Robust 4C-seq data analysis to screen for regulatory DNA interactions

Harmen J G van de Werken[1,2,6], Gilad Landan[3,4,6], Sjoerd J B Holwerda[1,2], Michael Hoichman[3,4], Petra Klous[1,2], Ran Chachik[3,4], Erik Splinter[1,2], Christian Valdes-Quezada[5], Yuva Öz[1,2], Britta A M Bouwman[1,2], Marjon J A M Verstegen[1,2], Elzo de Wit[1,2], Amos Tanay[3,4] & Wouter de Laat[1,2]

Regulatory DNA elements can control the expression of distant genes via physical interactions. Here we present a cost-effective methodology and computational analysis pipeline for robust characterization of the physical organization around selected promoters and other functional elements using chromosome conformation capture combined with high-throughput sequencing (4C-seq). Our approach can be multiplexed and routinely integrated with other functional genomics assays to facilitate physical characterization of gene regulation.

Recent systematic efforts to map chromatin features along chromosomes[1] have identified hundreds of thousands of putative regulatory sites in the human and mouse genomes. With an estimated 25,000 genes per respective genome, this suggests that multiple sites regulate each gene and that the great majority of regulatory interactions occur over long chromosomal distances. Understanding gene regulation in complex eukaryotic genomes is therefore dependent on detailed mapping of physical contacts between genomic elements such as promoters and enhancers. Such mapping has been revolutionized through the advent of chromosome conformation capture (3C) technology[2] and 3C-based methods[3]. However, existing strategies either provide low genome-wide resolution[4–6] or are designed for biased mapping of specific regulatory interactions[7]. A cost-effective and high-resolution methodology to identify and quantify interactions between selected genomic sites and unknown regulatory sequences is still unavailable, and this lack has prevented researchers from incorporating physical considerations into most studies of gene regulation. Here we present a modified version of 4C technology[8,9] to provide a solution to this problem.

Our modified high-resolution 4C-seq protocol (**Fig. 1a**) involves two rounds of DNA digestion with restriction enzymes having 4-bp specificity. After a cross-linking step to capture genomic interactions, primary enzyme digestion with a 4-bp rather than 6-bp cutter[10] increases the pool of fragment ends that can be analyzed by over tenfold. This greatly enhances the statistical power of the 4C analysis, enabling robust identification of specific interactions based on many ligation events rather than one or two ligation junctions. Subsequent ligation of cross-linked restriction fragments typically results in long (>2 kb; **Fig. 1b** and **Supplementary Fig. 1**) DNA concatemers containing multiple (often more than ten) restriction fragments, which presumably were all components of a single cross-linked chromatin aggregate. The size of these concatemers precludes efficient amplification and sequencing. We therefore include a second round of digestion (in contrast to published 4C strategies[11]) using a different 4-bp cutter. This greatly increases the complexity of the amplified library and the robustness of contact profiles, and it is crucial for the reproducibility of the method. After ligation-induced circularization of the short fragments, we perform inverse PCR using primers designed to target a primary restriction fragment of interest (the 'viewpoint fragment') and amplify its ligated partners. By using primers that include dangling Illumina adaptor sequences[10], the inverse-PCR products are immediately ready for sequencing, with the sequence read consisting of the forward inverse-PCR primer, the restriction site and the near end of the ligated fragment. Inverse PCR is applied to approximately 500,000 cells per experiment, theoretically interrogating 1 million ligation products per viewpoint. A typical experiment therefore requires high-throughput sequencing of no more than 1–2 million reads.

We developed an analysis pipeline for mapping and normalizing 4C-seq data (**Supplementary Fig. 2**) that uses two complementary strategies. We take advantage of the rich fragment pool (about 6–8 fragment ends per 1 kb) to generate statistically robust semiquantitative contact maps in the 10-kb to 1-Mb region surrounding the viewpoint; contact intensities span 3–4 orders of magnitude. To all regions more remote from the viewpoint, we apply a statistical enrichment approach (**Supplementary Fig. 2d**), using an estimated probabilistic background model to compute expected total coverage for fragment ends in genomic windows, and compare this expected coverage to the observed number of sequenced ligation products. In both strategies, we simultaneously analyze contacts at multiple scales[12], using the high-resolution restriction site grid to quantify contact intensities in genomic windows varying in size from as little as a few kilobases to as much as several megabases (Online Methods).

[1]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW), Utrecht, The Netherlands. [2]University Medical Center Utrecht, Utrecht, The Netherlands. [3]Department of Computer Science and Applied Mathematics, Weizmann Institute, Rehovot, Israel. [4]Department of Biological Regulation, Weizmann Institute, Rehovot, Israel. [5]Instituto de Fisiología Celular, Departamento de Genética Molecular, Universidad Nacional Autónoma de México, México Distrito Federal, Mexico. [6]These authors contributed equally to this work. Correspondence should be addressed to A.T. (amos.tanay@weizmann.ac.il) or W.d.L. (w.delaat@hubrecht.eu).

