doi:10.1038/nature12593

Preparation of $CD4^+$ T_H1 cells

A single cell suspension was prepared from spleens of male C57BL/6 mice between 8-10 weeks old in RPMI1640 medium with L-glutamine (Life Technologies) supplemented with 5% fetal calf serum (FCS) (PAA), Penicillin-Streptomycin (Life Technologies), 5 x 10^{-5} M 2-mercaptoethanol and 1 µg/ml anti-CD3 (eBioscience). After the cells were expanded in culture at 37˚C for 3 days, $CD4^+$ T-cells were purified using magnetic beads coated with anti-CD4 (Miltenyi Biotec). The purified cells were further cultured at 0.5 - 1.0 x 10^6 cells/ml in RPMI1640 medium with L-glutamine, 5% FCS, Penicillin-Streptomycin, 2-mercaptoethanol and 20 ng/ml IL-2 (Sigma) for about a week. The purity of the culture was assessed by fluorescence activated cell analysis after staining with FITC-labelled anti-CD4 (Miltenyi Biotec) and 4,6-di-amidino-2-phenyl-indole (DAPI).

Preparation of Hi-C library from single cells

Five to ten million T_H1 cells were fixed with 2% formaldehyde for 5 min at room temperature before quenching with 127 mM glycine. The cells were washed with PBS and the nuclei were prepared in 10 mM Tris-HCl (pH 8), 10 mM NaCl, 0.2% IGEPAL CA-630 with protease inhibitor cocktail (Roche) for 30 min on ice with intermittent agitation. The nuclei were washed with 1.2 x NEBuffer 3 (New England Biolabs; 60 mM Tris-HCl (pH 7.9), 120 mM NaCl, 12 mM $MgCl_2$, 1.2 mM DTT) and suspended in 400 µl of 1.2 x NEBuffer 3. Six µl of 20% SDS was added and incubated at 37˚C for 60 min with constant agitation, then 40 µl of 20% Triton X-100 was added and incubated at 37˚C for 60 min with constant agitation. Next, restriction enzyme 1 (RE1), 30 µl of 50 U/µl Bgl II (Dpn II for some libraries; New England Biolabs), was added and incubated at 37˚C for overnight with constant agitation. To label the digested DNA ends, 1.5 µl of 10 mM dCTP, 1.5 µl of 10 mM dGTP, 1.5 µl of 10 mM dTTP, 37.5 µl of 0.4 mM biotin-14-dATP (Life Technologies) and 10 µl of 5 U/µl DNA polymerase I, large (Klenow) fragment (New England Biolabs) were added and incubated at 37˚C for 45 min with occasional mixing. To ligate the biotin-labelled DNA ends, 750 µl of 10 x T4 DNA ligase reaction buffer (New England Biolabs), 75 µl of 100 x BSA (New England Biolabs), water and 50 µl of 1 U/µl T4 DNA ligase (Life Technologies) were added to the nuclei suspension to make the total volume 7.5 ml, and incubated at 16˚C for overnight. Then the nuclei were washed with PBS, single nuclei were picked up under the microscope using mouth-controlled glass capillaries to transfer to individual tubes containing 50 µl of PBS, and crosslinks were reversed by incubating at 65˚C overnight. To bind the biotin-labelled DNA from each cell to Dynabeads M-280 streptavidin (Life Technologies) magnetic beads, the washed beads were suspended in 50 µl of 2 x BW buffer (10 mM Tris-HCl [pH 8], 1 mM EDTA, 2 M NaCl), added to the single nucleus sample and incubated at room temperature for 60 min with constant mixing. After washing the beads with 1 x BW buffer (5 mM Tris-HCl [pH 8], 0.5 mM EDTA, 1 M NaCl) and 10 mM Tris-HCl (pH 8), the beads were incubated at 37˚C with RE2 (10 U of Alu I (New England Biolabs)), in 1 x NEBuffer 4 for 60 min. Then the beads were sequentially treated with i) 5 U of Klenow Fragment (3'-5' exo-) (New

England Biolabs) in 1 x NEBuffer 2 and 0.2 mM dATP, 37˚C for 30 min, and ii) 800 cohesive end units of T4 DNA ligase (New England Biolabs) in 1 x T4 DNA ligase buffer and 0.6 µM of customised Illumina adapter with 3 bp of identification tag (see below for sequences), room temperature for 60 min, each with constant mixing and the same washes before and after each reaction. Then the beads were re-suspended in amplification mixture containing 1 x Pfx amplification buffer (Life Technologies), 2 mM $MgSO_4$, 0.4 mM each dNTPs, 1 µM each library amplification primers (see below for sequences) and 2 U of Platinum Pfx DNA polymerase (Life Technologies), and amplified with the following programme (94˚C for 2 min, 25 cycles of [94˚C for 15 sec, 62˚C for 30 sec, 72˚C for 1 min], 72˚C for 10 min). After amplification, streptavidin-coated beads were removed, and the single cell Hi-C library was purified with Agencourt AMPure XP magnetic beads (Beckman Coulter), eluted in 15 µl of 10 mM Tris-HCl (pH 8.5), loaded on a 2% agarose gel, DNA between 300 and 700 bp was retrieved from the gel and eluted in 10 - 20 µl of 10 mM Tris-HCl (pH 8.5). The quality of the library was assessed with Agilent 2100 Bioanalyzer (Agilent), and the library was subjected to 2 x 40 bp paired-end high-throughput sequencing by Genome Analyzer IIx or HiSeq 1000 (Illumina).

