J. Genomics tools for unraveling chromosome architecture. *Nat. Biotechnol.* 28, 1089–1095 (2010).
13. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromosome interactions. *Nature* 485, 376–380 (2012).
14. Nora, E.P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485, 381–385 (2012).
15. Zhang, Y. et al. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* 148, 908–921 (2012).
16. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* 43, 1059–1065 (2011).
17. ENCODE Project Consortium et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816 (2007).
18. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution’s caution: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* 100, 11484–11489 (2003).
19. Weiher, C. et al. Three-dimensional arrangements of centromeres and telomeres in nuclei of human and murine lymphocytes. *Chromosome Res.* 11, 485–502 (2003).
20. Alcobia, I., Quina, A.S., Neves, H., Clode, N. & Parreira, L. The spatial organization of centromeric heterochromatin during normal human lymphopoiesis: evidence for ontogenetically determined spatial patterns. *Exp. Cell Res.* 290, 358–369 (2003).
21. Ding, C. & He, X. K-means clustering via principal component analysis. *Proc. Intl. Conf. Machine Learning (ICML 2004)* 225–232 (2004).
The pipeline we present here contains three sequentially applied modules: (i) alignment of read pairs from Hi-C to the genome; (ii) filtering at the level of aligned read pairs and restriction fragments; and (iii) iterative correction for biases in visibility for filtered and binned data. Iterative correction is a method for obtaining relative contact probabilities. Code and documentation for our Hi-C data analysis library are publicly available at an online repository, https://bitbucket.org/mirnylab/hiclib, and are also accessible through the lab website, http://mirnylab.mit.edu/.

Read alignment. Mapping of Hi-C ligation products is complicated by the possibility that either side of a Hi-C read pair can cross a ligation junction. Because ligation junctions may appear anywhere within a read, we use a general, multistep, parameter-free approach to independently map the two sides of each paired-end read. We start by truncating each read to 25 bp (that is, keeping 25 bp at the 5’ end of the read) and mapping them to the genome. Reads that do not map uniquely are then extended to 30 bp and re-mapped. This process is repeated with 5-bp steps until the maximum read length (here, 75 bp) is achieved, with uniquely mapped reads being collected at every step (Supplementary Fig. 1). To align reads, we use Bowtie2 (ref. 22) with increased stringency (--score-min -L 0.6,0.2--very-sensitive) to yield the maximum number of DS reads; all other settings were kept at default values.

Aligned reads are then assigned to restriction fragments by their 5’ end position. All reads starting within 3 bp of the restriction site and crossing it are considered to start exactly at this restriction site and are assigned to the fragment they point toward.

Read-level and fragment-level filtering. After alignment, our pipeline discards read pairs with two unmapped sides and removes read pairs corresponding to repeated instances of the same DNA molecule (which may result from PCR amplification). We then analyze the position and direction of each mappable read from each read pair to separate molecular byproducts from informative DS reads.

If both reads from the read pair occur in the same fragment, they either represent a self-circularized ligation product (if reads face in opposite directions) or an unligated ‘dangling end’ product (if reads face toward each other; Supplementary Fig. 1). In contrast, read pairs that represent informative chromatin contacts map to different restriction fragments. We additionally remove read pairs from neighboring fragments that face toward each other and are separated by less than the maximum length of molecules in the analyzed Hi-C library (here, 500 bp), as these can be contaminated by spuriously pulled down short DNA molecules.

A subset of mapped read pairs represent ligations of fragments that did not arise from canonical restriction digestion but are the product of star (noncanonical) activity of the enzyme or physical breakage of chromatin (Supplementary Fig. 2); this subset is further called ‘random breaks’. Furthermore, true enzymatic breaks may appear as random breaks because of differences between a studied genome and the reference genome or because of poor assembly of the reference genome (Supplementary Fig. 2). We identify random breaks as reads that are further from their restriction site than the maximum molecule length in the library.

The maximum library length may be determined by gel analysis or from size distribution of dangling ends (Supplementary Fig. 12). In principle, this criterion is necessary but not sufficient to label random break molecules, but allows consistent analysis of SS and DS random breaks. However, if the decision to remove all random breaks from the library is made, a more stringent criterion should be applied (see ref. 16).

To analyze the properties of random break reads, we compared heatmaps, decay of the contact probability, and cis-to-trans ratios between read pairs from random breaks and enzymatic cut sites (Supplementary Figs. 2 and 13). For HindIII Hi-C, heatmaps and contact probability decays are similar for these two classes of reads. However, we note that random breaks tend to form more interchromosomal contacts, suggestive of a higher number of intermolecular ligations. This may be due to the existence of relatively fewer compatible ligation partners for molecules with noncanonical ends in the same cross-linked protein-DNA complex, which biases these molecules toward a greater number of intermolecular ligations. Noise levels of read pairs coming from enzymatic cuts and random breaks may both vary greatly for different experimental conditions, as seen by different cis-to-trans ratios of enzymatic and random breaks (Supplementary Fig. 2).

For HindIII, both cis-to-trans ratios are higher, indicating that data obtained with NcoI is more strongly affected by noise from intermolecular ligations in solution.