RECEIVED 1 JUNE; ACCEPTED 17 AUGUST; PUBLISHED ONLINE 9 SEPTEMBER 2012; CORRECTED ONLINE 21 SEPTEMBER 2012 (DETAILS ONLINE); DOI:10.1038/NMETH.2173

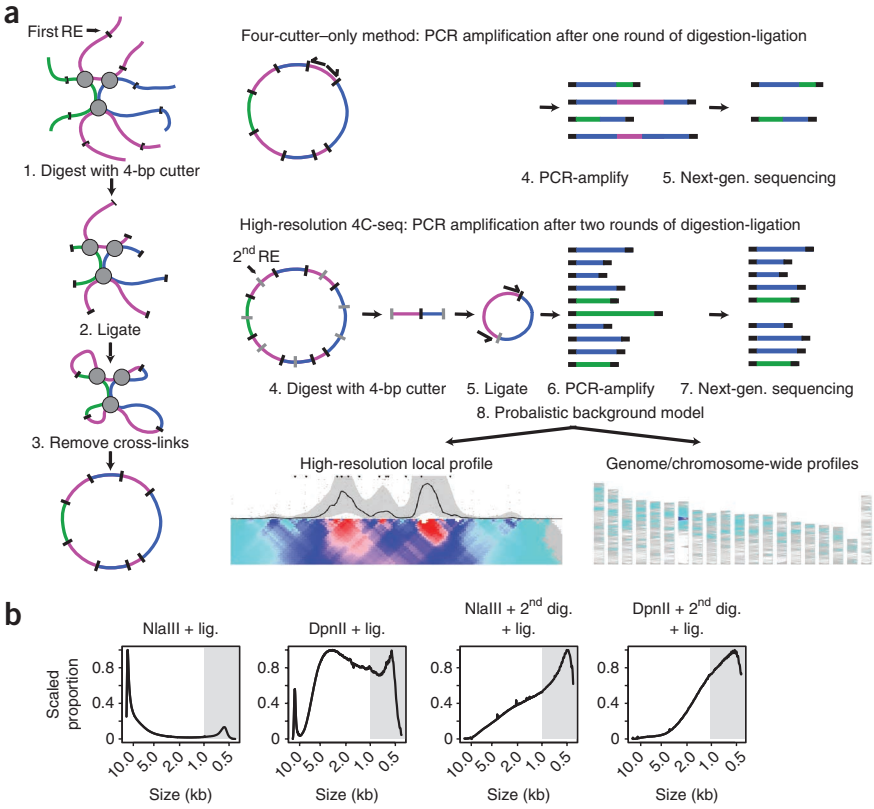

Figure 1 | 4C-seq. (**a**) Diagram showing the importance of including two rounds of digestion and ligation in high-resolution 4C-seq. RE, restriction enzyme; gen., generation. (**b**) Graphical representation of DNA size distribution after the first (left two graphs) and second (right two graphs) rounds of digestion and ligation. The shaded area indicates DNA fragment sizes that are PCR amplifiable and can be sequenced (see also **Supplementary Figs. 1** and **2**). Lig., ligation; dig., digestion.

To validate the approach, we first applied high-resolution 4C-seq to the well-characterized β-globin locus (**Fig. 2**). Chromosomal contacts within the β-globin cluster were previously mapped using 3C[13] and 3C–carbon copy (5C) technologies[14], and contacts with regions elsewhere in the genome were mapped with 4C technology[8]. The *Hbb-b1* gene is highly active in fetal liver cells, and its tissue-specific upregulation depends on a locus control region (LCR) composed of five hypersensitive sites (HS1–HS5) located ~30 kb upstream of the gene. We generated contact profiles using viewpoints next to the *Hbb-b1* gene and the HS2 element of the LCR in mouse fetal liver and fetal brain control. We obtained contact intensities for ~1,000 fragment ends in the 150-kb β-globin domain, compared to a few dozen in previously described 3C, 5C or low-resolution 4C data sets. This level of detail allows for the unbiased identification of the LCR (**Fig. 2a**). The data suggest that the gene-proximal side (HS1–HS2) is the most prominently interacting part of the LCR (**Supplementary Fig. 3**). A domain of preferred contacts exists that extends from the beginning of the LCR to the most distal active β-globin gene, *Hbb-b2*. A reciprocal profile generated using a viewpoint positioned at HS2 shows preferred contacts with the active *Hbb-b1* and *Hbb-b2* genes and demarcates the same domain in fetal liver. The defined domain and preferred contacts therein are absent in fetal brain, where the locus is inactive (**Fig. 2b**). CCCTC-binding factor (CTCF) sites flanking the β-globin locus were previously shown to interact with each other[15]. Our experiments confirm these interactions (**Supplementary Fig. 4**), identify a new interacting CTCF site (3′HS2) (**Supplementary Fig. 5**) and reveal that the single topological entity identified by 3C technology actually separates into two hierarchical structures: a chromatin loop that brings together CTCF sites flanking the locus and the regulatory interactions between the LCR and β-globin genes. We also applied our method to the α-globin locus[16–18]. The data confirm known long-range interactions and show that the α-globin locus adopts a defined domain topology only when active (**Supplementary Figs. 6** and **7**).

