Single cell Hi-C reveals cell-to-cell variability in chromosome structure

Takashi Nagano[#1], **Yaniv Lubling**[#2], **Tim J. Stevens**[#3], **Stefan Schoenfelder**[1], **Eitan Yaffe**[2], **Wendy Dean**[4], **Ernest D. Laue**[3], **Amos Tanay**[2], and **Peter Fraser**[1]

[1]Nuclear Dynamics Programme, The Babraham Institute, Cambridge, UK

[2]Department of Computer Science and Applied Mathematics and Department of Biological Regulation, Weizmann Institute, Rehovot, Israel

[3]Department of Biochemistry, University of Cambridge, UK

[4]Epigenetics Programme, The Babraham Institute, Cambridge, UK

[#] These authors contributed equally to this work.

Abstract

Large-scale chromosome structure and spatial nuclear arrangement have been linked to control of gene expression and DNA replication and repair. Genomic techniques based on chromosome conformation capture assess contacts for millions of loci simultaneously, but do so by averaging chromosome conformations from millions of nuclei. Here we introduce single cell Hi-C, combined with genome-wide statistical analysis and structural modeling of single copy X chromosomes, to show that individual chromosomes maintain domain organisation at the megabase scale, but show variable cell-to-cell chromosome territory structures at larger scales. Despite this structural stochasticity, localisation of active gene domains to boundaries of territories is a hallmark of chromosomal conformation. Single cell Hi-C data bridge current gaps between genomics and microscopy studies of chromosomes, demonstrating how modular organisation underlies dynamic chromosome structure, and how this structure is probabilistically linked with genome activity patterns.

Chromosome conformation capture[1] (3C) and derivative methods (4C, 5C and Hi-C)[2-6] have enabled the detection of chromosome organisation in the 3D space of the nucleus. These methods assess millions of cells and are increasingly used to calculate conformations of a range of genomic regions, from individual loci to whole genomes[3,7-11]. However, fluorescence in situ hybridisation (FISH) analyses show that genotypically and phenotypically identical cells have non-random, but highly variable genome and chromosome conformations[4,12,13] probably due to the dynamic and stochastic nature of chromosomal structures[14-16]. Therefore, whilst 3C-based analyses can be used to estimate

Correspondence and requests for materials should be addressed to PF (peter.fraser@babraham.ac.uk) for the single cell Hi-C method, AT (amos.tanay@weizmann.ac.il) for the statistical analysis, or EDL (e.d.laue@bioc.cam.ac.uk) for the structural modelling..

Author Contributions TN and PF devised the single cell Hi-C method. TN performed single cell Hi-C and DNA FISH experiments. SS carried out ensemble Hi-C experiments. WD microscopically isolated single cells. YL, EY and AT processed and statistically analyzed the sequence data. TJS and EDL developed the approach to structural modelling and analysed X chromosome structures. TJS wrote the software for 3D modeling, analysis and visualisation of chromosome structures. TN, YL, TJS, EDL, AT and PF contributed to writing the manuscript, with inputs from all other authors.

Data deposited in NCBI's Gene Expression Omnibus (Nagano *et al.*, 2013) and are accessible through GEO Series accession number GSE48262 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48262).

The authors declare that they have no competing financial interests.

an average conformation, it cannot be assumed to represent one simple and recurrent chromosomal structure. To move from probabilistic chromosome conformations averaged from millions of cells towards determination of chromosome and genome structure in individual cells, we developed single cell Hi-C, which has the power to detect thousands of simultaneous chromatin contacts in a single cell.