Analyses of Hi-C ligation junctions

The efficiency of biotin labelling at the Hi-C ligation junctions in the single cell Hi-C protocol was analysed for two Hi-C ligation products. DNA was extracted from ~ 5 x 10^6 cells after ligation step with or without ligase. Hi-C products (1 and 2) were amplified by PCR from the DNA using primers 1F and 1R for the product 1 and 2F and 2R for the product 2 (see below for sequences), and they were visualised on a gel after the same amount was digested with either Not I, Bgl II or Cla I. Properly labelled Hi-C ligation junctions can be digested with Cla I but not with Bgl II, while the ligation junctions without proper labelling can be digested with Bgl II but not with Cla I. Not I doesn't digest the PCR products. This feature allowed us to analyse the Hi-C ligation junctions.

Primer sequence for 1F, 5'-CCCTTGTCTTTCCTATGTCTCACCTG-3';

1R, 5'-GATGAGGGCTGAAGGAGAATTAAAG-3';

2F, 5'-TGTTGTTCTGTTTCCTCGAAAGAC-3';

2R, 5'-GTGATTTTTACGCCTGGAACGTA-3'.

Sequences of customised Illumina adapters

The oligonucleotides (F) and (R) should be annealed to prepare the adapter for each identification tag. p indicates 5' phosphate modification, and * indicates 5'-3' phosphorothioate linkage.

adapter with [CAA] tag;

(F) 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAA*T-3'

(R) 5'-pTTGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

adapter with [TAA] tag;

(F) 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAA*T-3'

(R) 5'-pTTAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

adapter with [TCA] tag;

(F) 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCA*T-3'

(R) 5'-pTGAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

adapter with [ACC] tag;

(F) 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTACC*T-3'

(R) 5'-pGGTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

adapter with [CCT] tag;

(F) 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCT*T-3'

(R) 5'-pAGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

adapter with [GTA] tag;

(F) 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTA*T-3'

(R) 5'-pTACAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

adapter with [CAG] tag;

(F) 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAG*T-3'

(R) 5'-pCTGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

adapter with [TCG] tag;

(F) 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCG*T-3'

(R) 5'-pCGAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

adapter with [ATA] tag;

(F) 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTATA*T-3'

(R) 5'-pTATAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

adapter with [TGC] tag;

(F) 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGC*T-3'

(R) 5'-pGCAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

adapter with [CTA] tag;

(F) 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTA*T-3'

(R) 5'-pTAGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

adapter with [GAG] tag;

(F) 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAG*T-3'

(R) 5'-pCTCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'

Sequences of library amplification primers

* indicates 5'-3' phosphorothioate linkage.

5'-
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGAT
C*T-3'

5'-
CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTC
CGATC*T-3'

DNA FISH

Five BAC clone DNAs, RP24-354M10, RP23-360F19, RP23-82A12, RP23-162J10 and RP23-92P6, were used as P1-P5, respectively. BACs were labelled with aminoallyl-dUTP by nick translation. After purification, 0.5 - 1 µg of labelled BAC DNA was coupled with one third vial of either Alexa Fluor 488 and/or Alexa Fluor 555 Reactive Dyes (Life Technologies) according to manufacturer's instructions. Twenty five ng of each dye-coupled BAC DNA probe were precipitated with 8 µg of mouse Cot-1 DNA and 10 µg of salmon testis DNA for each hybridisation. For chromosome paint DNA FISH, each of the P1-P5 probe precipitates was dissolved in 8 µl of Mouse Chromosome X Whole Painting Probe D-1420-050-FI (MetaSystems) and used for hybridisation. For double-label DNA FISH, precipitated probes were dissolved in 10 µl of 50% formamide, 10% sodium dextran sulfate, 1 x SSC and used for hybridisation.

Approximately 5 x 10^5 CD4+ T_H1 live cells were suspended in 50 µl of ice-cold PBS and spotted onto poly-L-lysine slides for 4 min to let the cells settle. The slides were immersed in 4% formaldehyde in PBS for 10 min to fix. The fixation was quenched with 0.1 M Tris HCl, pH7.4 for 10 min, and the cells were permeabilised with 0.1% saponin, 0.1% triton X-100 in PBS for 10 min. After washing, the slides were immersed in 20% glycerol in PBS for 20 min and subjected to three cycles of freezing/thawing in liquid nitrogen, followed by washing in PBS. Then the slides were incubated in 0.1N HCl for 30 min, washed in PBS, permeabilised again in 0.5% saponin, 0.5% triton X-100 in PBS for 30 min, washed again, and immersed in 50% formamide in 2 x SSC for 10 min. For hybridisation, the probe mixtures (prepared as above) were applied onto the cells, covered with coverslips, heated at 78˚C for 2 min and incubated at 37˚C overnight. The next day the slides were washed in 50% formamide in 2 x SSC for 15 min at 45˚C, followed by 0.2 x SSC at 63˚C for 15 min, 2 x SSC at 45˚C for 5 min, and 2 x SSC at room temperature for 5 min. The cells were finally counterstained with DAPI and coverslips were mounted with Vectashield mounting medium (Vector Laboratories).

Images of the DNA FISH signals were captured and analysed with LSM510 META confocal microscope (Carl Zeiss), LSM Image Browser (Carl Zeiss), IMARIS software (Bitplane) and MetaCyte automated imaging system (MetaSystems).