4C-seq can also be used to study interchromosomal and remote intrachromosomal interactions with high accuracy. Multiscale analysis quantifies the intensity of *cis* interactions between a viewpoint and chromosomal domains that vary in size between 10 kb and 5 Mb (**Supplementary Fig. 8**). We used a statistical model to estimate the probability of observing contacts between each fragment end and the viewpoint and computed the ratios between the expected and observed number of contacts in genomic windows of variable sizes. Comparison of the full chromosomal contact maps for different β-globin viewpoints that are within 50 kb of each other reveals highly consistent profiles in fetal liver. Remote (inter- or intra-) chromosomal contacts therefore appear to reflect the preferred neighborhood of larger chromosomal domains rather than specific interactions between regulatory sites. However, contact profiles are remarkably different in fetal brain cells. We observed similar cell type–dependent configurations for other tissue-specific genes (**Supplementary Fig. 8**) as well as for interchromosomal contacts formed by β-globin viewpoints (**Supplementary Fig. 9**). We note that, despite superior library complexity and resolution, the contact specificities for interchromosomal interactions are still modest, with the maximal enrichment detected as sixfold over the background (compared to 1,000-fold or more for local looping interactions). Still, the fact that genome-wide contact profiles are highly reproducible between different viewpoints in the same locus shows that the technique and its associated analysis also accurately identify contacts with the remainder of the genome.

We next examined two uncharacterized genes that are active in distinct tissues. *Oct4* is a key pluripotency gene expressed in embryonic stem cells (ESCs). A 4C-seq profile from a viewpoint positioned near the transcription start site (TSS) of *Oct4* (**Fig. 2c**) reveals specific contacts with a domain located approximately 17 kb upstream of the TSS, at a larger distance from the promoter than the known distal enhancer (−2 kb) and proximal enhancer (−1 kb)[19]. A luciferase reporter assay demonstrated that this segment drives increased reporter gene expression specifically in ESCs (**Fig. 2d**). Four other surrounding sequences

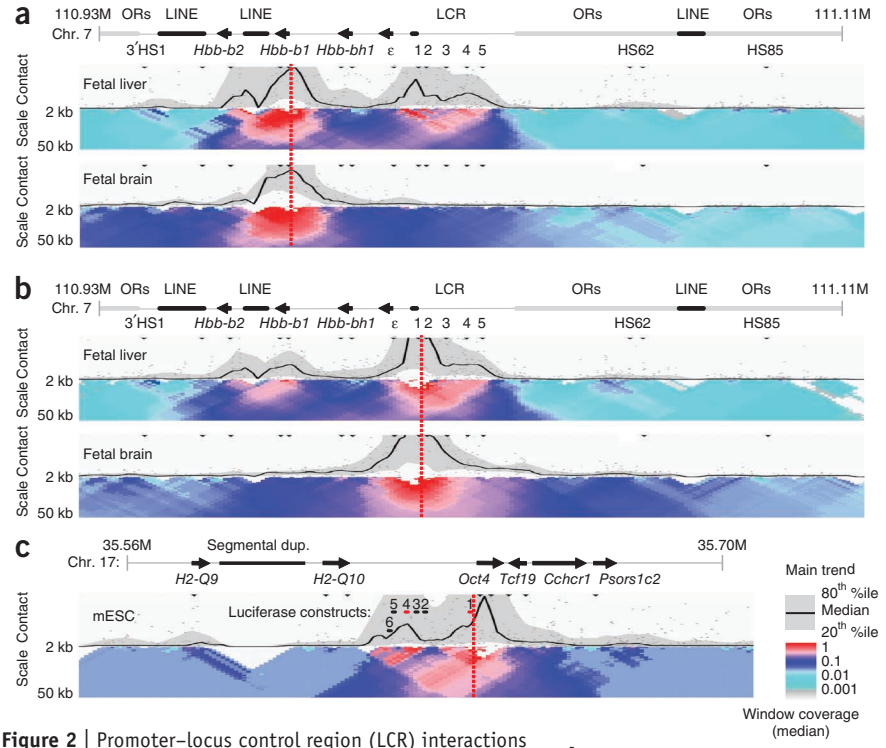

Figure 2 | Promoter–locus control region (LCR) interactions in the active β-globin cluster. (**a**) Contact profiles of the active (middle; fetal liver) and inactive (bottom; fetal brain) *Hbb-b1* gene, using a 5-kb window size in the main trend subpanel. Dashed red lines indicated viewpoint position. Chr., chromosome; ε, *Hbb-y* (εy); ORs, olfactory receptor genes; LINE, long interspersed nuclear element; #M, genomic coordinates in millions of base pairs. (**b**) Contact profiles for the hypersensitive site (HS) 2 element within the LCR in the active (middle; fetal liver) and inactive state (bottom; fetal brain), using a 5-kb window size in the main trend subpanel. (**c**) Contact profile of the *Oct4* promoter in mouse embryonic stem cells (mESCs), using a 7-kb window size in the main trend subpanel. Sequences capable of driving luciferase activity are marked red, and cloned sequences that are incapable are marked black. Dup., duplication. (**d**) Luciferase reporter assays demonstrating enhancer activity in ESCs of the upstream site found to interact with *Oct4*. Error bar indicates s.e.m.; $n = 9$ (independent experiments, each analyzed by triplicate PCR) for *Oct4* – 17 kb. For all other regions, $n = 2$ and individual data points are indicated with circles. $^*P \leq 0.05$ (one-tailed Student's *t*-test on log-transformed data).

without obvious contacts with the *Oct4* promoter did not show this activity in ESCs, despite enrichment for H3K4me1 (monomethylation of histone H3 lysine 4) at some of these segments (**Supplementary Fig. 10**).

