

RESEARCH

Open Access



Comparison of Hi-C results using in-solution versus in-nucleus ligation

Takashi Nagano^{*†}, Csilla Várnai[†], Stefan Schoenfelder, Biola-Maria Javierre, Steven W. Wingett and Peter Fraser^{*}

Abstract

Background: Chromosome conformation capture and various derivative methods such as 4C, 5C and Hi-C have emerged as standard tools to analyze the three-dimensional organization of the genome in the nucleus. These methods employ ligation of diluted cross-linked chromatin complexes, intended to favor proximity-dependent, intra-complex ligation. During development of single-cell Hi-C, we devised an alternative Hi-C protocol with ligation in preserved nuclei rather than in solution. Here we directly compare Hi-C methods employing in-nucleus ligation with the standard in-solution ligation.

Results: We show in-nucleus ligation results in consistently lower levels of inter-chromosomal contacts. Through chromatin mixing experiments we show that a significantly large fraction of inter-chromosomal contacts are the result of spurious ligation events formed during in-solution ligation. In-nucleus ligation significantly reduces this source of experimental noise, and results in improved reproducibility between replicates. We also find that in-nucleus ligation eliminates restriction fragment length bias found with in-solution ligation. These improvements result in greater reproducibility of long-range intra-chromosomal and inter-chromosomal contacts, as well as enhanced detection of structural features such as topologically associated domain boundaries.

Conclusions: We conclude that in-nucleus ligation captures chromatin interactions more consistently over a wider range of distances, and significantly reduces both experimental noise and bias. In-nucleus ligation creates higher quality Hi-C libraries while simplifying the experimental procedure. We suggest that the entire range of 3C applications are likely to show similar benefits from in-nucleus ligation.

Background

Chromosome conformation capture (3C) and its various derivatives such as 4C (circularized chromosome conformation capture), 5C (carbon-copy chromosome conformation capture) and Hi-C have emerged as standard tools to analyze the three-dimensional organization of the genome in the nucleus [1, 2]. These methods have been extensively used for addressing various biological questions and subject to further technical developments [3–8], contributing substantially to our understanding of nuclear genome organization. All these methods depend on a simple principle of proximity-dependent ligation where DNA ends in cross-linked, restriction-digested, solubilized chromatin complexes are subjected to re-ligation in dilute solution intended to favor stochastic

re-ligation of nearby DNA fragment ends in the same complex [9]. DNA fragment pairs that are ligated are the direct readout and form the basis for 3C-based assay results. Frequent ligation between two fragments is interpreted to indicate that those fragments were in spatial proximity in vivo in a significant proportion of cells at the time of cross-linking, thus providing clues to the three-dimensional organization of the genome. Therefore, proximity-dependent ligation is a critical step in 3C and all of its derivative methods such as 4C, 5C and Hi-C.

It has been regarded as essential to extensively dilute the solubilized cross-linked chromatin prior to ligation to prevent non-specific ligation due to chance inter-molecular collisions. However, during our recent development of single-cell Hi-C [10] we modified the original Hi-C procedure and carried out the ligation step within preserved nuclei. Although the chromatin is not physically diluted when the ligation takes place in this modified

* Correspondence: takashi.nagano@babraham.ac.uk; peter.fraser@babraham.ac.uk

[†]Equal contributors

Nuclear Dynamics Programme, The Babraham Institute, Cambridge CB22 3AT, UK

procedure, we confirmed a high correlation between the results of the original “in-solution ligation” and our modified “in-nucleus ligation” procedures [10]. This is consistent with Comet et al. [11], who found that dilution prior to ligation is not essential to observe characteristic 3C profiles, and Gavrilov et al. [12], who found that most of the chromatin remains insoluble in diluted 3C samples and that the bulk of the 3C signals come from chromatin ligation in this insoluble fraction.

We observed that Hi-C coverage appeared to be more uniform in single-cell Hi-C, suggesting that in-nucleus ligation may actually improve Hi-C results [10]. In-nucleus ligation Hi-C has also been employed by Sofueva et al. [13] and Rao et al. [14]. In particular, Rao et al. reported interaction maps at higher resolution after deep sequencing than previous in-solution ligation methods, further suggesting that in-nucleus ligation may lead to improved results. Here we extensively compare the two ligation methods side by side, and find that in-nucleus ligation provides more consistent ligation frequency over the full range of genomic distances, and produces data with significantly less bias and significantly less technical noise.

Results

An overview of the Hi-C library method employing either in-solution ligation or in-nucleus ligation is shown in Additional file 1. In-solution Hi-C ligation is essentially as described by Lieberman-Aiden et al. [15] and includes a sodium dodecyl sulfate (SDS) treatment to lyse nuclei and solubilize chromatin complexes prior to ligation. The in-nucleus ligation procedure omits the SDS/lysis step and executes ligation in preserved nuclei.

To compare the two methods, we split cell pellets into different aliquots and carried out the Hi-C procedure in

parallel; fixation, restriction digestion and biotin fill-in, up until the ligation step. In half the samples we lysed the cells for in-solution ligation, and with the other half we performed in-nucleus ligation. Downstream Hi-C library preparation steps after the ligation were identical for both aliquots of cells. We created Hi-C libraries in this way from two biological replicates each of mouse foetal liver cells (mouse-1 and mouse-2) and human embryonic stem (ES) cells (human-1 and human-2). We also sequenced a random ligation library prepared by reversal of the cross-links and purification of the DNA prior to ligation.