Single cell Hi-C

We modified the conventional or "ensemble" Hi-C protocol[3] to create a method to determine the contacts in an individual nucleus (Fig. 1a, Supplementary Information). We used male, mouse, spleenic $CD4^+$ T cells, differentiated *in vitro* to T helper (Th1) cells to produce a population of cells (>95% $CD4^+$), of which 69% have 2n genome content, reflecting mature cell withdrawal from the cell cycle. Chromatin cross-linking, restriction enzyme (Bgl II or Dpn II) digestion, biotin fill-in and ligation were performed in nuclei (Fig. 1a and Extended Data Fig. 1a) as opposed to ensemble Hi-C where ligation is performed after nuclear lysis and dilution of chromatin complexes[3]. We then selected individual nuclei under the microscope, placed them in individual tubes, reversed cross-links, and purified biotinylated Hi-C ligation junctions on streptavidin-coated beads. The captured ligation junctions were then digested with a second restiction enzyme (Alu I) to fragment the DNA, and ligated to customized Illumina adapters with unique 3 bp identification tags. Single cell Hi-C libraries were then PCR amplified, size selected and characterized by multiplexed, paired-end sequencing.

De-multiplexed single cell Hi-C libraries were next filtered thoroughly to systematically remove several sources of noise (Extended Data Fig. 1b-f, Supplementary Information). Hi-C in male diploid cells can theoretically give rise to at most two ligation products per autosomal restriction fragment end, and one product per fragment end from the single X chromosome. Using Bgl II, the total number of distinct mappable fragment-end pairs per single cell cannot therefore exceed 1,201,870 (Extended Data Fig. 1g, Supplementary Information). In practice, deep sequencing of the single cell Hi-C libraries demonstrated that following stringent filtering our current scheme allows recovery of up to 2.5% of this theoretical potential, and has identified at least 1000 distinct Hi-C pairings in half (37/74) of the cells. Deep sequencing confirmed saturation of the libraries' complexity, and allowed elimination of spurious flow cell read pairings and additional biases (Extended Data Tables 1-3). Based on additional quality metrics we selected ten single cell datasets, containing 11,159-30,671 distinct fragment-end pairs for subsequent in-depth analysis (Extended Data Fig. 1h-l). Visualization of the single cell maps suggested that despite their inherent sparseness, they clearly reflect hallmarks of chromosomal organization, including frequent *cis*-contacts along the matrix diagonal and notably, highly clustered *trans*-chromosomal contacts between specific chromosomes (Fig. 1b).

Single cell and ensemble Hi-C similarity

We used the same population of $CD4^+$ Th1 cells to generate an ensemble Hi-C library. Sequencing and analysis[17] of 190 million read pairs produced a contact map representing the mean contact enrichments within approximately 10 million nuclei. The probability of observing a contact between two chromosomal elements decays with linear distance following a power law regime for distances larger than 100 kb[3,18]. We found similar regimes for the ensemble, individual cells and a pool of 60 single cells (Fig. 1c). Moreover, after normalizing the matrices given this canonical trend, comparison of intra-chromosomal interaction intensities for the pool and ensemble, by global correlation analysis of contact enrichment values at 1 Mb resolution generates a highly significant correspondence (Fig. 1d). This is emphasized by the high similarity observed in comparisons of individual chromosomes from ensemble and pooled Hi-C maps (Fig. 1e). In summary, despite different experimental procedures and sparse nature of the single cell matrices, the pooled matrix retains the most prominent properties of the ensemble map, confirming the validity of the approach and prompting us to further explore the similarities and differences among the individual cell chromosomal conformations.