Satb1 is a gene that is highly active in thymocytes. Using 4C-seq, we find that the *Satb1* TSS preferentially contacts a gene-poor chromosomal domain in thymocytes that spans over 600 kb upstream and includes multiple contact hotspots (**Supplementary Figs. 11a** and **12a–c**). The domain was identified reciprocally when assaying contacts from a remote putative contacting element 470 kb away from the gene. In fetal brain, where *Satb1* is over an order of magnitude less active (**Supplementary Figs. 12** and **13**), this interaction domain is completely absent, with only weak residual contact noticeable with an element 650 kb upstream (**Supplementary Fig. 12d**). The same is true in ESCs (**Supplementary Fig. 11a**) that express *Satb1* at even lower levels (nearly 3 orders of magnitude less than thymocyte expression levels). These results suggest that the large *Satb1* proximal domain acts as a regulatory scaffold that facilitates the high expression of *Satb1* in T cells. To examine the regulatory activity of the contact hotspots, we tested a series of distal sites in a luciferase reporter assay. Contacting elements at a distance of 253 kb, 470 kb and 649 kb from the TSS were found to substantially boost reporter gene expression in lymphoid cells (**Supplementary Fig. 11b**).

It is important to determine how an effective promoter viewpoint should be positioned relative to the TSS when screening for promoter-enhancer interactions. We therefore studied contact profiles derived from viewpoints positioned within a few kilobases around the *Satb1* and *Oct4* TSSs. The data show that chromosomal contacts around TSSs can be surprisingly position specific (**Supplementary Figs. 14** and **15**). For enhancer screening purposes, positioning the viewpoint immediately upstream of the TSS seems to be preferred. Controls using viewpoints further upstream and downstream of the TSS can be used to provide more data on the regional architecture of the TSS-enhancer domain.

Physical or topological domains around active genes were recently demonstrated globally in Hi-C experiments[4–6]. Effective mapping of functionally relevant contacts within these domains requires high-resolution strategies as presented here. Compared to (semi-)quantitative 3C, which is currently the method of choice to assay contacts between nearby regulatory sequences, the integrated 4C-seq strategy and pipeline are easier (in that only one primer pair and no control template are needed), much more robust (in the β-globin locus, we analyzed nearly 1,000 independent ligation events, compared to 15–20 junctions analyzed in a typical 3C experiment) and unbiased to pre-chosen genomic partners.

The 4-seq analysis pipeline, and a genome-wide 4C primer database, can be downloaded from http://compgenomics.weizmann.ac.il/tanay/?page_id=367/.

METHODS

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

Accession codes. 4C-seq data is available on GEO: GSE40420.

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

ACKNOWLEDGMENTS

We would like to thank R. Palstra, E. Yaffe and other members of our labs for help and input, G. Geeven for testing the 4C-seq pipeline and H.T. Timmers (Netherlands Proteomics Centre and University Medical Center Utrecht) and J.P.P. Meijerink (Erasmus Medical Center-Sophia Children's Hospital) for providing cells. This work was supported by grants from the Modelling Hepatocellular Carcinoma (MODHEP) Seventh Framework Program (FP7) collaborative project to W.d.L. and A.T., grants

from the Dutch Scientific Organization (NWO) (91204082 and 935170621), a European Research Council Starting Grant (209700, '4C') and InteGeR FP7 Marie Curie Initial Training Networks contract (no. PITN-GA-2007-214902) to W.d.L., and the Israeli Science Foundation integrated technologies grant to A.T.

AUTHOR CONTRIBUTIONS

H.J.G.v.d.W. designed, performed and analyzed experiments and wrote the manuscript; E.S. helped design experiments; G.L. helped design and analyze experiments and wrote the manuscript; P.K., S.J.B.H., B.A.M.B., M.J.A.M.V., Y.Ö. and C.V.-Q. performed experiments; M.H., R.C. and E.d.W. helped analyze experiments; and A.T. and W.d.L. designed experiments, supervised the project and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2173. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Shen, Y. *et al. Nature* **488**, 116–120 (2012).
2. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. *Science* **295**, 1306–1311 (2002).
3. de Wit, E. & de Laat, W. *Genes Dev.* **26**, 11–24 (2012).
4. Lieberman-Aiden, E. *et al. Science* **326**, 289–293 (2009).
5. Sexton, T. *et al. Cell* **148**, 458–472 (2012).
6. Dixon, J.R. *et al. Nature* **485**, 376–380 (2012).
7. Li, G. *et al. Cell* **148**, 84–98 (2012).
8. Simonis, M. *et al. Nat. Genet.* **38**, 1348–1354 (2006).
9. Zhao, Z. *et al. Nat. Genet.* **38**, 1341–1347 (2006).
10. Splinter, E. *et al. Genes Dev.* **25**, 1371–1383 (2011).
11. Lower, K.M. *et al. Proc. Natl. Acad. Sci. USA* **106**, 21771–21776 (2009).
12. de Wit, E., Braunschweig, U., Greil, F., Bussemaker, H.J. & van Steensel, B. *PLoS Genet.* **4**, e1000045 (2008).
13. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. & de Laat, W. *Mol. Cell* **10**, 1453–1465 (2002).
14. Dostie, J. *et al. Genome Res.* **16**, 1299–1309 (2006).
15. Splinter, E. *et al. Genes Dev.* **20**, 2349–2354 (2006).
16. Vernimmen, D., De Gobbi, M., Sloane-Stanley, J.A., Wood, W.G. & Higgs, D.R. *EMBO J.* **26**, 2041–2051 (2007).
17. Zhou, G.L. *et al. Mol. Cell. Biol.* **26**, 5096–5105 (2006).
18. Baù, D. *et al. Nat. Struct. Mol. Biol.* **18**, 107–114 (2011).
19. Yeom, Y.I. *et al. Development* **122**, 881–894 (1996).