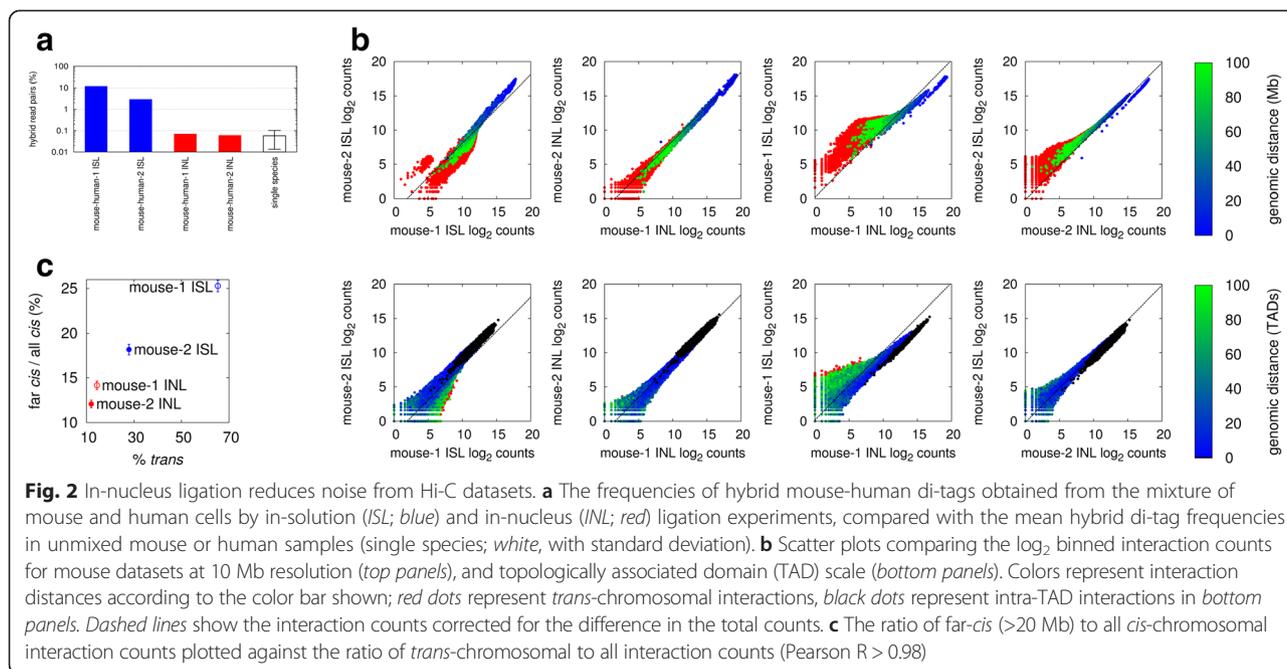
In-nucleus ligation reduces noise

After paired-end sequencing of the Hi-C libraries, the resulting FASTQ files were mapped against either the mm9 or hg19 genome assemblies using HiCUP [16], a Hi-C bioinformatics pipeline for aligning Hi-C reads and removing commonly encountered experimental artefacts. Table 1 shows the numbers of total di-tags and mapped di-tags for each library along with the breakdown of unique di-tags including the percentages of intra-chromosomal (*cis*) and inter-chromosomal (*trans*) di-tags for each dataset. The most obvious initial observation is that the in-nucleus datasets have consistently, markedly lower rates of *trans*-chromosomal interactions (10–14 %) compared with their corresponding in-solution ligation datasets (26–65 %). The relatively high percentage of *trans*-chromosomal contacts from in-solution ligation is consistent with several previously published in-solution ligation Hi-C datasets [3, 4, 10, 13–15, 17–20] (Fig. 1). As expected, the random ligation dataset had greater than 90 % *trans*-chromosomal contacts (Table 1 and Fig. 1). We also compared published data from tethered conformation capture (TCC) [4] and found the percentage of

Table 1 Datasets in this study

Dataset name	Total di-tags	Mapped di-tags	Percent _{re-ligation}	Unique di-tags	Percent _{cis}	Percent _{trans}	Percent _{hybrid}
Mouse-1 ISL	255,402,369	173,621,670	2.17	123,691,456	34.7	65.3	0.02
Mouse-2 ISL	131,079,448	83,009,025	3.85	32,939,021	72.1	27.9	0.05
Human-1 ISL	31,550,241	19,444,276	5.61	11,128,672	41.3	58.7	0.08
Human-2 ISL	27,518,246	17,060,401	7.79	6,994,931	74.5	25.5	0.03
Mouse-human-1 ISL	40,398,555	22,667,233	3.16	17,137,291	38.2	61.8	11.9
Mouse-human-2 ISL	2,928,112	1,405,001	3.84	488,340	72.9	27.1	2.9
Mouse-1 INL	242,699,915	167,175,980	3.49	108,733,338	85.6	14.4	0.16
Mouse-2 INL	134,813,901	85,804,344	6.80	37,259,526	88.0	12.0	0.02
Human-1 INL	34,860,295	22,113,388	10.40	11,457,531	85.8	14.2	0.07
Human-2 INL	33,500,802	20,280,022	14.03	7,777,001	90.0	10.0	0.03
Mouse-human-1 INL	33,762,441	19,318,810	4.16	11,533,994	85.6	14.4	0.06
Mouse-human-2 INL	2,920,332	1,423,890	7.53	573,366	88.5	11.5	0.07
Mouse RL	36,268,145	25,581,886	9.25	4,920,447	7.1	92.9	-

INL in-nucleus ligation, ISL in-solution ligation, RL random ligation



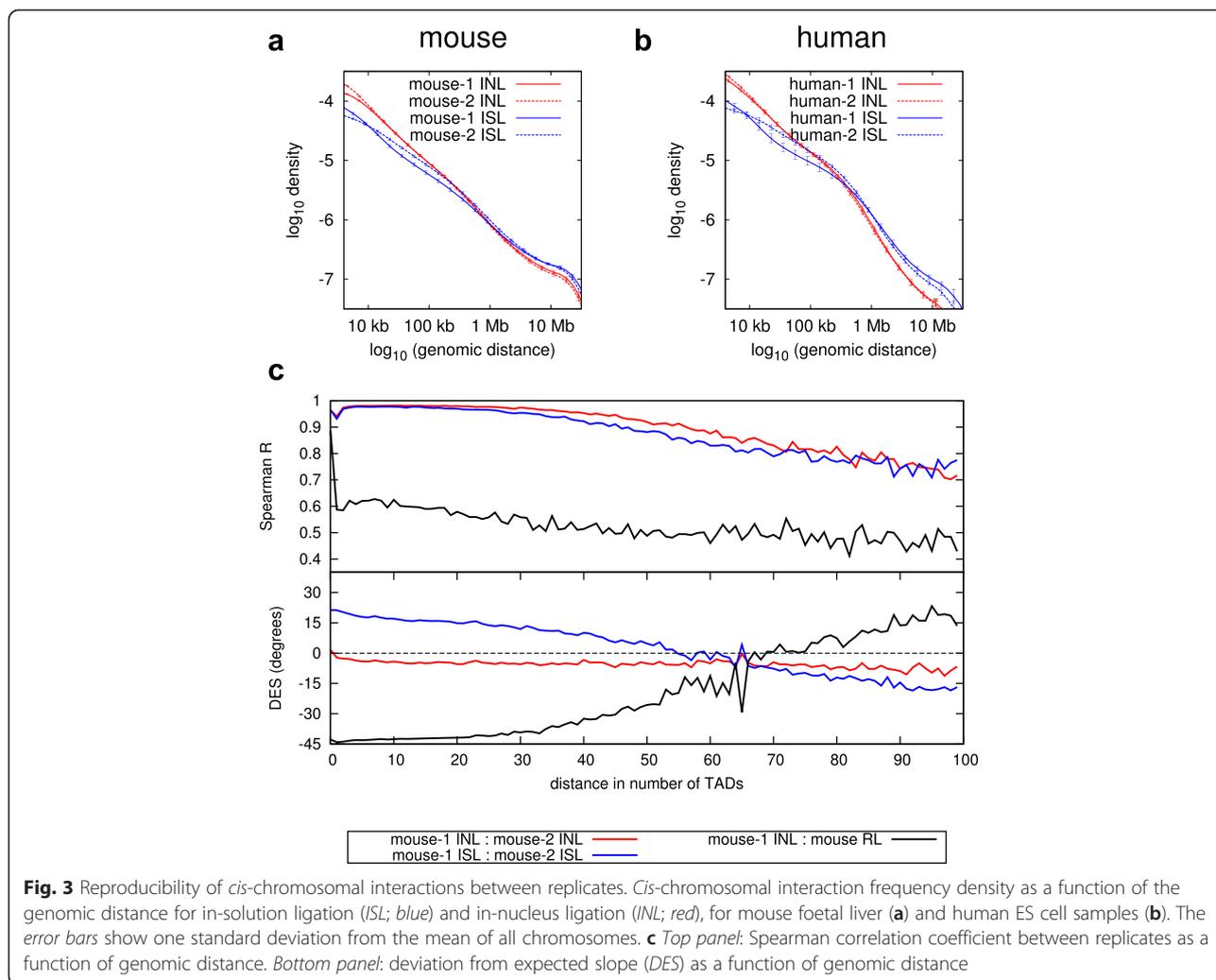
If the increased *trans*-chromosomal interactions generated during in-solution ligation Hi-C are caused by random ligations, far *cis*-chromosomal interactions should also contain a significant amount of noise. To test if this is the case, we compared the binned interaction counts, and topologically associated domain (TAD; see below) level interaction counts, in the raw Hi-C data between the mouse in-solution and in-nucleus ligation datasets. We used these datasets because they had the highest coverage. With increasing genomic distance we found a significant increase in the *cis*-chromosomal interaction log counts for the in-solution ligation datasets compared with the in-nucleus datasets in both the binned and TAD level analyses (Fig. 2b). This diverged from what would be expected purely from the ratio of the total number of interactions of the two datasets. We also observed this trend in comparisons of the human datasets (data not shown). Further inspection revealed a strong correlation between the ratios of *trans*-chromosomal to all interactions, and the ratios of far *cis*- (>10 Mb) and all *cis*-chromosomal interactions (Pearson R > 0.98 for the mean; Fig. 2c). Thus, in-nucleus ligation produces significantly fewer long-range *cis*-chromosomal di-tags (>10 Mb) compared with in-solution ligation, which is consistent with the hypothesis of reduced random ligation noise during in-nucleus ligation.