Intra- and inter-domain contacts

A key architectural feature of ensemble Hi-C datasets is their topological domain structure[18-20]. As expected 1403 domains were identified in the Th1 cell ensemble Hi-C map[18] (Supplementary Information Table 1, and Supplementary Information). We used the ensemble domains to ask whether the same domain structure can be observed at the single cell level. Visual inspection of the domain structure overlaid on individual intra-chromosomal contact maps (Fig. 2a), and global statistical analysis of the ratios between intra- and inter-domain contact intensities in individual cells (approximately 2-fold enrichment on scales of 100 Kb to 1 Mb, Fig. 2b and Extended Data Fig. 2a), both supported the idea that domains are observed consistently in the single cell maps. To test whether domain structures are variable between individual cells, we estimated the distributions of intra-domain contact enrichments across cells and compared it to the distributions derived from reshuffled maps. We reasoned that cell-to-cell variation in intra-domain contact intensities would result in an increase of the variance of this distribution compared to the expected variance resulting from sampling contacts in uniformly (shuffled) intra-domain contacts. The data however (Fig. 2c), showed that the distributions for the intra-domain enrichments in real cells are not more varied than expected (Kolmogorov-Smirnov $p < 0.52$). A similar observation was derived by comparison of the correlations between intra-domain contact enrichments for pairs of real and pairs of reshuffled maps (Extended Data Fig. 2b). While this analysis cannot quantify variability in the high-resolution internal structure of domains, the data suggests that domain intactness is generally conserved at the single cell level.

Visual comparison of whole chromosome contact maps (Fig. 2d) suggested that unlike intra-domain interactions, inter-domain contacts within single cell chromosomes are structured non-uniformly. The maps showed large-scale structures as indicated, for example, by specific insulation points separating chromosomes into two or more mega-domains in a cell-specific fashion. To rule out the possibility that this can be explained by sparse sampling of contacts in each single cell map we again used reshuffled controls. In each map (real or randomized) we quantified the frequency of loci that strongly polarize the matrix into two weakly connected submatrices (using an insulation score; Supplementary Methods). We confirmed that single cell maps indeed show many more such loci than reshuffled maps (Fig. 2e and Extended Data Fig. 2c). The reshuffled controls made by mixing contacts from different single cell maps, are in fact similar to sparse versions of the ensemble map, which do not show specific structure at the intra-domain level. Along similar lines, the correlation in contact intensities between domains on the same chromosome in pairs of single cell maps is lower compared to reshuffled controls (Fig 2f). Taken together, these data show that domains form a robust and recurrent conformational basis that is evident in each of the single cells. However, inter-domain contacts are highly variable between individual cells, suggesting large-scale differences in higher-order chromosome folding that are obscured in ensemble maps, averaged over mil1ar rm((Figeci, toyaoathat)Tmmof).o2 mil1ame chromosome incW single-copy, male X chromosome. We used intra-chromosomal contacts as distance restraints and calculated structural models using a simulated annealing protocol to condense a particle-on-a-string representation of individual chromosomes from random initial conformations (Supplementary Information), to produce both fine-scale and low-resolution models, with backbone particles representing either 50 or 500 kb of the chromosome, respectively. For fine-scale calculations, each intra-chromosomal contact restrained its precise position on the chromosome, while low-resolution calculations combined contacts into larger bins. Tests of our simulation protocol demonstrated that restraint density was the most important parameter for modeling (Extended Data Fig. 3a, b). Hence, from the ten high-quality single cell datasets, we selected six with the largest numbers of intra-chromosomal X contacts, plus one with a lower number of contacts (cell-9) for contrast.

Repeat calculations starting from random positions generated 200 X chromosome models for each cell at both scales. The fine-scale models displayed very low numbers of restraint violations (Extended Data Fig. 3c). We introduced an estimated average unit DNA distance highly variable inter-domain structure of the X chromosomal territory, some of its key organisational properties are robustly observed across the cell population.

Domains at the interface

Overlaying data from *trans*-chromosomal contacts on the X chromosome models demonstrates that *trans*-chromosomal contacting regions are strongly enriched toward the inferred surface of the models (Fig. 3h), providing further validation. These observations prompted us to further explore the structural characteristics of interfaces between chromosomal territories, and the relationships between such interfaces and the domain structure of the territory itself. We found that *trans*-chromosomal contact enrichments of domains vary across cells (Fig. 4a), showing a significant difference between the mean contact enrichment per domain in the real and reshuffled maps (p < 1.2e-9, Kolmogorov-Smirnov test). The higher variance of the distribution for the real data suggests some domains are more likely to contact elements on other chromosomes. Previous work has suggested that active genomic regions on the sub-domain scale often loop out of their chromosome territories[24], which might imply less defined local domain structures and disassociation from their chromosome territory. However, our analysis shows that *trans*-contacting domains retain domain organisation, as demonstrated by the intra-domain contact regions from those that are Lamin-associated, although both types of domains tend toward while maintaining some key local (domain) and global (depth from surface) organisational features.