ONLINE METHODS

General considerations with respect to modeling and interpreting 4C-seq experiments. The raw experimental readout from a single 4C-seq experiment consists of 500,000–3 million reads containing the ligation junctions between the viewpoint-fragment end (where the forward PCR primer is positioned) and any other primary restriction-fragment end. To map the latter, we developed a specialized 4C-seq genome mapping algorithm that controls for sequencing errors and nonunique sequences while considering the high coverage (100×–100,000×) of fragment ends that are proximal to the viewpoint fragment (see "Mapping and filtering sequence reads" below). The number of reads mapped by the algorithm to each fragment end defines the experiment's coverage profile and represents the intensity of contacts between each restriction fragment and the viewpoint fragment, combined with multiple stochastic and systematic noise factors. The challenge in 4C-seq analysis is to create a robust scheme to normalize these noise factors while enabling maximal resolution of the derived contact profiles.

Interchromosomal and remote intrachromosomal contacts are observed with small probability throughout the genome in an overall spatially uncorrelated fashion (**Supplementary Fig. 2a**). In such remote regions, the observation of multiple reads for the same fragment end is not more informative than a single read (**Supplementary Fig. 2a**,**b**) and can be assumed to represent stochastic amplification events. In support of this, we find that long-range and interchromosomal contact profiles from replicate experiments are reproducible at the level of larger genomic regions (>100 kb) but not at the level of the single fragment. This is similar to the behavior of Hi-C matrices that show a typical background contact probability of 0.001% between any two fragment ends[5,20]. Conclusions on remote contacts in such settings are therefore obtained by pooling together statistics from hundreds of fragment ends (for 4C) or dozens of thousands of fragment pairs (in Hi-C), such that a sufficient number of reads can be aggregated. In contrast, local contacts between the 4C viewpoint and fragments in its chromosomal vicinity (i.e., <1 Mb distance), are recovered with high probability, and the number of reads mapping to individual fragment ends in this region is highly informative (**Supplementary Fig. 2c**). The effective resolution of a 4C experiment therefore relies on the proper quantitative normalization of the coverage profile near the viewpoint.

The computational normalization and visualization of 4C-seq contact profiles is achieved using two complementary strategies. A quantitative approach to contact intensity is applied in the region surrounding the viewpoint, which generates a normalized contact profile that represents intensities within 3–4 orders of magnitude. A statistical-enrichment approach (subsequent sections and **Supplementary Fig. 2**) is applied to all other regions; this approach uses an estimated probabilistic background model to compute expected total 4C coverage for fragment ends in selected genomic windows and compares the expected coverage to the observed number of sequenced ligation products. In both strategies, analysis is done using an approach that simultaneously captures interactions at multiple scales, reminiscent of the previously published domainograms[12]. The high-resolution restriction site grid provides the opportunity to examine contact intensities in genomic windows varying in size from as little as a few kilobases to as much as several megabases.

To quantify contact strength, we analyze and visualize medians of normalized coverage (in the viewpoint's vicinity) or enrichment of observed versus expected number of reads (for the rest of the genome). Such quantities describe contact intensity rather than statistical significance (i.e., *P* value) of nonbackground behavior. This approach prevents the bias toward larger genomic windows (more data points) that is commonly introduced when using *P* values to visualize contact intensity. Instead, *P* values are used only when testing the robustness of the intensity statistics. This unbiased approach ensures that statistics at different genomic scales are directly comparable and interpretable.

Preparation of 4C-template. 4C templates were prepared essentially as described previously[21]. In brief, primary tissue was isolated, single cells suspensions were made, chromatin was cross-linked with 2% formaldehyde for 10 min at room temperature, nuclei were isolated and cross-linked DNA was digested with a primary restriction enzyme recognizing a 4-bp restriction site. This was followed by proximity ligation after which cross-links were removed. A secondary restriction enzyme digestion was performed with a 4-bp restriction enzyme recognizing a different sequence than the primary enzyme, followed again by proximity ligation. Typically, 200 ng of the resulting 4C template was used for the subsequent PCR reaction, of which 16 (total: 3.2 µg of 4C template) were pooled and purified for next-generation sequencing. The PCR products were purified using two columns per sample of the High Pure PCR Product Purification Kit (Roche cat. no. 11732676001). The kit separates the PCR products that are larger than 120 bp from the adaptor-containing primers (which are respectively ~75 nucleotides (nt) and ~40 nt in size). Similar results were obtained with products from a single PCR reaction (200-ng template).