In-nucleus ligation improves reproducibility

The results shown in Fig. 2b suggest that in-nucleus ligation produces data that are more reproducible between replicates. To examine this in greater detail we compared the distribution of all *cis*-chromosomal interactions, often referred to as the powerlaw curves. We

observed that the curves for in-nucleus ligation-derived datasets are significantly more reproducible between replicates compared with those of in-solution ligation-derived datasets (Fig. 3). Notably, the in-nucleus replicates are highly reproducible across the entire range, from 10 kb to 100 Mb distance, whereas the in-solution replicates show significant divergence over broad ranges. As expected from our analyses shown in Fig. 2, the in-nucleus ligation generates significantly fewer *cis*-chromosomal di-tags over long distances (>10 Mb) compared with in-solution ligation, and also showed an increase in close *cis* di-tags. In fact, the powerlaw curves corresponding to in-nucleus ligation maintain a more uniform slope over the entire range of distances.

We also assessed reproducibility between replicates by calculating the Spearman correlation coefficient (R), and the deviation from the expected slope (DES) (see “Materials and methods”). In a perfectly reproducible experiment we would expect the Spearman correlation coefficient to be R = 1 and the deviation from the expected slope to be DES = 0. As can be seen in Fig. 3c the Spearman coefficient by itself can be misleading since it suggests that the random ligation dataset correlates well with the other Hi-C datasets. However, considering the Spearman correlation together with the DES provides a better comparison. We note that both the in-nucleus and in-solution datasets have remarkably good Spearman correlation coefficients, suggesting high reproducibility at multiple length scales. However, the DES calculation shows that the in-solution ligation comparison deviates substantially from the expected slope, whereas the in-nucleus ligation comparison deviates only marginally,



indicating improved reproducibility. Thus, in combination with the results shown in Fig. 2 we conclude that in-nucleus ligation provides more consistent results with more uniform coverage of *cis*-chromosomal interactions compared with in-solution Hi-C.

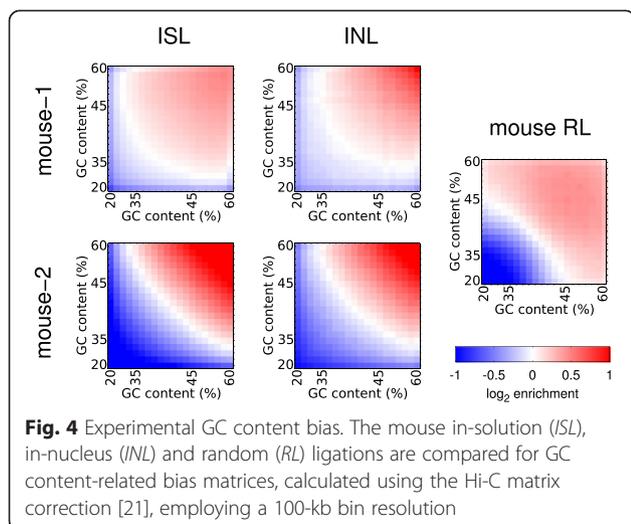
Reduced fragment length bias

Hi-C experiments are subject to systematic experimental biases, such as restriction fragment length and the GC content biases [21] that can affect the coverage of particular restriction fragments. These biases can be corrected by normalization of the Hi-C matrix containing the binned interaction data by inferring the bias contributions explicitly [21] or by a matrix balancing algorithm [22]. We found above that in-nucleus ligation produces more consistent *cis*-chromosomal interactions than in-solution ligation when considering unnormalized data. We tested if in-nucleus ligation improves the systematic biases discovered by Yaffe and Tanay [21]. We found little difference in the GC bias comparing in-

solution and in-nucleus ligation (Fig. 4). Since the GC content bias is mainly created during library amplification PCR [23], this bias would be expected using both methods. However, we found that in-nucleus ligation results in a remarkable reduction in fragment length bias (Fig. 5a, b), generating more uniform ligation between fragments regardless of length with the exception of extremely short (<100 bp) fragments. TCC appears more uniform compared with in-solution ligation, but it still exhibits a strong bias between short and long fragment lengths (Fig. 5c). We conclude that in-nucleus ligation effectively removes restriction fragment length bias, producing more consistent results between replicates.

Improved reproducibility of normalized Hi-C data

To investigate the reproducibility in close *cis*-, far *cis*- and *trans*-chromosomal di-tags in the normalized data, we plotted the normalized Hi-C matrices for chromosome 9 with the first principle component indicating the A and B compartments defined by Lieberman-Aiden