Overview of sequence analysis

Sequenced single cell Hi-C libraries are similar to standard Hi-C (ensemble Hi-C) libraries in their general structure, but different in their statistical properties and critical noise factors. For example, single cell maps are sparse, and accurate probabilistic modeling as we routinely perform with ensemble Hi-C is impractical when thousands instead of millions of data points are available. On the other hand, the single cell libraries are sequenced to near-saturation, making it possible to identify and filter multiple noise factors that cannot be approached when analyzing ensemble Hi-C data. We have therefore implemented a specialized protocol to analyze raw single cell Hi-C sequences and generate reliably filtered contact maps. We next provide a step-by-step description of the analysis procedure, following with more details on the analysis of control and calibration experiments supporting some of the key aspects of our noise filtering approach.

1. De-multiplexing and mapping

Paired-end reads of each multiplexed sequencing lane were paired-up and separated to single cell datasets based on the 3 bp unique identification tags prefixing each read. Reads lacking a perfect match index were discarded. Read-pairs of each single cell were then processed independently. We mapped reads to the *Mus musculus* genome (assembly mm9) using the Maq program with standard parameters. Mapping was performed separately for the two ends of each paired-read. We retained for further analysis only pairs of reads with unique hits of high quality ($q \geq 30$) at both ends for further analysis.

2. Fragment association, pair classification

Given the coordinate and orientation of each mapped read, the first RE1 (Bgl II or Dpn II) restriction site downstream of the read was identified and used as the fragment-end (fend) associated with the read. By the design of the experiment, the vast majority of the mapped reads (94.9-99.7%, median 99%) mapped exactly at a RE2 (Alu I) restriction site. We retained read-pairs in which both ends mapped to the first RE2 site upstream of RE1 site, which we named normal ligation read-pairs (**Extended Data Fig. 1b**). Read-pairs with over 10 bp of total distance from the first RE2 sites in their respective fragments (classified as missed RE2) were discarded. Pairs of reads mapped to the same RE1 fragment facing each other (no ligation), not facing each other (self ligation), or pairs facing the same RE1 site (re-ligation) were also discarded. This resulted in a list of normal-ligated read-pairs associated with distinct pairs of RE1 fragment ends, which are denoted here as fend-pairs.

3. Read count filtering

Following mapping and filtering of imperfect ligation products, we identified the set of distinct fend-pairs and the coverage (number of read-pairs) associated with each of these putative contacts. Unlike ensemble Hi-C libraries, in single cell Hi-C the chance of forming a ligation between two fragment ends more than once is extremely small, and forming more than two ligation products is impossible for uniquely mapped sequences, since only two copies of each autosomal fragment are present in the nucleus. Nevertheless, following amplification of the single cell products and extensive sequencing, we observed a broad

distribution of fend-pair coverage (**Extended Data Fig. 1k**). These variable levels of coverage are disregarded by subsequent analysis, freeing single cell Hi-C analysis from the significant levels of systematic amplification bias as in ensemble Hi-C analysis. Nevertheless, in order to make sure that there is no remaining amplification bias that may result in non uniform coverage, we tested the contact coverage across the genome and found it is highly uniform (**Extended Data Fig. 1l**). Following a series of control experiments (described below), we estimated that a significant fraction of the fend-pairs covered (i.e., sequenced) only once (constituting a minor fraction of the total number of reads, but a significant fraction of the unique fend-pairs) represent likely pairing errors of the sequencing machine. We therefore further filter all fend-pairs that were covered only once.

4. Low mappability filtering and simplified fend-pair count normalization

Following the steps described above, single cell Hi-C maps included distinct valid ligation products. As in any Hi-C map, the density of contacts in these maps is limited by the density of RE1 fragments with sufficiently unique termini. In the single cell protocol the density of unique termini beginning at the RE2 sites (Alu I) adjacent to RE1 sites are of importance. We have identified all fends for which the upstream Alu I site starts a non-unique sequence of 36 bp, and marked these as non-unique. All fend-pairs associated with one or more of such reads were not considered when computing the potential number of fend-pairs associating two chromosomal regions of interest. Furthermore, we searched for regions with significantly high fend-pair coverage and excluded a sub-telomeric region in chromosome 11 (3,000,000-3,100,000) which according to the Hi-C data is involved with frequent contacts with all mouse chromosomes, probably due to incorrect assembly. Finally, RE1 fragments that completely lack an Alu I site (blind fragments) were identified and marked as inaccessible by the present protocol. Such fragments are relatively rare when using Bgl II as RE1, but become more abundant when using a more frequent cutter for RE1 (**Extended Data Fig. 1g**). Given these definitions, we used simple computation of the number of valid fend-pairs to normalize the number of observed contacts in a given two-dimensional bin of the Hi-C matrix. Importantly, although this approach did not directly correct for some of the biases we identified before in Hi-C[1], our studies suggest that the entire single cell Hi-C procedure, with saturated sequencing and complete control over amplification biases, generate results of comparable or even better quality than those derived using the ensemble approach.

5. Validating the single cell origin of the maps

Validating that a sample consists of a single cell that is not in midst of replication is essential. A non-replicating male mouse cell has single copies of the sex chromosomes (X and Y) and 19 dual copies of the autosomes. Theoretically, an autosomal fend can participate in at most two ligations, and a sex chromosome fend in only one. In contrast, when more than one cell is being assayed, or when a cell is assayed during S-phase, we can expect more than two ligation products to be associated with a single fend. In practice, mapping noise and sequencing errors may result in a small number of fends that are covered more than two times, and we wished to validate statistically that maps we retained for further analysis contain, if at all, only a small number of such errors. To this end, we computed the number of autosomal fends with more than two hits for each map, and compared it to a binomial distribution that represent the

expected number of fends covered by k fend-pairs given sequencing of a given number of reads:

$$\Pr(coverage = k) = \binom{N}{k}\left(\frac{1}{m}\right)^{k}\left(1-\frac{1}{m}\right)^{N-k}$$

where N is the sequencing depth and m the total number of potential fends (**Extended Data Fig. 1d**). In practice this model is conservative, since it assumes we sample fends uniformly, while in fact various factors increase or decrease the probability of covering an individual fragment end pair, but nevertheless we used it to discard samples for which the ratio between the binomial expectation and observed number of triply covered autosomal fends was high. Importantly, we found that discarding read-pairs that were covered once (as described above) dramatically reduced (~ 40-100%) the number of autosomal fends with more than two hits. We reason that this effect is due to pairing errors during sequencing as we discuss below, and that such pairing errors are *a priori* affecting only fends that are part of valid clusters on the flow cell, thereby creating a bias favoring triple coverage in genuine single cell sources. It is therefore critically important to remove such spurious pairs through filtering of fend-pairs that are represented by a single read.