Online Methods

Male Th1 cells were fixed and subjected to modified Hi-C, in which nuclei were maintained through restriction enzyme digestion, biotin fill-in labelling and ligation. Single nuclei were isolated and processed to prepare single cell Hi-C libraries for paired-end sequencing.

Sequences were mapped to the mouse genome, and abnormal read pairs were discarded. Read pairs that occurred only once (without duplication) in the library sequencing were removed. We chose 10 single cell datasets for further in-depth analyses based on several quality criteria (see Supplementary Information). To validate the single cell Hi-C procedure, we pooled the single cell Hi-C datasets and compared them to ensemble Hi-C dataset prepared from approximately 10 million cells essentially as described[3]. We created reshuffled datasets by randomly redistributing contacts of the analyzed single cells to create the same number of cells with the same number of contacts in each cell as a control to statistically analyse the variation among single cell datasets.

We reconstructed three-dimensional X chromosome structure models using restrained molecular dynamics calculations employing a simulated annealing protocol. A combination of unambiguous distance restraints from the X intra-chromosomal contacts in the single cell

5. Simonis M, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet. 2006; 38:1348–54. [PubMed: 17033623]

6. Zhao Z, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet. 2006; 38:1341–7. [PubMed: 17033624]

7. Duan Z, et al. A three-dimensional model of the yeast genome. Nature. 2010; 465:363–7. [PubMed: 20436457]

8. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol. 2012; 30:90–8. [PubMed: 22198700]

9. Marti-Renom MA, Mirny LA. Bridging the resolution gap in structural modeling of 3D genome organization. PLoS Comput Biol. 2011; 7:e1002125. [PubMed: 21779160]

10. Tanizawa H, et al. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic Acids Res. 2010; 38:8164–77. [PubMed: 21030438]

11. van de Werken HJ, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods. 2012; 9:969–72. [PubMed: 22961246]

12. Osborne CS, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet. 2004; 36:1065–71. [PubMed: 15361872]

28. Misteli T. Beyond the sequence: cellular organization of genome function. Cell. 2007; 128:787–800. [PubMed: 17320514]

29. Branco MR, Pombo A. Intermingling of Chromosome Territories in Interphase Suggests Role in Translocations and Transcription-Dependent Associations. PLoS Biol. 2006; 4:e138. [PubMed: 16623600]

30. Chuang CH, et al. Long-range directional movement of an interphase chromosome site. Curr Biol. 2006; 16:825–31. [PubMed: 16631592]

31. Chubb JR, Boyle S, Perry P, Bickmore WA. Chromatin motion is constrained by association with nuclear compartments in human cells. Curr Biol. 2002; 12:439–45. [PubMed: 11909528]

32. Dundr M, et al. Actin-dependent intranuclear repositioning of an active gene locus in vivo. J Cell Biol. 2007; 179:1095–103. [PubMed: 18070915]