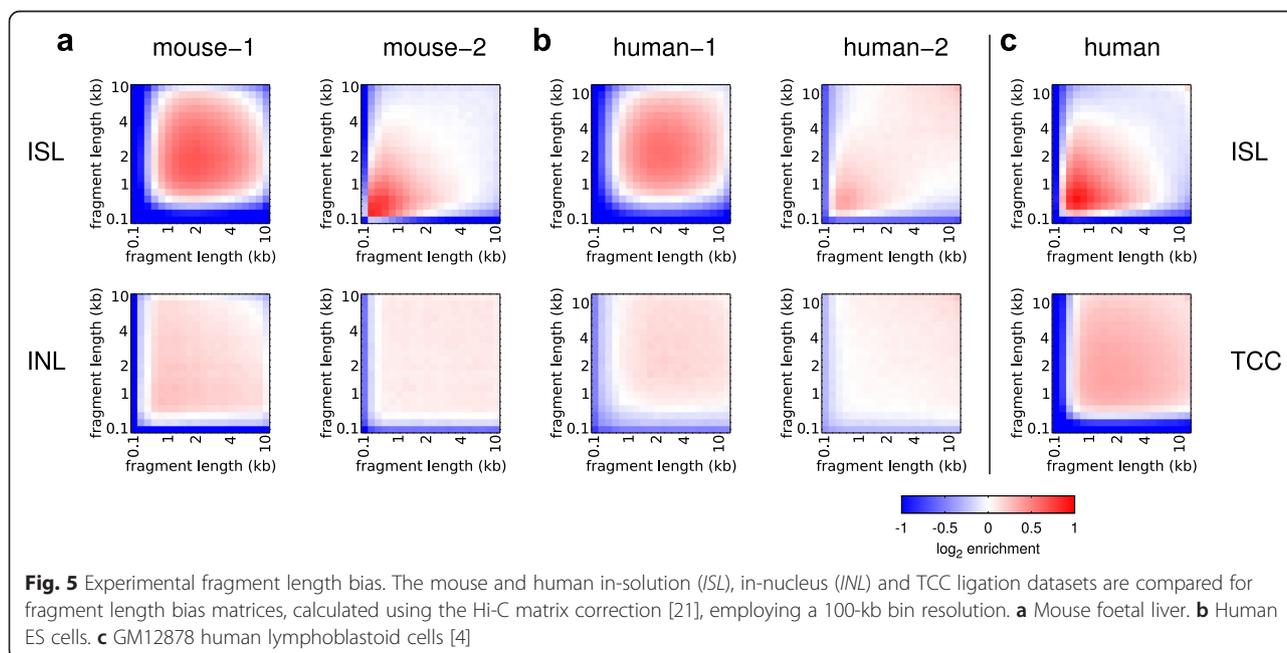
et al. [15] (Fig. 6). We found that the matrices were similar and the compartments were identical at 1-Mb resolution; however, we noted that the in-nucleus ligation matrices appeared sharper, which is particularly obvious at long range. To investigate this further, we plotted the coverage-corrected (Fig. 7a–f), and coverage- and distance-corrected (Fig. 7g–l) Hi-C matrix element values from the different datasets against each other, as in Fig. 2. All show high reproducibility for close *cis*-chromosomal di-tags (blue dots) whereas far *cis*- and *trans*-chromosomal di-tags (green and red dots) show high reproducibility only

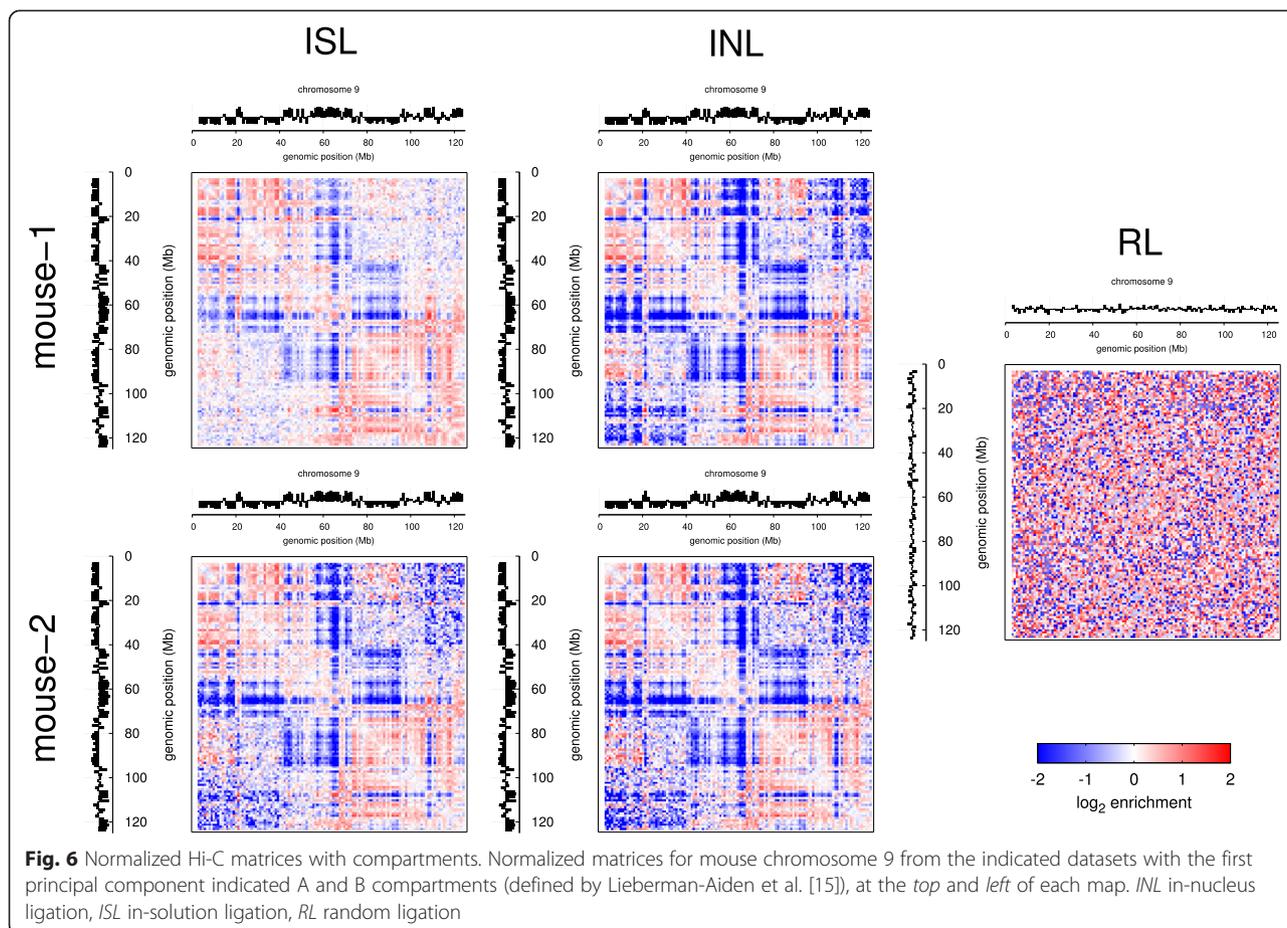
with in-nucleus ligation (Fig. 7a, g). In contrast, the in-solution ligation datasets display poor reproducibility for far *cis*- and *trans*-chromosomal di-tags (Fig. 7b, h). We observed similar effects when comparing the raw interaction counts (Fig. 2b), although it is more pronounced after correction for technical bias.

Comparison of the Hi-C data with random ligation data shows a clear shift in the scatter plot toward the horizontal axis, away from the marked diagonal (Fig. 7c, d, i, j). This trend is due to the increased noise in the random ligation. When comparing in-solution with in-nucleus datasets we also observed a similar shift away from the diagonal, suggesting that the in-solution ligation datasets have increased noise (Fig. 7e, f, k, l). In fact, this shift away from the diagonal is already apparent when comparing the two in-solution datasets (Fig. 7b, h). Together with the observations shown in Figs. 1, 2 and 3, these results suggest that in-nucleus ligation results in reduced noise and increased reproducibility of the normalized data.