4C-seq primer design. 4C primer pairs carry additional 5′ overhangs composed of the adaptor sequences (obtained from Illumina technical support) necessary for Illumina single-read sequencing (GA-II and Hi-seq 2000). The strategy therefore produces sequencing reads (36-mers, in this study) composed of the 4C primer sequence (20 nucleotides, specific to a given viewpoint) followed by 16 nucleotides that identify a captured sequence. The reading primer always hybridizes to, and ends at the 3′ side of, the first restriction recognition site. This design ensures analysis of only primary ligation events and provides sufficient sequence information to unambiguously identify most captured sequences (that is, the mappability of 16-mers directly adjacent to a given 4-bp site is 68%, using NlaIII and DpnII as the restriction enzyme combination). The nonreading primers, with sizes between 18 and 27 bp, were designed at a distance of ≤120 bp from the secondary restriction site.

PCR primers were designed taking into account the following additional rules. The viewpoint fragment preferably had a size of at least 500 bp to allow efficient cross-linking to other DNA fragments. The fragment end (the nucleotide sequences of the viewpoint fragment between the primary and secondary restriction site to which both 4C primers hybridize) was at least 300 bp and preferably more than 350 bp to allow efficient circularization during the second ligation step. Primer3 (ref. 22) was used to find the optimal primer pair for a given viewpoint fragment, with the following adaptations to the default settings: optimal

temperature of 55 °C, minimum of 45 °C and maximum of 65 °C; GC content between 35 and 65%. Primers were checked against the mouse genome with MegaBLAST[23] (settings -p 88.88 -W 12 -e 1 -F T), which requires primers on the reading side to be matched uniquely in the genome and primers on the nonreading side to have a maximum of three perfectly matching BLAST high-scoring segment pairs (HSP). Both primers were also required to have fewer than 30 HSPs with an identity of at least 88.89% (16 of 18 bp). **Supplementary Table 1** shows all the primers used in this study. Moreover, a genome-wide 4C primer database was constructed that implements all rules and can be downloaded from our website (http://compgenomics.weizmann.ac.il/tanay/?page_id=367/).

Mapping and filtering sequence reads. **Supplementary Table 2** describes all experiments included in this work and provides statistics on their read counts. At least 10–20 different 4C experiments (different viewpoints, or the same viewpoint but barcoded) were mixed and sequenced simultaneously in one Illumina GA-II or HiSeq 2000 lane. Sequence tags generated by the 4C-seq procedure are prefixed by the 4C reading primer that includes the restriction site sequences. We therefore separate multiplexed 4C-seq libraries according to the prefix and extract their suffixes for further processing. The algorithm for mapping of suffixes to the genome was designed given the following main considerations:

1. Valid suffixes should begin at a primary restriction site and continue with its downstream sequence (i.e., they should be mappable to one of the experiment's fragment ends).
2. The expected coverage profile is highly nonuniform, and fragment ends that are proximal to the viewpoint are likely to be covered dozens to thousands of times. The number of reads mapped to each fragment end represents significant information in the viewpoint region (<1 Mb), but very little information out of this region (**Supplementary Fig. 2a**,**b**).
3. Even though the sequencing error rate is low, ligation junctions occur thousands of times and may give rise to dozens of copies with particular mismatches. When such mismatches create variants that are mappable to the genome (by chance), substantial coverage of remote fragment ends may be incorrectly inferred.

A special-purpose mapping algorithm based on these considerations is described next. We assume sequence tags of length L_{tag} (read length minus the length of the primer and restriction site) are given and that base-calling probabilities are provided for each tag. We define the precision of each sequence tag as the product of estimated base-calling probabilities. The probability of any specific mismatch is computed by multiplying the base-error probabilities at the variable positions. The algorithm proceeds along the following steps:

1. Constructing a fragment end index: all restriction sites in the genome are identified, and the L_{tag} bp sequences both upstream and downstream are indexed using a hash table.
2. Computing interim coverage for well-separated fragment ends: a fragment end is classified as 'well separated' if all other fragment ends of size L_{tag} in the genome differ from it in at least two positions. We regard sequence tags with precision >0.9 that map perfectly to well-separated fragment ends as unambiguous and compute an interim coverage profile (denoted interim_j) for each such fragment end j by looking it up in the hash table and summing the total precision of tags mapped to it. We note that the precision threshold is adjustable and may need to be changed for longer reads with low base-calling probabilities toward the end of the read.
3. Computing fragment-end mapping weights: we define the mapping weight of each well-separated fragment end j as its interim coverage computed in the previous step. For all other fragment ends, the mapping weight equals the distance-weighted geometric mean of interim coverage on well-separated fragment ends in a window of size W around the fragment end i

$$\text{weight}(i) = \exp\left(\frac{1}{Z(W,i)}\sum_{d(i,j)<W}\left(1-\frac{d(i,j)}{W}\right)\log(\text{interim}_j)\right)$$