We further filter the data at the level of individual restriction fragments. We first remove from consideration all fragments shorter than 100 bp, because they are comparable to the read length. We also remove fragments longer than 100 kb ($P < 10^{-5}$ to occur in the randomized genome once) as they likely represent poorly assembled or repeated regions of the genome. We then remove the top 0.5% of fragments with the greatest number of reads, as they may be prone to PCR artifacts or represent fragile regions of the genome or genome assembly errors (Supplementary Fig. 14). This cutoff was determined by considering the ratio of cis to total reads per fragment; we find that this ratio has a non-monotonic behavior as the function of number of reads per fragment and starts rapidly decreasing after our cutoff value of 0.5%.

This exact cutoff was specific to the Hi-C data set, and both TCC and mouse data sets have cutoffs closer to 0.1% or 0.2%. We also observe significant differences between heatmaps reconstructed from the 0.2% highest-count fragments and the second-highest 0.2% of fragments. Removing these high-count fragments suppresses blowouts in the interchromosomal contact map and yields consistent eigenvectors.

Iterative correction and filtering of binned data. We bin filtered reads at either 1-Mb or 200-kb resolution, assigning DS reads to a genome-versus-genome heatmap and SS reads to a genome-wide vector. Reads are assigned to bins by the centers of their fragments, as most reads reflect contacts occurring anywhere within a fragment. To study relative contact probabilities of long-range interactions, we remove from consideration contacts within a bin and with adjacent bins, as these contacts occur at scales finer than the given resolution. We then remove the 2% of bins having the fewest number of contacts, to avoid blowouts of these bins during iterative correction. This cutoff is determined by analyzing the variance of the bins as a function of coverage (Supplementary Fig. 15) and may be set less stringently for future data sets with...
higher coverage. For between-data set comparisons, we remove from consideration any bin that does not pass the filters or does not have any counts in either data set.

We perform iterative correction on the resulting contact maps to obtain biases \( B_i \) and ‘true’ \( T_{ij} \) relative contact probabilities by explicitly solving the system of equations:

\[
O_{ij} = B_i B_j T_{ij} \sum_{i=1}^{N} B_i = 1
\]

The summation is over elements of column \( i \), \( O_{ij} \) is the map of DS reads and \( N \) is the number of bins. Because an experiment represents a sample from a distribution of possible interactions, the observed interaction frequency is a realization \( O_{ij} \sim f(\mathcal{E}_{ij}) \) from some distribution \( f \) (for example, a Poisson distribution).

For a range of distributions, the maximum-likelihood solution for \( B_i \) under these constraints is the solution of the above equation, which can be found by the simple iterative procedure described below.

After the vector of biases is computed, the corrected map \( \tilde{T}_{ij} \) of relative contact probabilities is obtained by

\[
\tilde{T}_{ij} = O_{ij}/(B_i B_j)
\]

(Anonymous Fig. 16 and Anonymous Note).

Algorithmically, the iterative correction is implemented as follows. We start by creating a working copy of the matrix \( O_{ij} \) denoted \( W_{ij} \) as the iterative process gradually changes this matrix to \( T_{ij} \). We initialize the iterative procedure by setting each element of the vector of total biases \( B_i \) to 1. We begin each iteration by calculating the coverage \( S_i = \sum_j W_{ij} \). Next, additional biases \( \Delta B_j \) are calculated by renormalizing \( S_i \) to have the unit mean \( \Delta B_j = S_j/\text{mean}(S_j) \). We then divide \( W_{ij} \) by \( B_j \Delta B_j \) for all \((i,j)\) and update the total vector of biases by multiplying by the additional biases. Iterations are repeated until the variation of the additional biases becomes negligible; at this point \( W_{ij} \) has converged to \( T_{ij} \).

For many maps, iterative correction converges in around 10 iterations (Anonymous Fig. 15). For certain cases, however, further iterations are required to pull in the outliers, particularly bins with poor visibility. Throughout the paper, we use 50 iterations for iterative correction; this is twice the number required for the worst case encountered. The initial uniform bias vector is chosen such that after the first iteration, the estimated bias vector is in fact the maximum-likelihood estimate of factorizable biases without the constraint of uniform visibility. We note that iterative correction can be extended to incorporate SS reads (Anonymous Fig. 4). SS reads are used to supplement DS reads in calculating the coverage profile.

Calculating scaling of the contact probability. To plot contact probability as a function of genomic separation (Fig. 2e), we first divide all possible separations (400 kb–150 Mb) into logarithmically sized bins, which grow in size by a factor of 1.2 and are rounded to the nearest integer. For each logarithmic bin, we then calculate the mean value of Hi-C contact map in this range of separations, excluding bins that were filtered out or had no counts. For this calculation, chromosomal arms are treated separately, and interchromosomal data were not considered.