6. Quality controls and selection of high quality matrices

We defined the following three conditions to select the ten highest quality single cell Hi-C matrices for in-depth analysis:

A) At least 10,000 unique fend-pair contacts.

B) The percentage of read-pairs that were sequenced only once (and therefore filtered) is less than 30%.

C) The number of autosomal fends with more than two hits should be less than the expected number according to the binomial model as described above.

Interestingly, all the samples that pass these criteria show a favorable ratio between contacts within elements of the same chromosome (*cis*) and *trans*-chromosomal contacts, an indication for a successful ensemble Hi-C experiment. We suggest that similar quality controls can be used routinely, although the thresholds employed can be adjusted depending on the goals of the analysis (i.e. if the goal is to maximize the number of cells while tolerating increasing level of noise). We indicate the complete statistics of the core set of single cell maps in **Extended Data Fig. 1h-j**.

Assessing Hi-C noise factors and their effect on single cell contact maps

In order to ensure single cell Hi-C maps are robust and as noise free as possible, we identified several potential sources of experimental errors in general HI-C experiments, and explored their effects in particular on our single cell Hi-C protocol. As we describe below, our control

experiments suggest that following the filtering procedure described above, our high-quality single cell contact maps contain a small, manageable level of noise, and that the protocol adaptation that were required to facilitate single cell manipulation and profiling, and in particular the in-nucleus ligation phase, did not affect negatively the specificity of the maps.

1. Intra-cellular spurious ligations

Intra-cellular spurious ligations might occur during one of the ligation phases - the in-nucleus fill-in and ligation or the Illumina adapter ligation - when ligations are formed between fragments that are not part of the same or two proximal chromatin aggregates. For instance, the second digestion with Alu I produces free floating DNA fragments which are later washed away, but some might randomly ligate to the Alu I digested fragment bound to a streptavidin-coated bead. Looking for a systematic way to seclude and exclude these spurious pairs we used the intra-chromosomal (*cis*) to *trans*-chromosomal ratio of different categories of fend-pairs as a measure for their quality. Since the probability of ligation in *cis* is much higher than ligation in *trans*, sets of fend-pairs with a high fraction of *trans* pairs are assumed to contain more spurious ligations. We found the "missed RE2" fend-pairs to have a high fraction of *trans* pairs (14-92%, median 71%) compared to the "normal ligation" fend-pairs (with 8-50% *trans*, median 21%). Filtering out those fend-pairs removed much of the noise in the experiment, which is reflected as uniform *trans* contacts (**Extended Data Fig. 1c**). We hypothesise that at a certain level, spurious ligation products are still present in the normal-ligation product pool, as they are in any Hi-C dataset, motivating the need for robust statistical analyses in modeling the contact matrix even when analyzing single cell matrices.

2. Inter-cellular spurious ligations

Inter-cellular spurious ligations might be formed when chromatin aggregates from foreign nuclei contaminate a cell's own aggregates. To test if such spurious ligations do occur and to what extent, we applied the single cell Hi-C strategy to a mixed population of human and mouse cells. By mixing the population at several different stages of the experiment and quantifying the levels of heterogeneous human-mouse ligation products, we were able to test the level of inter-cellular contamination introduced into the experiment (**Extended Data Table 2**). Overall the frequencies of spurious pairs (either human-human pairs in mouse libraries, mouse-mouse in human libraries, or human-mouse in either) were very low, ranging within 0.007 to 0.103%. We estimate these values mostly represent mapping problems, since in control experiments, mapping of pure mouse material to a combined human-mouse genome generated comparable levels of heterogeneous products, presumably due to mapping imperfections. We therefore concluded that inter-cellular contamination is not a significant source of noise in our experiments.

3. Sequence pairing errors

One possible, although somewhat unexpected source of noise in the single cell Hi-C map is imperfect read-pairing during the sequencing process. To test this idea, we added control phiX174 DNA to single cell Hi-C mouse libraries just before sequencing and searched for spurious mouse-phiX174 read-pairs in the sequencing read outs. We detected very low, but

consistent levels of spurious read-pairs, which allowed us to estimate a global rate of error in pairing at around 0.1% (0.04% to 0.012%, **Extended Data Table 3**). Spurious pairs are further filtered out by the requirement that both read-ends must begin with identical matching identification tags (which will happen in probability 1/n, where n is the number of multiplexed cells in the lane, usually around 9-12 in this study). While this represents a low error rate, it can generate a significant number of distinct pairs (a few thousands pairs per cell) when sequencing is saturating the single cell library complexity and when most of the products are sequenced many times. Since these spurious pairs are created after amplification, each of them appears only once. The fraction of *trans* pairs among these singly covered read-pairs is consistently higher than that of read-pairs covered more than once (average across cells of 34% versus 24%), reinforcing our observation that this group of read-pairs contains more spurious pairs than the rest of the read-pairs. Since the sparse single cell maps are extremely noise sensitive, we prefer to lose some valid pairs in the process of eliminating spurious ones, and therefore decided to discard all singly covered read-pairs. In summary, we require that at least two reads support each ligation event, and by that effectively eliminate the flow cell pairing errors.