Sharper structural features

Hi-C results provide information on chromatin organization into TADs [18, 24, 25]. TAD boundaries are defined by the observation that *cis*-chromosomal interactions within each TAD are more abundant over those connecting adjacent TADs. Our finding that in-nucleus ligation improves Hi-C data quality by reducing noise and bias prompted us to test if in-nucleus ligation datasets define these structural features more clearly. We identified TAD boundaries by calculating





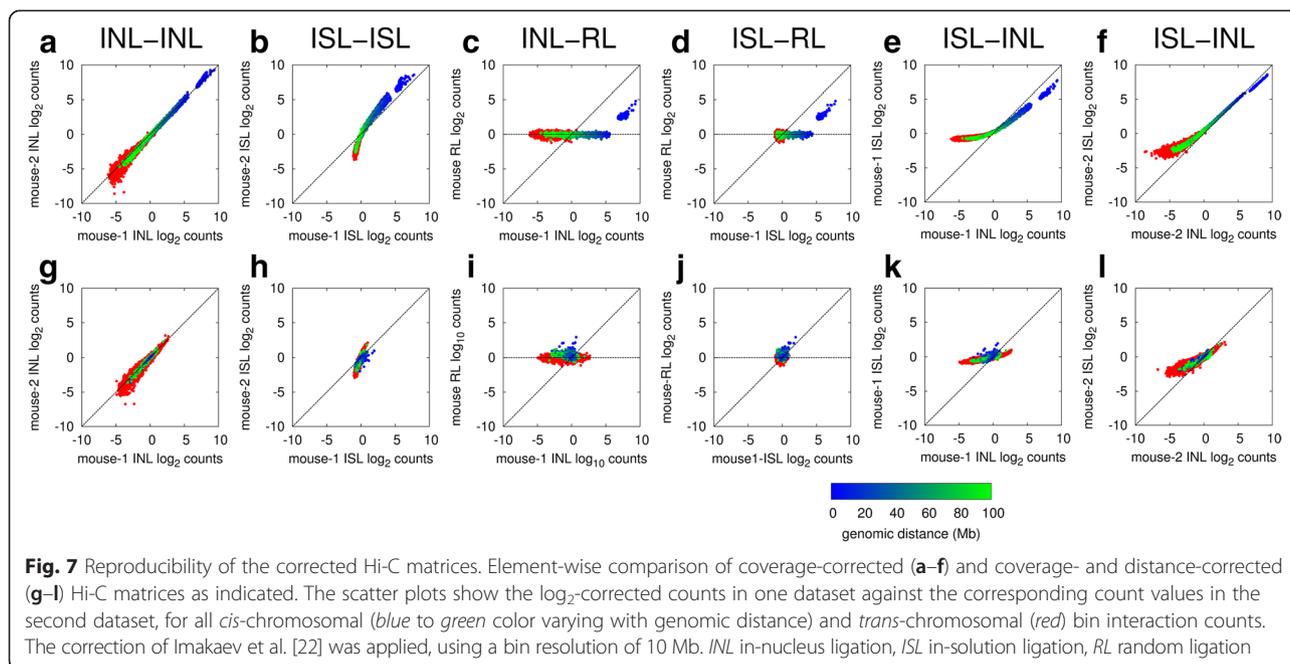
the directionality index of Hi-C interactions in the two in-nucleus mouse foetal liver replicates and their corresponding in-solution datasets. We found that approximately 60 % of the TAD boundaries previously identified by Dixon et al. [18] in mouse ES cells were within 100 kb of the boundaries we identified with the in-solution ligation datasets, whereas 70 % were within 100 kb of our identified in-nucleus ligation boundaries. We identified 2448 TAD boundaries that were consistent between the two in-nucleus datasets. In contrast, we found 1334 TAD boundaries that were consistent in both in-solution datasets. We selected 547 TAD boundaries that were defined by all four datasets and analyzed the distribution of di-tags and interaction directionality indices around these boundaries (Fig. 8a). The in-nucleus ligation results consistently show stronger depletion of interactions across the selected boundaries, demonstrating a more robust recognition of these structural features. The boundaries found only by in-solution ligation were weak boundaries, very close to the detection threshold (Fig. 8b). Although these boundaries were not detected by both in-nucleus replicates, on average they appear equally apparent with in-nucleus ligation. As expected, the boundaries found only by in-nucleus ligation Hi-C did not, on average,

reach the detection threshold with in-solution ligation (Fig. 8c). These results are in agreement with the greater number of boundaries identified consistently by in-nucleus ligation (2448) compared with in-solution ligation (1334). We conclude that in-nucleus ligation is capable of highlighting domain boundaries and potentially other structural features more reproducibly.

Discussion

We have shown that in-nucleus ligation results in significantly improved Hi-C data in many aspects. We found a significant reduction in technical noise arising from spurious ligation events, and increased reproducibility between replicates for *trans*-chromosomal and *cis*-chromosomal interactions at all distances. In addition, in-nucleus ligation effectively eliminates the fragment length bias normally found in conventional Hi-C data. All of these improvements appear to lead to cleaner data, allowing for a more robust detection of TAD boundaries, as an example.

We think many of these aspects are related to one another. For example, the reduced technical noise coming from spurious ligations could be because the ligatable fragment ends are physically more constrained