where

$$Z(W,i) = \sum_{d(i,j)<W}\left(1-\frac{d(i,j)}{W}\right)$$

and $d(i,j)$ is the genomic distance between fragment ends i and j, interim_j is the coverage of well-separated fragment end j and W is 200 kb or the minimal size for which $Z(W,i) \geq 6$ if $Z(200\text{ kb}, j) < 6$. This ensures that non–well-separated fragment ends that are extremely distant from well-separated fragment ends use a larger window and are weighted using a robust geometric mean.
4. Computing the coverage profile: given the mapping weights defined above, each read is distributed among its potential originating fragment ends according to the sequence and base-calling probabilities of the read as well as the mapping weights of fragment ends with the same or a similar sequence. Specifically, given a sequence tag s, the mapping vote for a fragment end i with sequence $\text{fes}(i)$ is computed as

$$\text{vote} = \frac{1}{Z}\Pr(\text{fes}(i)|s)\times\text{weight}(i)$$

where $\Pr(\text{fes}(i)|s)$ is the probability that the read s originated from the fragment end with sequence $\text{fes}(i)$, calculated using the read's base-calling errors. The normalization factor Z is computed by summing over all fragment ends with sequences that are not too far from the read sequence

$$Z = \sum_{\Pr(\text{fes}(i)|s)>s}\Pr(\text{fes}(i)|s)\times\text{weight}(i)$$

Here we used an ε value of 0.0001, but this can be modified according to read length and overall sequence quality, as it significantly affects the algorithm's run-time performance (by determining the number of fragment ends the algorithm must examine per read). Moreover, fragment ends with sequences that appear more than five times in the genome, and their

doi:10.1038/nmeth.2173

associated reads, are filtered out. The final coverage profile is defined by distributing each read among fragment ends according to the mapping votes.

In general, although the mapping algorithm takes full account of nonunique fragment ends (which, as described above, are resolved in a way that can also affect the expected coverage of unique fragment ends), the analysis included in the present work is based on coverage statistics for only unique fragment ends.

Our C++ implementation of the above mapping algorithm can complete mapping of a 2 million–read experiment within approximately 10 min on a single core of a Linux machine using up to 6 GB of RAM. We note that longer reads can further reduce mapping ambiguity and allow application of standard mapping algorithms for 4C-seq.

Construction of a background model for remote intra- and interchromosomal contacts. Several steps in the 4C-seq experimental protocol are prone to systematic biases that may influence the distribution of coverage inferred by the mapping algorithm described above. Of these, factors affecting ligation and amplification efficiency include the restriction fragment length, its GC content and the size of the 4C amplification product as determined by the linear genomic distance between the primary restriction site and the nearest secondary restriction site. Broad enrichment patterns (such as those presented in **Supplementary Figs. 8** and **9**) represent relatively mild contact preferences (two- to threefold enrichment) of large genomic windows. Because some of the potential sources of bias in coverage are distributed nonuniformly in the genome and may differ significantly between large genomic windows, it is particularly important to normalize raw coverage before studying global contact trends.

We define the fragment length associated with each fragment end as the distance between the two primary restriction sites forming the fragment. This length is binned into the ranges 0–50, 50–100, 100–200, 200–300, 300–400, 400–500 and >500 and is denoted by $\mathrm{fl}(i)$. We define the fragment end length as the distance between the primary restriction site and the nearest secondary restriction site, binned into the ranges 0–100, 100–500, and >500 and is denoted by $\mathrm{fe}(i)$. Furthermore, we define the variable $\mathrm{bl}(i)$ as indicating whether a fragment end belongs to a fragment that lacks a secondary restriction site. Such fragment ends are denoted as 'blind' and are expected to generate longer 4C products than nonblind fragment ends (**Supplementary Fig. 2e**). More specifically, the 4C-seq products that map to blind fragment ends are dependent on the location of a secondary restriction site in another ligated fragment (which is part of the longer concatemer generated by the initial 3C procedure). Given these parameters, and empirical coverage profile $\mathrm{cov}(j)$ (defined as 1 if fragment end j is covered by a 4C product and 0 otherwise), we estimate the background probability of coverage for a fragment end i as

$$\mathrm{expcov}(i) = \frac{\#\{j \in \mathrm{frag.ends\ s.t.\ cov}(j)=1, \mathrm{fe}(j)=\mathrm{fe}(i), \mathrm{fl}(j)=\mathrm{fl}(i), \mathrm{bl}(j)=\mathrm{bl}(i)\}}{\#\{j \in \mathrm{frag.ends\ s.t.\ fe}(j)=\mathrm{fe}(i), \mathrm{fl}(j)=\mathrm{fl}(i), \mathrm{bl}(j)=\mathrm{bl}(i)\}}$$

Some considerations involving model estimation should be noted:

1. The model is estimated from fragment ends that are unique in the genome (see description of mapping strategy above). Normalization and computation of statistics in windows can be done using the nonunique fragment ends as well.
2. Only the binary (yes/no) coverage profile $\mathrm{cov}(j)$ is used by the model. Multiple-read coverage was shown empirically not to be more informative than single-read coverage for remote interactions (**Supplementary Fig. 2a**).
3. Separate background models are calculated for intra- and interchromosomal contacts, as the chromosomal territory effect results in remote intrachromosomal interactions at least three- to fivefold more enriched than interchromosomal interactions, even at a distance of >100 Mb from the viewpoint. A region of 10 Mb around the viewpoint is discarded when estimating the model for intrachromosomal contacts because coverage in this region cannot be assumed a priori to be uniform.