Testing sequencing saturation

To test if the sequencing depth applied to our single cell libraries is sufficient to approach the limits of the library complexity, we studied the distribution of read ligation product coverage in our samples. In addition, several libraries were selected for re-sequencing, and the correlation in coverage value, together with the number of new valid fend-pairs (e.g. having cumulative coverage larger than 1) was explored. We find that adding 4-13.8 million reads to libraries of size 1.1-2.4 million reads resulted in addition of 13-47% new valid fend-pairs. In particular, we found that fend-pairs covered only once in one library, were re-sequenced in a second run of the libraries in 34-71% of the cases, suggesting the frequency of sequence pairing errors resulting in spurious products within uniquely covered fend-pairs is around 30-70% (**Extended Data Table 1**).

Ensemble map analysis and domain identification

Ensemble Hi-C maps were pre-processed and analyzed as described[1]. We then modeled the ensemble Hi-C map as previously described for *Drosophila* maps[2], computing optimized scaling factors throughout the genome. The topmost 5th percentile of 20 kb-smoothed inferred scaling factors were used to demarcate domains. We discarded domains with general low expected coverage (less than 10,000 potential mappable intra-domain fend-pairs or in poorly mappable regions), resulting in 1403 domains with a median length of 1.13 Mb.

Generating a pooled single cell Hi-C map and comparing it to an ensemble Hi-C map

We pooled together fend-pairs from high quality maps (QC criteria A and B but not C in the above; Bgl II as RE1), combining 2,295,341 fend-pairs from 60 samples. Technical biases

were then corrected using a previously described model[1]. The technical features learnt and corrected were fragment length (from one Bgl II site to the next), fragment end length (from one Bgl II site to the upstream Alu I site), and fragment end GC content. The technical biases found were very small compared to the ensemble Hi-C biases, supporting the idea that many of the common Hi-C biases are adequately accounted for during the single cell Hi-C processing pipeline.

We then used the pooled and ensemble maps to estimate the dependency between contact probability and distance, using the ratio between observed and potential fend-pairs within the same chromosome and separated by genomic distances:

F(d) ~ number of observed fend-pairs at distance bin d / number of potential fend-pairs at distance bin d

The similarity between the ensemble Hi-C and the pool of single cell Hi-C samples contact maps was next estimated by computing the number of contacts observed in 1 Mb square bins, and normalizing the coverage by the F(d), d being the distance of the bin from the diagonal.

Controlling analysis using reshuffled single cell maps

We used reshuffled single cell Hi-C maps to control for the statistical conclusions on cell-to-cell contact map variability and insulation score (**Fig. 2c-2f**, **Extended Data Figs. 2b**, **2c**, **3c**, **4d** and **6b**). Reshuffling was designed to preserve the essential properties of the single cell maps in the randomized dataset, including the fend coverage depth, the distribution of coverage per specific fends, the typical association between genomic distance and contact probability, and the ratio between inter- and intra- chromosomal contact distributions. To this end, we generated a pool of contacts (fend-pairs) from all 9 single cell datasets for which Bgl II was used as RE1, and repartitioned them to reshuffled cells using the original cell fend-pair counts. This procedure was repeated 50 times to allow for computation of the background distribution of each of the single cell map property that was observed.

A different reshuffling approach was used to assess the mean *trans* enrichment of domains (**Fig. 4a**). We were interested to test if some domains have consistently higher or lower coverage. Simply shuffling fend-pairs among the set of single cell maps might have little effect on that attribute, since the coverage of domains that are highly enriched or depleted in all the single cell maps will remain the same after the reshuffling. Therefore we reshuffled the fend-pairs within each cell independently, repeating the following for each fend-pair. For a fend-pair (f_1, f_2), where f_1 is on a certain chromosome, we randomly selected another fend-pair (f_1', f_2') where f_1' is on the same chromosome as f_1. We than switched f_2 and f_2', which resulted in two new fend-pairs: (f_1, f_2') and (f_1',f_2). This method preserves the fend-pair coverage of each fend as well as the number of fend-pairs between each pair of chromosomes, which specifically maintains the *cis-trans* coverage ratio.

For the control of **Fig. 2b** we calculated the intra/inter ratio with shifted domains. A random number was added to the domains coordinates of each chromosome separately,

modulus the chromosome length, which resulted in cyclically shifting the domains, preserving their number, sizes and order (with 1000 repetitions).

Computing insulation score

The insulation score of a fend aims to capture how many contacts are formed between the chromosomal regions upstream and downstream of it (crossing it). The number of actual contacts crossing over a fend is compared to the expected number assuming a uniform distribution of the contacts within the same chromosome across all possible fend-pairs in it. We consider contacts between fends distanced up to 40 Mb from each other.

The insulation score for fend *s* is:

$$f(s) = -\log_{10} \frac{O_s + C}{\frac{P_s}{\sum_{t \in chr(s)} P_t} \cdot \sum_{t \in chr(s)} O_t + C}$$

Where O_s is the number of contacts crossing over fend *s*, P_s is the total number of all possible fend-pairs crossing over fend *s*, *chr(s)* is the fend *s* chromosome, and *C* is a regularisation factor (C = 0.05). We compared the distribution of insulation scores in the real cells by calculating insulation scores for 50 sets of reshuffled cells.