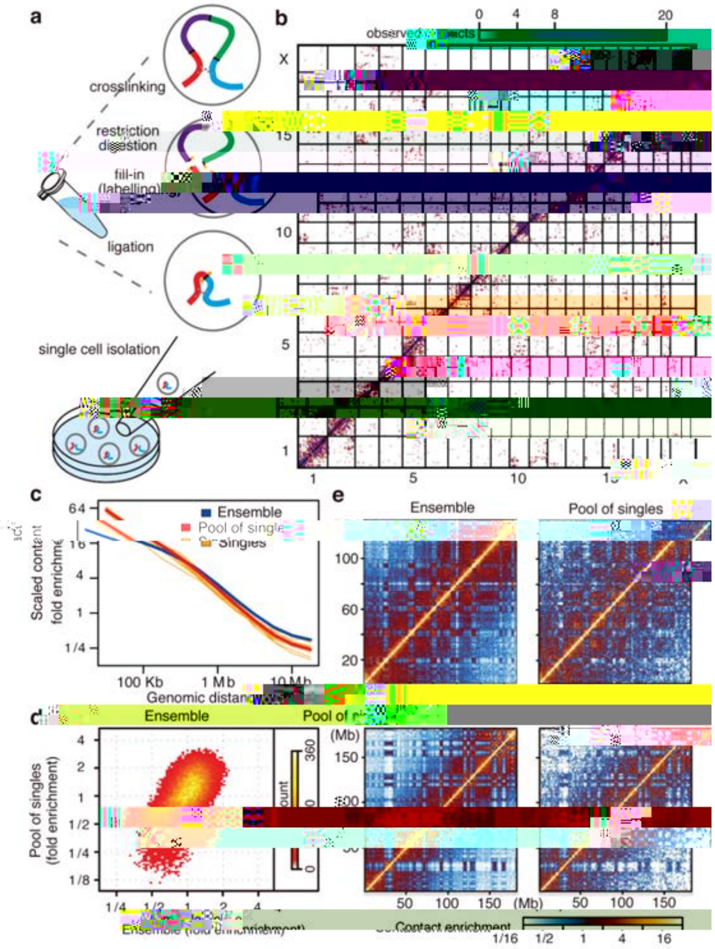

Figure 1. Single cell and ensemble Hi-C

a, Single cell Hi-C method. **b**, Single cell Hi-C heatmap (cell-5), coverage for 10 Mb bins. **c**, Contact enrichment versus genomic distance, from ensemble Hi-C, pool of 60 single cells and 10 individual cells, scaled to normalise sequencing depths. **d**, Normalising by the trends in **c**, intra-chromosomal contact enrichments for 1 Mb square bins, comparing ensemble and pooled single cell Hi-C (Spearman correlation = 0.56). **e**, Intra-chromosomal contact enrichment maps of ensemble and pooled single cell Hi-C, for chromosome 10 (top) and chromosome 2 (bottom), using variable bin sizes.