Obtaining a corrected interchromosomal contact map. To obtain a corrected binned interchromosomal contact map, we start with a filtered genome-wide contact map. We then substitute intrachromosomal contacts with pseudo-contacts, which simplifies further analysis and permits the use of multiple statistical tools. Pseudo-contacts for every intrachromosomal position \((i,j)\) of the map are chosen randomly and uniformly from interchromosomal contacts of the corresponding row and column, and are added symmetrically. The resulting genome-wide map is then iteratively corrected. Because original pseudo-counts are affected by biases, we redraw pseudo-counts from the iteratively corrected map and repeat iterative correction. This procedure is repeated three times and ensures consistent and unbiased pseudo-counts. Pseudo-contacts obtained this way are visually indistinguishable from the rest of the map and introduce negligible noise to the computed eigenvectors (Spearman \( r = 0.998 \) between two realizations). This method for generating pseudo-contacts preserves the structure of the sampling noise and thus avoids the introduction of spurious correlations between bins on the same chromosome. A similar pseudo-count approach can be used to mask visible translocations in rearranged genomes. For the mouse Hi-C data set, chromosome 13 has three clear translocations; correlations with genomic features and average inter-arm maps were calculated after masking these translocations. Omitting mouse chromosome 13 from the analysis led to similar results.

Eigenvector analysis of interchromosomal contact map. Eigenvector analysis of a corrected interchromosomal contact map \( T \) involves expanding the matrix as a sum of outer products between eigenvectors, \( E_i E_i \), weighted by their eigenvalues:

\[
T_{ij} = \sum_k \lambda_k E_i^k E_j^k + \langle T \rangle
\]

where \(<T>\) denotes the mean value of the matrix, and the magnitude of the eigenvalue \( \lambda_k \) describes the amount of information captured by the corresponding eigenvector \( E_i \), where \( k \) runs from 1 to \( N \). Eigenvectors are then sorted by the absolute value of their eigenvalues, and eigenvectors corresponding to the three largest eigenvalues, \( \lambda_1, \lambda_2 \) and \( \lambda_3 \), are used for further analysis (Fig. 3). Iterative correction is a key prerequisite for eigenvector expansion; performing eigenvector expansion (or principal-component analysis, PCA) on the raw data entangles biases and eigenvectors, making the result nontransparent and bias dependent. Moreover, \( \lambda_3 \) is clearly interpretable as the solution to a linear model of chromatin interaction preferences.

We note that eigenvector expansion of iteratively corrected interchromosomal data is mathematically equivalent to PCA of iteratively corrected data, but not to PCA of uncorrected data or PCA of the correlation matrix. The best way to illustrate this connection is to follow through the steps of PCA performed on a symmetric matrix \( A \) with unit marginals, where the sum over any row or column of \( A \) equals 1. The covariance of matrix \( A \) is simply a dot product of \((A - \text{mean}(A))\) with itself. As taking the dot product of a matrix with itself does not change its eigenvectors, eigenvectors obtained from PCA of matrix \( A \) will be equivalent to the eigenvectors of \( A \) with squared eigenvalues. If the matrix was not iteratively corrected, this relationship does not hold.

To test the statistical significance of eigenvectors, we re-sample the interchromosomal contact map, perform iterative correction and compute the resulting eigenspectra; this procedure is repeated for 1,000 independent realizations. Re-sampled data for contact matrix element \((i,j)\) is generated by drawing from the distribution
of interchromosomal contacts from row \( i \) and column \( j \); this is the same procedure described for generating intrachromosomal pseudo-contacts.

**Obtaining an average inter-arm interchromosomal contact map.** To obtain an average interchromosomal inter-arm interaction map (Fig. 4a), we first calculate the genome-wide set of bins with nonzero number of interchromosomal contacts. The intersection of these bins on the map denotes the portion of the interchromosomal map where a contact between two bins can be observed, creating an expected genome-wide map of 1s (observable) and 0s (unobservable). We then extract all autosomal observed and expected inter-arm interchromosomal maps, and keep only those that have more than 20 nonzero rows and columns. Individual maps are then rescaled to an 80 × 80 square using linear interpolation, keeping the sum constant during the rescaling procedure. Finally, maps are summed together and the enrichment map is obtained by dividing observed by expected.

**Evolutionary conservation of chromatin organization.** To calculate between-species correlations of \( E^1 \), we make use of syntenic genome-versus-genome chained alignments (http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/vsMm9/syntenicNet)\(^{18} \). We then bin these alignments to 1 Mb, and create a one-to-one map of human to mouse bins. For each human bin we find the mouse bin with the greatest number of syntenic base pairs; we do likewise for each mouse bin. We then keep the intersection of these two sets. We then calculate between-eigenvector correlations using this subset of bins. Despite differences in cell types between human and mouse data sets, we observed a high interspecies correlation of \( E^1 \). Future studies with matched cell types may reveal even higher between-species correlations. To control for the possible confounding effect of GC content in syntenic regions, we perform a permutation test stratified by GC content. For each species, we sort eigenvectors by GC content and then exchange nearest neighbors. This preserves the exact correlation structure between \( E^1 \) and GC content for each species but otherwise randomizes the eigenvector. For this analysis, mouse eigenvectors were calculated after omitting the rearranged chromosome 13.

22. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).