Epigenetic annotation of domains

For the association of domains with H3K4me3 we extracted the topmost 1st percentile peaks of H3K4me3 ChIP-Seq data smoothed by a 50 bp sliding window on T_H1 cells[3], and calculated the peak density per domain (**Fig. 4d**). Note that 27% of the domains have no H3K4me3 peaks at all. These are therefore always used when analyzing inactive domains.

For the association of domains with the nuclear lamina (**Extended Data Fig. 5f-i**) we calculated the mean mESC LaminB1-DamID enrichments[4] of the probes within the domain.

Defining and exploring chromosomal interfaces

To allow robust identification of interfaces between chromosomes in the single cell contact maps, we associated each *trans*-chromosomal contact with the two *cis* domains encompassing its fends, and used only pairs of domains with at least two contacts.

To analyse chromosomal interfaces between linearly adjacent domains (**Fig. 5b**), we considered each of the two loci involved separately for every *trans* contact and classified all *trans* contacts located within 200 - 600 kb from one fend and up to 4 Mb from the other fend.

If nearby contacts connected the same two domains as the contact associated with the fend under consideration, we classified them as "domain-domain". If the nearby contacts connected different domains on the other chromosome we classified them as "domain-chromosome". Contacts connecting different domains on both chromosomes were classified as "chromosome-chromosome". Contacts near domain boundaries (up to 100 kb) were ignored, since the boundary position is not exact. These observations were first normalized by the expected counts from a random uniform distribution of the cell's *trans*-chromosomal contacts. The domain-domain and domain-chromosome ratios were then scaled by the chromosome-chromosome ratio to show the enrichments. Contacts were binned to 100 kb square bins for the analysis.

To generate a high-resolution identification of complex chromosomal interfaces, we visualised contacts between selected pairs of chromosomes, highlighting a configuration involving interactions between multiple domains (**Extended Data Fig. 7f**).

X chromosome structural model calculation and analysis

Three-dimensional (3D) particle-on-a-string models of X chromosome structures were calculated using restrained molecular dynamics simulations, where the experimental distance restraints were derived from the Hi-C data. The calculations employed a simulated annealing protocol to generate energy minimized structural models that satisfy all the experimental restraints. A combination of unambiguous distance restraints from intra-chromosomal contacts in the single cell Hi-C data was used, together with anti-distance restraints between regions that were found not to contact each other in the ensemble Hi-C data derived from millions of cells. To filter out noise the unambiguous distance restraints were derived from a subset of the experimental contacts, which were found to be part of a self-supporting sub-network. A contact was identified as supported if within 2 Mb of each of the contacted loci there was either another contact or two further interactions which both contact the same intermediate 2 Mb region. In the simulated annealing based structure calculations a series of initially extended structures with random conformations were slowly cooled from a simulated high temperature until the temperature was near to zero Kelvin during the calculation of a particle dynamics trajectory (see below for more details).

In order to calculate the structure of the X chromosome in individual cells we have used the simplest possible "beads-on-a-string" polymer model and have made minimal assumptions. Currently, the data is too sparse to calculate structures at the resolution of either a nucleosome, or indeed the length of the restriction fragments that are cross-linked in the Hi-C process. We have, therefore, calculated beads-on-a-string structures at two illustrative resolutions, where each backbone particle represents either 50 kb or 500 kb of DNA.

Fine-scale structural models were primarily constructed of particles representing the backbone path of the chromosome which were placed at regularly spaced 50 kb intervals. These were restrained with an upper distance limit of 0.15 μm between adjacent particles (and no lower limit), in order to generate a beads-on-a-string polymer model of the chromosome backbone. This backbone distance scaling (0.15 μm per 50 kb) was chosen to give a folded chromosome territory of roughly the right size and was found to give no restraint violations. We then inserted extra particles into this backbone wherever we observed a single cell Hi-C contact. These extra particles were placed at their precise sequence position in the chromosome backbone, intermediate to the backbone particle positions (and possibly other neighbouring Hi-C contact particles). Such intermediate particles were restrained to their neighbouring particles with upper distance limits set from their sequence separation with a scale of 3.0 μm per megabase, corresponding to an appropriate fraction of the 0.15 μm per 50 kb backbone spacing.

The Hi-C contact restraints used a flat-bottomed harmonic potential and were set to a distance of 60 nm, and had lower and upper distance limits of 30% and 200% of the target distance, respectively. The distance scale used for the data-derived restraints (and also for the inter-particle repulsive energy term) was based on the close packing of two chromatin fibres, and the fractional restraint limits are based on the size range of the fragments in the DNA

libraries that were sequenced. As illustrated by the violation analysis (**Extended Data Fig. 3c**), the upper limit is clearly the most important parameter. At the current level of precision the overall chromosome conformations are not sensitive to the actual restraint distance used, as long as it is proportionately small compared to the spacing between backbone particles. (Test calculations with closer distance restraints, by a factor of 10, show no practical differences in the structures.)

Additional repulsive distance restraints were derived from regions observed *not to* interact in the ensemble Hi-C dataset. These non-interaction restraints were assigned a lower distance limit of 0.30 μm and no upper limit and were included where a given 250 kb region of the chromosome had no observed contacts with another region of the same size. All particles were subject to a general repulsive term during the low temperature steps of the annealing protocol, to prevent their superposition.

Based on the statistical analyses (see above) and the presence of high numbers of *cis* restraints within the X chromosome (> 200), we chose the six highest quality single cell datasets for performing structure calculations. The cell-9 dataset was also included for comparison, as an example of a dataset that passed the statistical quality checks, but has a smaller number of restraints (126).