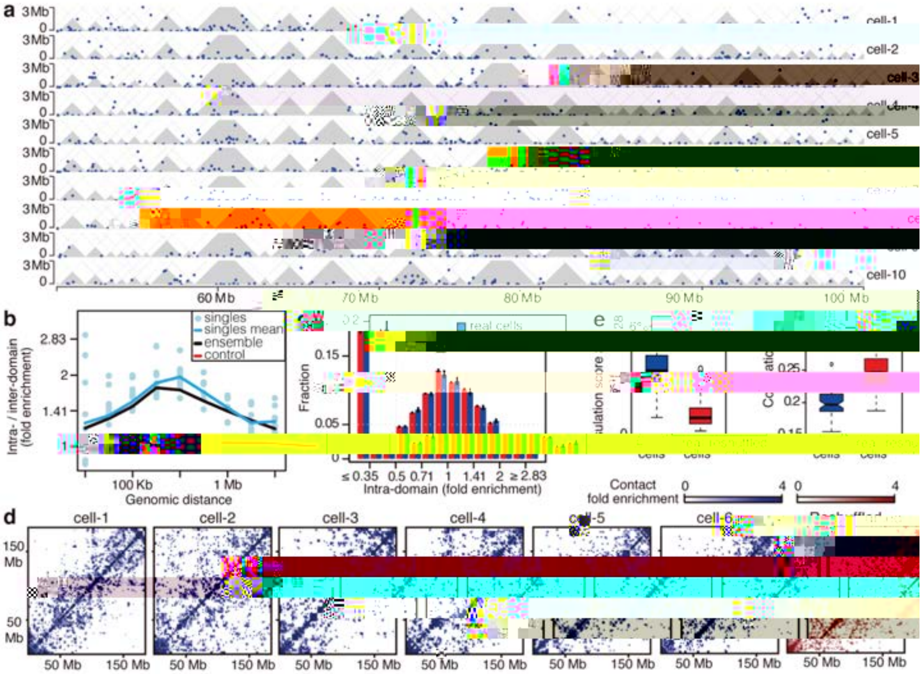

Figure 2. Conserved intra-domain, but not inter-domain structure in single cells

a, Individual intra-chromosomal contacts of 50 Mb region of chromosome 2 up to 3 Mb distance (blue dots), domains (grey). **b**, Ratios between intra-domain and inter-domain contact enrichments over genomic distance. Control is combined trend of 10 single cells calculated by repeatedly shifting the domains randomly. **c**, Distribution of intra-domain contact enrichments per domain from 9 cells (where Bgl II was used) and reshuffled datasets (black bars, standard errors). **d**, Maps of inter-domain contacts intensities for chromosome 2 from individual cells and reshuffled controls using variable bin sizes. **e**, Distribution of percentage of loci with high insulation scores in single vs. reshuffled cells. **f**, For all pairs of single cells, the correlations between inter-domain contact numbers of all pairs of domains within the same chromosome were computed. Shown are the distributions of these correlations in the real and reshuffled cells.