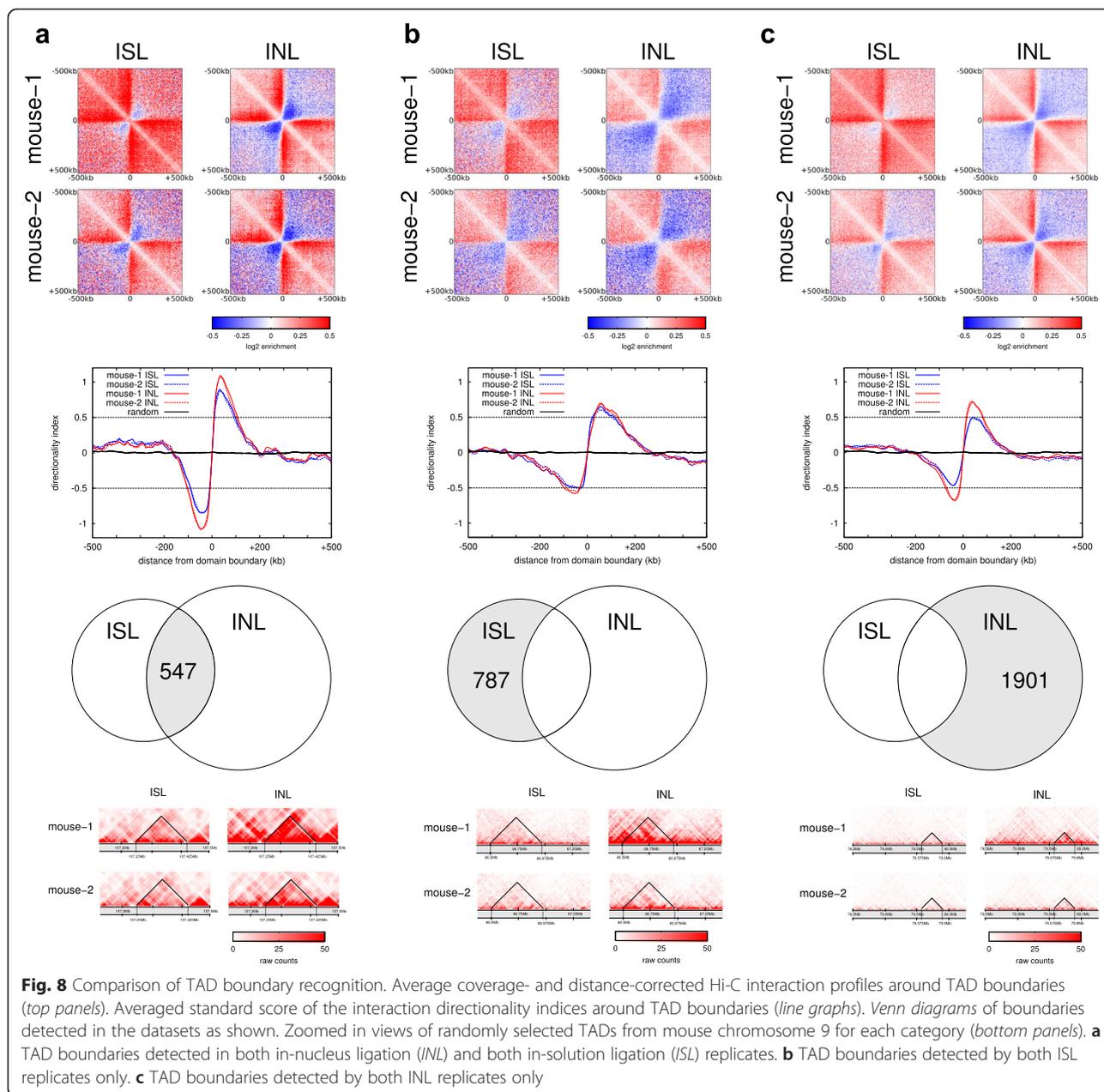
within the preserved nuclei, and thus only fragment ends in close nuclear proximity are capable of ligation. We have shown that a large amount of measurable, hybrid di-tags appear in mixing experiments using in-solution ligation, demonstrating that conventional in-solution ligation in a single sample will result in a significant amount of un-measurable spurious di-tags in *cis* and *trans*. Since these artefactual events will much more likely occur between different chromosomes, distal regions from the same chromosome or indeed between complexes from different cells, eliminating these spurious ligation events results in less noise and improved reproducibility for *trans*-chromosomal and far *cis*-chromosomal interactions. Since spurious ligations are un-measurable in a single sample, we suggest that an indicator of Hi-C library quality is a reduced percentage of *trans* di-tags.

We suggest that the results point to the benefits of a nuclear microenvironment during the ligation process. Fragment ends undergoing in-nucleus ligation may have reduced mobility compared with in-solution ligation due to molecular crowding in the nucleus. Their potential movements during the ligation process would be more restricted within a confined space, thereby increasing the chances of their ligation to spatially proximal fragment ends (Additional file 3a). In this scenario, fragments would not necessarily need to be directly cross-linked to each other to be ligated, but they simply need to be in spatial nuclear proximity. Anything that relaxes the spatial constraints between fragment ends will potentially compromise proximity-dependent ligation and result in greater noise (Additional file 3b). For example,

the harsh solubilization process (65 °C in the presence of SDS) prior to in-solution and TCC ligation would be expected to initiate cross-link reversal and liberate cross-linked complexes and small fragments, taking them out of their nuclear context. During ligation in solution, fragment ends would have greater mobility, potentially leading to a fragment length bias. During TCC ligation soluble complexes are immobilized, perhaps leading to a reduced incidence of spurious ligation (and reduced *trans* ratio), but the relaxation of spatial constraints due to nuclear disruption may lead to greater fragment end mobility, which could, for example, lead to a fragment length bias. In contrast, fragment length bias is not observed with in-nucleus ligation (Fig. 5), presumably because fragment end mobility is restricted to a confined nuclear space. This probably explains why we consistently observed more re-ligation for in-nucleus replicates (Table 1; $P = 0.01$, t-test).

Conclusions

We conclude that in-nucleus Hi-C ligation offers significant improvements over conventional in-solution Hi-C. In-nucleus ligation yields cleaner results with less technical noise characterized by lower ratios of *trans* di-tags. In-nucleus ligation also provides greater reproducibility between replicates at all distance scales, and effectively eliminates fragment length bias. These improvements lead to greater power to discern structural features such as TAD boundaries. Our results suggest that in-nucleus ligation will provide improved results for all experiments employing 3C-based techniques that rely on proximity-dependent ligation (3C, 4C, 5C and Hi-C).



Materials and methods

Cells

H9 (WA09; WiCell) human ES cells were maintained using Pluripro fully defined media and matrix (Cell Guidance Systems). Approximately 50 million cells (at passage 56) were harvested with Accutase (Life Technologies), suspended in Pluripro media and directly processed for fixation.

Mouse foetal livers were dissected from C57BL/6 mouse embryos at day 14.5 (E14.5) of development. Foetal liver cells were suspended in DMEM (Dulbecco's modified Eagle minimal essential medium; Life Technologies) supplemented with 10 % foetal bovine serum,

filtered through a cell strainer (70 μm) and directly fixed by addition of formaldehyde.

Hi-C

Except for the ligation step, Hi-C was performed essentially as described in Lieberman-Aiden et al. [15], with some modifications.

Thirty to 50 million cells were fixed in 2 % formaldehyde for 10 min, quenched with 0.125 M glycine, spun down (400 × g, 5 min) and washed once with phosphate-buffered saline. The cells were incubated in 50 ml permeabilization buffer (10 mM Tris-HCl pH 8, 10 mM NaCl, 0.2 % Igepal CA-630, Complete EDTA-free

protease inhibitor cocktail [Roche]) for 30 min on ice with occasional agitation, spun down ($650 \times g$, 5 min, $4^\circ C$), and the cell pellets were resuspended in $358 \mu l$ of $1.25 \times$ NEBuffer2 (NEB) per 5 million cell aliquot. We added $11 \mu l$ of 10 % SDS to each aliquot, followed by an incubation at $37^\circ C$ for 60 min with continuous agitation (950 rpm). To quench the SDS, $75 \mu l$ of 10 % Triton X-100 was then added per aliquot, followed by an incubation at $37^\circ C$ for 60 min with continuous agitation (950 rpm). To digest chromatin, 1500 U of HindIII (NEB) was added per aliquot and incubated at $37^\circ C$ overnight with continuous agitation (950 rpm). After digestion, restriction sites were filled in with Klenow (NEB) in the presence of biotin-14-dATP (Life Technologies), dCTP, dGTP and dTTP (all $30 \mu M$) for 60 min at $37^\circ C$.

For in-solution ligation, $86 \mu l$ of 10 % SDS was added per aliquot and incubated at $65^\circ C$ for 30 min with continuous agitation (950 rpm), followed by addition of 7.61 ml of ligation mix ($745 \mu l$ of 10 % Triton X-100, $820 \mu l$ of $10 \times$ T4 DNA ligase reaction buffer [NEB], $82 \mu l$ of 10 mg/ml bovine serum albumin [NEB] and 5.965 ml water) per aliquot and incubation at $37^\circ C$ for 60 min with occasional agitation. For in-nucleus ligation, 7.61 ml of ligation mix ($820 \mu l$ of $10 \times$ T4 DNA ligase reaction buffer [NEB], $82 \mu l$ of 10 mg/ml bovine serum albumin [NEB] and 6.71 ml water) was added per aliquot (compared with the in-solution ligation, SDS addition and incubation at $65^\circ C$ were omitted). For the ligation reaction (both in-solution and in-nucleus variants), $50 \mu l$ of $1 U/\mu l$ T4 DNA ligase (Life Technologies) was added per aliquot, followed by incubation at $16^\circ C$ for 4 h.