Simulated annealing force field and protocol

A simple force field was employed to move the particles, and thus allow condensation of the structural model during the annealing schedule. The simulated force field used two terms: a general repulsive term that prevents the superposition of particles; and a distance restraint term to restrain the backbone path, the data-derived interactions and the non-interacting regions.

General particle repulsion

A general repulsive force occurring within a limiting inter-particle distance of 60 nm was set to be proportional to the square of the distance inside this limit. Here the force constant was weaker than the force constant for the distance restraints by a factor of 25 ($k_2 = 25$, $k_1 = 1$). As described below the general repulsive term was not used for the initial high-temperature stages of the annealing protocol.

$$F_{rep} = k_1 (d - d_{lim})^2, \text{ if } d < d_{lim}$$

$$F_{rep} = 0, \text{ if } d \geq d_{lim}$$

Distance restraints

The distance restraints used a flat-bottomed potential, which is harmonic (i.e. the force is proportional to the separation) at large distances. The potential was divided into four distinct regions relative to the upper and lower bounds of the distance restraint as follows:

1) If restrained particles are closer than the lower restraint limit then the force is repulsive and proportional to the square of the difference between the inter-particle distance

and the lower limit.

$F_{restraint} = k_2 (d - d_{lower})^2$, if $d < d_{lower}$

2) Inside the upper and lower bounds of the distance restraint no force is applied.

$F_{restraint} = 0$, if $d_{lower} \leq d \leq d_{upper}$

3) For inter-particle distances above the upper restraint limit, but not exceeding an extra 30 nm, the force is attractive and proportional to the square of the difference between the inter-particle distance and the upper limit.

$F_{restraint} = -k_2 (d - d_{upper})^2$, if $d_{upper} < d < d_{upper} + 30$

4) For inter-particle distances above the upper restraint limit, and exceeding the extra 30 nm, the force is attractive and proportional to the difference between the inter-particle distance and the upper limit plus 15 nm.

$F_{restraint} = -k_2 (d_{upper} + 15 - d)$, if $d \geq d_{upper} + 30$

Structure calculations started from roughly circular extended conformations of the chromosome backbone particles. The circular diameter was 6.0 μm and the randomization was uniform in the radial direction ± 0.3 μm. The choice of a circular conformation gave slightly quicker convergence during the annealing protocol, likely because in this arrangement each particle will initially have a clear and direct path to any other particle, to which it may be restrained.

Annealing protocol

The structure calculation used the following annealing schedule (in all instances), to decrease the effective temperature (see Table below). The full general repulsive term was used only in the final stages and (when used), the anti-distance restraints derived from non-interaction data were applied at all stages of the protocol. Within each stage the temperature was decreased uniformly over a given number of steps and at each temperature a number of 'motion' steps was performed where particle forces, velocities and positions were updated. All particles were simulated to have unit masses.

Start Temp (K)	End Temp (K)	Temp Steps	Motion Steps	Repulsive Scale
10000	5000	4	1000	0.00
5000	3000	10	1200	0.01
3000	1000	30	1200	0.10
1000	300	70	1200	0.50
300	25	100	3400	1.00

Start Temp (K)	End Temp (K)	Temp Steps	Motion Steps	Repulsive Scale
25	0.01	3	22000	1.00

The annealing schedule is significantly longer, in terms of temperature and motion steps, than required for some of the data sets to converge (to consistent bundles of structural models).

Low-resolution models

Simpler, low-resolution models of the X chromosome structures were generated by grouping the single-cell Hi-C interaction points into 500 kb regions. In these calculations the backbone path of the chromosome was represented by particles separated by 500 kb, which also formed the anchor points for data-derived distance restraints. In contrast to the fine-scale structures, there were no particles intermediate to the regularly spaced backbone. Although this results in particles that may be restrained by multiple, potentially antagonistic, restraints it yields structures which are better determined and where the overall chromosome conformation is more easily visualised. In addition, no repulsive non-interaction restraints (i.e. derived from the population Hi-C data) were used in the calculation of these simpler models.

The annealing schedule and energy terms used for generating low-resolution models were the same as that used in the fine-scale models. For data derived (single-cell Hi-C) contacts the restraint distances were, however, set differently as proportional to one over the number of interactions squared (i.e. between the same 500 kb regions), with lower and upper restraint limits of ± 20 % either side. For the backbone path the restraint distance was set to be the same as a single interaction, but with no lower limit and an upper limit at the restraint distance (i.e. modified by +0% and -100%, respectively). The distance scale was set for the structure calculation so that the backbone restraint distance and the general repulsive term was the same, in model coordinates, as the in fine-scale models. The structural models were then scaled after the calculation so that the average size of the X chromosomes for all samples, along the longest axis, was the same as for the fine-scale models.

Structures from shuffled data sets

A randomly shuffled dataset was created to investigate what the calculated structures would look like in the absence of an underlying chromosome conformation to guide the interactions. The shuffled interaction dataset was made from the distance restraints of the single-cell dataset with the largest number of intra-chromosomal contacts for the X chromosome (cell-1). The complement of interacting sequence locations was preserved but the pairings of positions to generate distance restraints was randomized. The probability of selecting any two positions to pair was scaled with one over the square root of the sequence separation, which arises from considering the proportion of routes with zero travel distance after a random walk in $3D^5$ with an increasing numbers of steps. The structure calculation was performed in the same manner as for the fine-scale (50 kb backbone) structures and models were generated for both different randomly shuffled datasets and for repeat calculations using the same shuffled dataset.