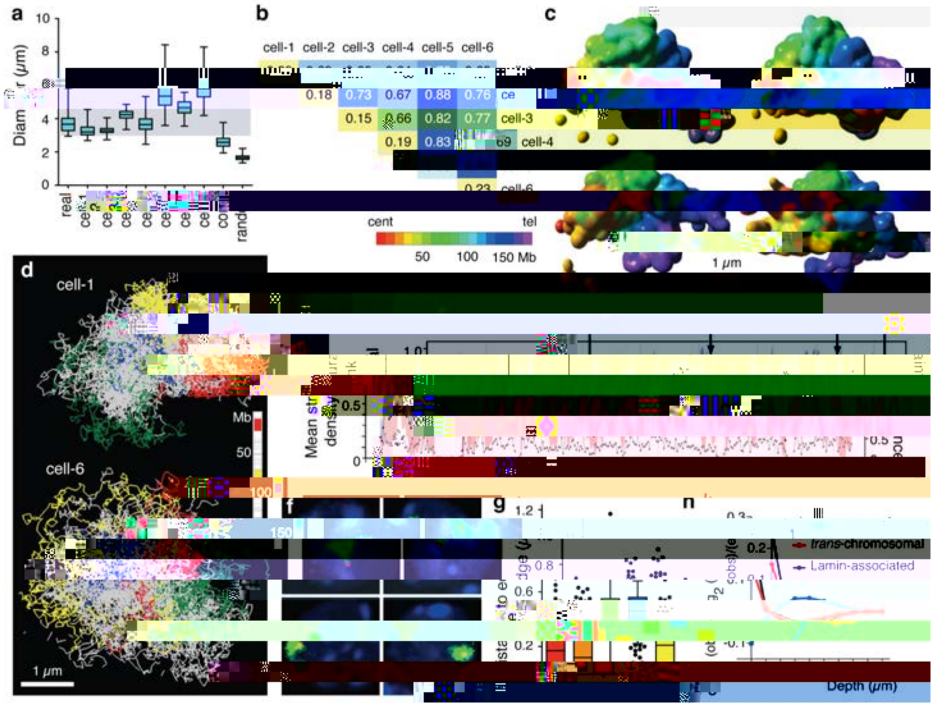

Figure 3. Structural modeling of X chromosomes

a, Distribution of longest diameter of X chromosome paint DNA FISH signals in 62 male Th1 cells (real), 200 structural models calculated for each single cell (cell-1 to -9), 200 structures from combined dataset (cell-1 and -2; comb) and 200 structures from 20 randomised cell-1 datasets (random; 10 calculations per dataset). Whiskers denote minimum and maximum. **b**, Average coordinate root-mean-square deviation (RMSD) values in microns comparing 200 low-resolution structural models for each cell and between cells. **c**, Four surface-rendered models of the X chromosome from cell-1, which are most representative of the data based on hierarchical clustering of pair-wise RMSD values (Supplementary Information). Scale bar, 1 μm. **d**, Structural ensembles of the four most representative fine-scale models for cell-1 and cell-6, with four large regions coloured. Scale bar, 1 μm. **e**, Mean structural density rank for 500 kb regions (black) from 6 × 200 fine-scale models from cell-1 to -6. Standard deviation (blue/pink). Abundance of intra-chromosomal restraints (grey, right axis). DNA FISH probes (P1-P5) are indicated. **f**, DNA FISH on Th1 cells. X chromosome paint (green) and specific locus signals (red). **g**, Distribution of DNA FISH distance measurements between signal centres for probes P1 - P5 and edge of the X chromosome territory in Th1 cells (n = 114, 113, 105, 115, 108 for P1-P5). Whiskers denote 10 and 90 percentiles. **h**, Enrichment of *cis*-, *trans*-contacts and Lamin-B1 associated domains at various depths on the chromosome models relative to null hypothesis of random positions.

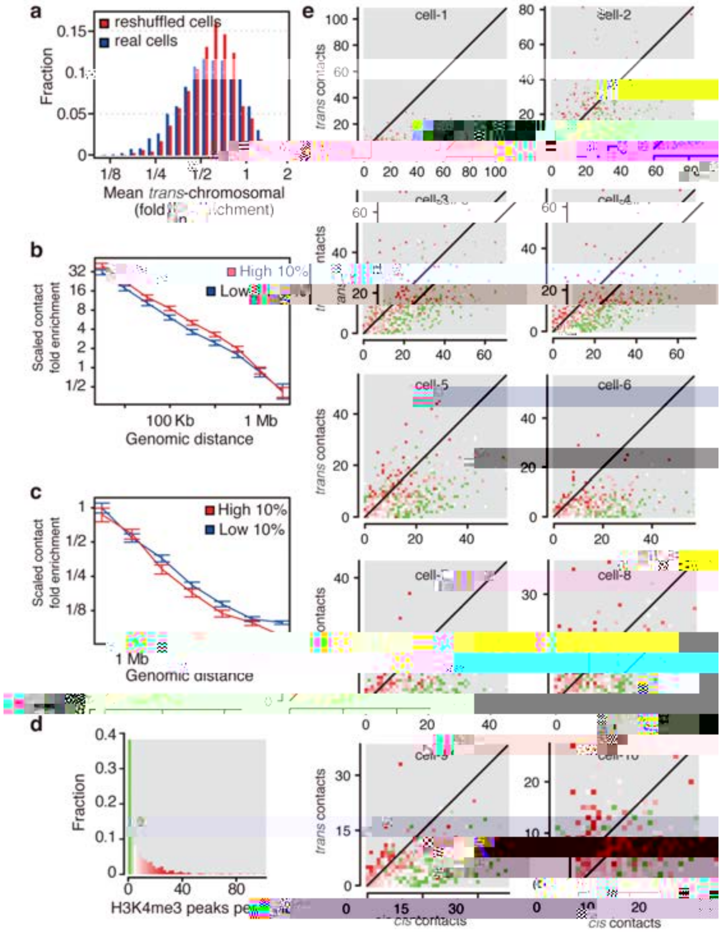

Figure 4. Active domains localise to territory interfaces

a, Distribution of *trans*-chromosomal contact enrichments of each domain averaged across real and reshuffled cells. Reshuffling maintains the number of *cis* and *trans* contacts within each cell and chromosome. **b**, Intra-domain contact enrichment over genomic distance for high vs. low *trans*-chromosomal contacting domains selected independently in each cell, with 95% confidence intervals. **c**, Same sets as in **b** but plotting the enrichment of inter-domain contacts. **d**, Distribution of H3K4me3 peak density in domains (number of peaks divided by size), color-coded according to density. **e**, Domains plotted according to number of *trans*- and *cis*-chromosomal (excluding intra-domain) contacts, color coded for H3K4me3 density as in **d**.