The cross-links were reversed by adding $60 \mu l$ of 10 mg/ml proteinase K (Roche) per aliquot and incubating at $65^\circ C$ overnight. After overnight incubation, another $60 \mu l$ of proteinase K per aliquot was added, followed by incubation at $65^\circ C$ for an additional 2 h. RNA was removed by adding $12.5 \mu l$ of 10 mg/ml RNase A (Roche) per aliquot and incubating at $37^\circ C$ for 60 min. DNA was isolated by a phenol (Sigma) extraction, followed by a phenol/chloroform/isoamylalcohol (Sigma) extraction and standard ethanol precipitation. The precipitated DNA was washed three times with 70 % ethanol, and dissolved in $25 \mu l$ TE per aliquot. Subsequently, all aliquots were pooled and the Hi-C DNA was quantified (Quant-iT Pico Green, Life Technologies). Biotin was removed from non-ligated restriction fragment ends by incubating 30–40 μg of Hi-C library DNA with T4 DNA polymerase (NEB) for 4 h at $20^\circ C$ in the presence of dATP. After DNA purification (QIAquick PCR purification kit, Qiagen) and sonication (Covaris E220), the sonicated DNA was end-repaired with T4 DNA polymerase, T4 DNA polynucleotide

kinase, Klenow (all NEB) and dNTPs in $1 \times$ T4 DNA ligase reaction buffer (NEB). Double size selection of DNA was performed using AMPure XP beads (Beckman Coulter), before dATP-addition with Klenow exo^- (NEB). Biotin-marked ligation products were isolated with MyOne Streptavidin C1 Dynabeads (Life Technologies) in binding buffer (5 mM Tris pH8, 0.5 mM EDTA, 1 M NaCl) for 30 min at room temperature, followed by two washes in binding buffer, and one wash in $1 \times$ T4 DNA ligase reaction buffer (NEB). Paired-end (PE) adapters (Illumina) were ligated onto Hi-C ligation products bound to streptavidin beads for 2 h at room temperature (T4 DNA ligase in $1 \times$ T4 DNA ligase reaction buffer [NEB], slowly rotating). After washes in wash buffer (5 mM Tris, 0.5 mM EDTA, 1 M NaCl, 0.05 % Tween-20) and binding buffer, the DNA-bound beads were resuspended in NEBuffer 2. Bead-bound Hi-C DNA was amplified with 12 PCR amplification cycles using PE PCR 1.0 and PE PCR 2.0 primers (Illumina). The concentration and size distribution of Hi-C library DNA after PCR amplification was determined by Bioanalyzer profiles (Agilent Technologies) and quantitative PCR, and the Hi-C libraries were paired-end sequenced on Illumina Hi-Seq 1000 or MiSeq platforms.

Mapping and filtering

The FASTQ paired-end read data were mapped against the appropriate reference genome (hg19, mm9 or an hg19/mm9 combined genome) and then filtered to remove frequently encountered experimental artefacts using the HiCUP [16] analysis pipeline developed at the Babraham Institute. After the filtering step, we calculated the difference of the ratio of the number of invalid di-tags relative to the uniquely mapped di-tags between the in-nucleus ligation and in-solution ligation datasets. For each di-tag category, we performed a t-test with the null hypothesis that the mean of the differences is 0, that is, there is no difference arising from the ligation step.

Proportion of hybrid mouse-human di-tags in the hybrid samples

For the mouse-human hybrid samples, we calculated the expected proportion of hybrid mouse-human di-tags (p_{hybrid}) in the Hi-C library, assuming random ligation and that the enzymatic restriction was complete:

$$p_{hybrid} = \frac{2n_{fend}^{mouse} n_{fend}^{human}}{(n_{fend}^{mouse} + n_{fend}^{human})^2}$$

where n_{fend}^{mouse} is the number of mouse fragment ends (the number of mouse cells multiplied by twice the number

of HindIII fragments in the mouse genome, 823,379), and $n_{\text{fend}}^{\text{human}}$ is the number of human fragment ends (the number of human cells multiplied by twice the number of HindIII fragments in the human genome, 837,163). In a sample containing a 5:1 ratio of mouse:human cells, $p_{\text{hybrid}} = 0.281$.

Powerlaw curves

We plotted the frequency of *cis*-chromosomal interactions at various genomic distances. The frequency density was obtained by binning the unique *cis*-chromosomal Hi-C di-tags, using 50 bins of equal size on a \log_{10} genomic distance plot.

Bias calculation

We quantified the extent to which the fragment length and the GC content of the fragment ends affect the read coverage using the hicpipe software (version 0.93) [26] developed by Yaffe and Tanay [21]. For each HindIII restriction fragment end, we calculated the fragment length, the GC content of the last 200 bp of the fragment end, and the mappability of the fragment. For the di-tags we used a segment length threshold of 500 bp, that is, we filtered out any di-tags where the sum of the distances from the read positions to the fragment ends where the ligation occurred was greater than this threshold. The algorithm binned the fragment lengths into 20 equally sized bins according to increasing fragment length. In turn, a 20×20 interaction matrix of these fragment length bins was used to describe the interaction bias between any two fragment ends. Similarly, a 20×20 interaction matrix was constructed using the GC content of the fragment ends. By performing a maximum likelihood optimization using the *trans*-chromosomal data (at 100 kb, 500 kb, 1 Mb and 10 Mb bin resolutions), we obtained the 20×20 interaction bias matrices describing the fragment length bias and the GC content bias.

Normalization of matrices

We calculated the coverage-corrected Hi-C matrices and the coverage-and-distance-corrected Hi-C matrices using the HOMER software [27] employing the algorithm described by Imakaev et al. [22]. It was assumed that the coverage of each bin should be the same in bias-free data, and that the observed Hi-C counts were the true counts multiplied by a factorizable bias (the factorizable bias of two interacting bins was the product of the bias contribution of the two individual bins).

The bias contribution vector and the true interaction matrix were optimized using an iterative approach, starting with the mapped filtered Hi-C data from HiCUP [16]. We used 1 and 10 Mb bin resolutions, excluding bins with coverage less than 20 % of the mean bin coverage, or more than 4 standard deviations away from the mean bin coverage.

Identification of compartments

We identified the compartments by calculating the first (or, for human samples, the first two) eigenvector(s) of the bin interaction profile correlation matrix for each chromosome, using the HOMER software [27]. The first eigenvector (or, for the human samples, the eigenvector related to the compartmental pattern as opposed to the chromosome arms) was aligned to active histone modification marks. This was done by multiplying the eigenvector by -1 if the Pearson correlation coefficient of the eigenvector and the H3K4me3 histone modification mark ChIP-seq [19, 28] profile was negative. The magnitude of the correlation coefficient was typically around 0.7. Chromosome bins with positive values in the eigenvector were considered to be in the A compartment, and bins with negative values to be in the B compartment. For the human chromosome 4, there was no clear separation between the first and second eigenvector profiles, so reads on human chromosome 4 were omitted from further analyses.