Combined data sets

To test whether different single-cell data sets are compatible with a single underlying chromosome conformation, fine-scale structure calculations were performed on combined data sets made by merging interactions from two different single-cell data sets using all possible combinations of cell-1 trough cell-4 (those with the largest amount of experimental Hi-C data). The pairwise interaction data was merged prior to the selection of self-supporting restraints for the structure calculation. All other aspects of the structure calculations were the same as for the 50 kb scale models.

Violation analysis and distance matrices

Restraint violation analyses of the X chromosome structures was performed by measuring the distances between particles that were directly restrained by single-cell Hi-C interaction data in the fine-scale models. Distance distributions were obtained for the models generated from the single-cell Hi-C contacts, the shuffled datasets and the combined dataset. If the model distance exceeded the upper bound of the distance restraint (120 nm) the restraint was deemed to be violated.

Distance matrices were created to give a visual map of sequence positions that are in close proximity in the structural models for comparison with the underlying data-derived restraints as well as the population Hi-C interaction data. For all datasets, the distance between backbone particles was recorded and averaged over the structural models. For the fine-scale structures the distances between 50 kb separated backbone particles were analysed directly. For larger scale analysis of whole chromosomes the particles were grouped into adjacent 250 kb regions and the minimum distance between any pair or particles was recorded. The array of distances was then plotted as a colour density matrix where the density was derived from the logarithm of the distance. For both single-cell and combined matrices, the maximum intensity of the colour scales (at the minimum distance) was bounded at 10 nm and the minimum intensity at 1.0 μm.

Structure comparisons

The coordinates of pairs of structural models were superposed using singular value decomposition (SVD), and the root-mean-square deviation (RMSD) between models was then used as a measure of coordinate variation. When comparing models derived from repeat calculations from the same dataset, we superposed the coordinates of the Hi-C contact particles. When comparing models derived from different datasets, we superposed the coordinates of the regularly spaced backbone particles (which are common between different structure ensembles).

Hierarchical clustering of pairwise RMSD values was then used to compare each model with all other models within a structural ensemble and also to compare pairs of complete structure ensembles derived from different datasets. A plot of the resulting matrix (with diagonal elements excluded), where a lighter shade represents a more similar structure, was used to illustrate the presence of reproducible conformational modes. Four representative but conformationally distinct models from each structure ensemble were obtained by using k-

medoids clustering. The medoid model of each cluster was used for graphical visualisation to illustrate the range of conformations generated by repeat calculations (**Fig. 3c**). The coordinates of the structural models are available online in a PDB format.

Density and depth analyses

The particle densities of positions within a structure were calculated to identify regions within the sequence that were consistently buried or exposed. To do this the density of a particle was initially calculated by taking the sum of one over the distances squared of the particle to all other particles. For a given structural model the particles were ranked according to this density and the density was expressed as a fraction in the interval [0,1]. The rank fractions from different models were then averaged over all the models from all the data sets (200 repeat calculations on six datasets).

In order to investigate the depth of particular loci relative to the predicted edge of the X chromosome territory, a surface was defined for each of the structural models. First, the 3D coordinate space was divided into a cubic grid and the structural density in each voxel was defined by summing radial contributions from each particle, using a normal volume distribution. Second, we defined a 3D density iso-surface as those voxels that border the region of significant density and recorded the depth of each structural particle (for each X chromosome structural model) as the minimum distance to this surface.

For the analysis of the structure depth of *cis*- and *trans*-chromosome locations, relative to a null hypothesis, relative entropy values were calculated. These are the components of the sum of Kullback-Leibler divergence. Here, for each category, the observed fraction (***Obs***) of points at a given structure depth is compared to the expected fraction (***Exp***) of points at the same depth resulting from a large random, uniform selection of sequence positions along the chromosome, which excluded unmappable regions. For a given depth bin the divergence (***d***) is plotted as:

$$d = Obs \log_2(Obs/Exp)$$

Hence, a value near zero indicates that the data closely match the null hypothesis or the amount of data was low.

Validation of calculation protocols with synthetic structures

In order to define the accuracy and precision that we might expect from the structure calculation protocol *per se*, given the current numbers of experimental restraints, we created synthetic datasets for a 3D Hilbert curve (a non-overlapping, compact fractal path) as a test structure. The Hilbert curve particle-on-a-string test structure was constructed from a cube of 4096 points where the separation between sequentially adjacent points is equivalent to 50 kb, corresponding in total to a large 204 Mb chromosome. A distance scale was also introduced to the Hilbert curve based on an estimate of 0.15 μm per 50 kb, as used in the fine-scale X

chromosome modelling. To investigate the effect of restraint density on the resulting structural models, spatially adjacent points were randomly chosen to provide a specific number of synthetic restraints. These distance restraints used a harmonic potential with a target distance equal to the spacing between adjacent Hilbert curve points.

Randomised extended conformations of 4096 points were then annealed using the synthetic restraints and the same simulated annealing protocol that was used for the X chromosome models. Multiple models were generated from the Hilbert curve data and superposed with one another and the original 3D Hilbert curve using SVD. For a given number of randomly selected restraints, the mean pair-wise RMSD between calculated models was used to estimate the coordinate precision and the mean RMSD between the calculated models and the original Hilbert curve was used to estimate the accuracy that may be expected from the structure generation procedure.

Supplementary references

1. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* **43**, 1059-65 (2011).
2. Sexton, T. et al. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458-72 (2012).
3. Deaton, A.M. et al. Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Res* **21**, 1074-86 (2011).
4. Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S. W. M., Solovei, I., Brugman, W., Gräf, S., et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Molecular cell, 38(4), 603–13 (2010).
5. A. Einstein, Investigations on the Theory of Brownian Movement (Dover, 1956).