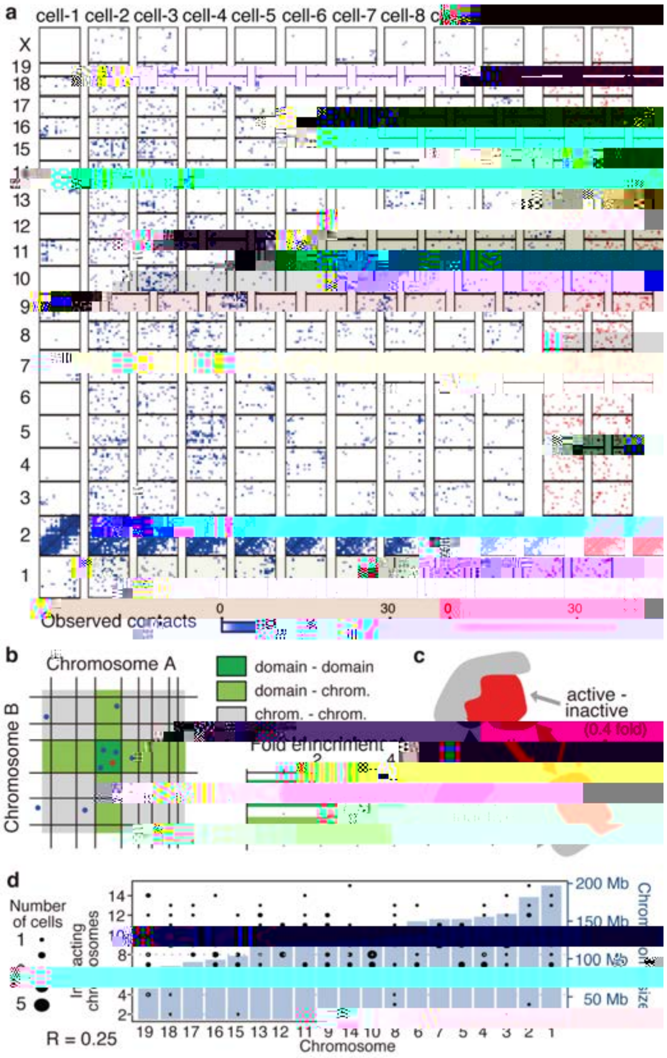

Figure 5. Chromosomal interfaces

a, All *trans*-chromosomal contacts formed by chromosome 2 in real cells (blue) and reshuffled (red). **b**, Schematic diagram of a chromosomal interface between linearly adjacent domains, their borders marked in black on two chromosomes, A and B. We considered each of the two contacting fragments of every *trans*-chromosomal contact and classified every nearby *trans*-chromosomal contact as domain-domain, domain-chromosome and chromosome-chromosome, the latter being used as background for normalisation (Supplementary Information). Contact under consideration (red), nearby contacts (blue). Fold enrichments shown for each group type (error bars, standard deviation). **c**, *Trans*-chromosomal contacts are highly significantly enriched between active domains (H3K4me3 enriched) or between inactive domains, but not mixed interaction (chi-square test; *p* = 5.8e-18; even after taking account of the generally higher connectivity of active domains). **d**, Bar graph depicting mouse autosomes ordered by size with number of interacting chromosomes per single cell (black circles depict the distribution over individual cells). Mean number of interacting chromosomes changes modestly (30%) with chromosome size, suggesting a highly organized territory structure with surface that is not scaling with chromosome length.