Compartment interaction bias among mouse–human hybrid reads

For the hybrid mouse–human di-tags, we assessed if there were any compartment-dependent non-random interactions, for instance, if mouse compartment A formed interactions preferentially with human compartment A. We counted hybrid di-tags in which both reads mapped to either compartment A or compartment B. We performed Fisher's exact test on these counts.

Scatter plots and measures of matrix reproducibility

We calculated the Spearman correlation of all *cis*- and *trans*-chromosomal interactions between different Hi-C experiments, at a 10-Mb bin resolution, as well as at a TAD level, using TADs as variable sized bins. In addition, we plotted each binned interaction count in one dataset against the corresponding interaction count in a second dataset. We colored the points of the plot according to the genomic distance of the interacting bins.

We subdivided the bin interaction count data according to the genomic distance of the interacting bins, and performed a linear fit on each of these datasets ($y = ax + b$, where a is the slope and b is the intercept). For each distance, we then corrected the slope for the Hi-C library sizes ($a_{\text{corr}} = a C_x / C_y$ where C_x and C_y are the total counts in the libraries shown on the x and y axes). The DES was then the angle between the corrected slope and the $y = x$ line:

$$\text{DES} = \text{atan}(a_{\text{corr}}) - \text{atan}(1).$$

A perfectly reproducible experiment would result in $\text{DES} = 0$ and a Spearman correlation $R = 1$.

Calculation of TAD boundaries

We calculated TADs in our coverage-corrected Hi-C matrices using the Hi-C domain finding tool of the HOMER software [27]. The algorithm defined directionality indices (DIs) as described in [18], based on the ratio of upstream and downstream interaction counts. We quantified the number of upstream and downstream interactions within an interaction distance of 1 Mb, using 25-kb overlapping bins with a step size of 5 kb. Bins with coverage less than 15 % of the mean bin coverage or greater than 4 standard deviations above the mean were excluded. This resulted in DI values at an effective 5-kb resolution (at the centre of each 25-kb window), which were further smoothed using a running average over a ± 25 kb window. Domain boundaries were then called where the smoothed DI was at a local extremum and at least 0.5 standard deviations away from the mean. Using the domains identified by HOMER, we called consensus TAD boundaries for in-solution ligation and in-nucleus ligation datasets, by keeping only TAD boundaries (rounded to the closest genomic position using a 25-kb resolution).

Hi-C interactions around TAD boundaries

We plotted the interaction directionality profile around the TAD boundaries using the average of the standard scores of the un-smoothed DI values, as a function of distance from the domain boundary upstream or downstream. A random control included 9686 randomly selected genomic positions. In addition, we plotted the coverage- and distance-corrected Hi-C interaction profiles around the consensus TAD boundaries using HOMER [27] and 25-kb overlapping bins with a step size of 5 kb.

Availability of supporting data

The datasets supporting the results of this article are available in the Gene Expression Omnibus (GEO) repository under accession number [GEO:GSE70181] [29].

Additional files

Additional file 1: Schematics of in-solution ligation and in-nucleus ligation Hi-C experimental protocols. (PDF 746 kb)

Additional file 2: Assessing randomness of hybrid di-tags. (PDF 52 kb)

Additional file 3: In-nucleus versus in-solution ligations. (PDF 337 kb)

Abbreviations

3C: chromosome conformation capture; 4C: circularized chromosome conformation capture; 5C: carbon-copy chromosome conformation capture; bp: base pair; DES: deviation from the expected slope; DI: directionality index; ES: embryonic stem; INL: in-nucleus ligation; ISL: in-solution ligation; kb: kilobase; Mb: megabase; PCR: polymerase chain reaction; PE: paired-end; SDS: sodium dodecyl sulfate; TAD: topologically associated domain; TCC: tethered conformation capture.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TN and SS conceived the study and designed the experiments, SS and BMJ performed the experiments, CV and SWW analyzed the data, TN, CV and PF wrote the manuscript with inputs from all other authors, and PF directed the entire project. All authors read and approved the final manuscript.

Acknowledgements

We thank Peter Rugg-Gunn for providing human ES cells, Kristina Tabbada and Dimitra Zante for sequencing the Hi-C libraries, and Yaniv Lubling, Eitan Yaffe and Amos Tanay for help with data analyses. This work was supported by the European Research Council, Advanced Grant 111608 and the Biotechnology and Biological Sciences Research Council UK grant BBS/E/B/000C0405 to PF.

Received: 12 May 2015 Accepted: 12 August 2015

Published online: 26 August 2015

References

- Hakim O, Misteli T. SnapShot: chromosome conformation capture. *Cell*. 2012;148:1068.
- Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet*. 2013;14:390–403.
- Selvaraj S, Dixon JR, Bansal V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol*. 2013;31:1111–8.
- 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.
- Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res*. 2014;24:1854–68.
- De Vree PJP, de Wit E, Yilmaz M, van de Heijning M, Klous P, Versteegen MJ, et al. Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nat Biotechnol*. 2014;32:1019–25.
- Jäger R, Migliorini G, Henrion M, Kandaswamy R, Speedy HE, Heindl A, et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun*. 2015;6:6178.
- Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre BM, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res*. 2015;25:582–97.
- Dekker J. The three “C”s of chromosome conformation capture: controls, controls, controls. *Nat Methods*. 2006;3:17–21.
- Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*. 2013;502:59–64.
- Comet I, Schuettengruber B, Sexton T, Cavalli G. A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber. *Proc Natl Acad Sci U S A*. 2011;108:2294–9.
- Gavrilov AA, Gushchanskaya ES, Strelkova O, Zhironkina O, Kireev II, Iarovaia OV, et al. Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. *Nucleic Acids Res*. 2013;41:3563–75.
- Sofueva S, Yaffe E, Chan W-C, Georgopoulou D, Vietri Rudan M, Mira-Bontenbal H, et al. Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J*. 2013;32:3119–29.
- Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014;159:1665–80.
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragozcy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326:289–93.
- HiCUP (Hi-C User Pipeline). <http://www.bioinformatics.babraham.ac.uk/projects/hicup>.
- Zhang Y, McCord RP, Ho Y-J, Lajoie BR, Hildebrand DG, Simon AC, et al. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*. 2012;148:908–21.

18. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485:376–80.
19. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, et al. A map of the *cis*-regulatory sequences in the mouse genome. *Nature*. 2012;488:116–20.
20. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep*. 2015;10:1297–309.
21. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*. 2011;43:1059–65.
22. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*. 2012;9:999–1003.
23. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*. 2011;12:R18.
24. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. 2012;485:381–5.
25. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*. 2012;148:458–72.
26. Hi-C pipeline for bias correction. <http://www.wisdom.weizmann.ac.il/~eitany/hicpipe>.
27. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 2010;38:576–89.
28. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74.
29. Datasets deposited at Gene Expression Omnibus, GSE70181. <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=wtmfywywjrsh&bip&acc=GSE70181>.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