Computing contact enrichment values for multiscale windows. Analysis and visualization of long-range contacts is done by computing enrichment over genomic windows

$$oe(x,y) = \mathrm{lr}\left(\sum_{i\in[x,y]} \mathrm{cov}(i), \sum_{i\in[x,y]} \mathrm{expcov}(i)\right)$$

where the summation is performed over all fragment ends in the genomic range $[x,y]$ and the lr function is defined as a regularized log ratio

$$\mathrm{lr}(o,e) = \log 2\left(\frac{o+\max(0,\mathrm{prior}-e)}{\max(e,\mathrm{prior})}\right)$$

and 'prior' is a regularization parameter that we empirically set to 4. This approach is appropriate for quantifying the intensity of contact enrichment across different window sizes. It does not guarantee statistical significance; we suggest that P values be used not to explore contact intensities but rather only to confirm hypotheses on patterns of such contacts. When required, P values for rejecting the background model can be easily generated using empirical likelihood ratio tests or normal approximation of the log likelihood distribution over genomic windows of a given size.

Normalization of quantitative contact profiles in the region proximal to the viewpoint. For analysis of proximal contacts, the inherent extreme variability in contact intensities in the viewpoint region undermines the uniformity assumption that allows the construction of the complex and parameter-rich background model described above. Empirical analysis suggests that the single most important factor affecting read count in the viewpoint region is whether a fragment is blind (lacking a second restriction site) or nonblind (**Supplementary Fig. 2c**). Different genomic windows have variable ratios of blind and nonblind fragments, as determined by the frequency of the secondary restriction sites within the window. This ratio may be coupled with various genomic features, such as the regional GC content, gene density and repeat density. As a result, without proper normalization,

regions that are rich in blind fragments will be biased toward lower 4C-seq coverage (**Supplementary Fig. 2e**). Such lower 4C-seq coverage may in turn be misinterpreted to be correlated with features that are associated with gene regulation. One must therefore normalize blind and nonblind 4C-seq coverage in a quantitative fashion that is robust to the variable density of fragment-end types and to the biased amplification of the ligation products involving these ends.

The normalization scheme for the contact profiles in the region proximal to the viewpoint is given a set of quantitative coverage profiles and assumes a priori that all of the profiles represent the same distribution of contact intensities. Distinct profiles are generated from the blind and nonblind fragments in the same experiment, and optionally from replicate experiments, or even from two experiments using different first restriction enzymes, assuming the viewpoints are located within a short genomic distance. The algorithm then performs several steps to combine the profiles:

1. Fragment ends are mapped back to genomic coordinates (using bins of 16 bp in our current implementation). Each fragment end is mapped to the center of the fragment, such that the 3′ and 5′ fragment ends are mapped to the same genomic bin. Nonunique fragment ends are masked.
2. We perform linear interpolation of the 3′ and 5′ coverage at fragment centers to generate coverage values for the remaining genomic bins. This is done to generate a fixed number of data points for each genomic interval and prevents systematic biases that can be caused by nonuniform distributions of blind or nonblind fragments. This results in four coverage profiles per track: 5′ nonblind, 3′ nonblind, 5′ blind and 3′ blind.
3. One interpolated profile is selected and all other profiles are quantile normalized to match its distribution. We then project the normalized interpolated profiles back onto the fragment end space.
4. Resulting profiles are combined by direct summation. The maximum median for all windows of size 5 kb (or as determined by the user) is identified and all medians are scaled by it. All depicted median values thus represent enrichment relative to the maximum attainable 5-kb median value.
5. Medians of normalized coverage for running windows of size 5 kb are generated (to plot the contact intensity trend). The 20th and 80th percentiles are also computed and depicted (percentiles can be determined by the user). For visualization purposes, median trends are weakly smoothed (using means of three consecutive data points), whereas the 20th and 80th percentile trends are more vigorously smoothed (using means of seven consecutive data points).
6. Medians are also calculated for sliding windows (2–50 kb) of linearly increasing size, displayed as color-coded multiscale diagrams, with values representing enrichment relative to the maximum attainable 12-kb median value.

It is also possible to use statistics other than the median to view contact profiles near the viewpoint. These include mean, geometric mean and variations that allow truncation of extreme values. Nonmedian statistics also support the use of standard deviation in place of percentiles. The options are supported by the pipeline but were not used in the analysis reported here.

We note that although empirical analysis indicates that additional factors beyond the blind/nonblind distinction (such as fragment length) may be correlated with systematic coverage biases in the region proximal to the viewpoint, such biases are small compared to the dynamic range of the typical contact profile (which spans 3–4 orders of magnitude). Therefore, these additional factors cannot be effectively used for further normalization in the region proximal to the viewpoint (which contains a limited number of fragment ends).

Luciferase assays. Candidate enhancers were PCR amplified with Phusion Taq polymerase with the primers listed in **Supplementary Table 3**. The Oct4–1.5kb (1,955 bp), Oct4–17kb (1,084 bp), Oct4–20kb (1,700 bp), Oct4–21kb (1,523 bp) and Satb1–649kb (1,406 bp) amplicons were first cloned into Clone Jet vector (Fermentas) with the blunt-end protocol and were then subcloned into the TATA-box–containing pGL4.10(luc2) plasmid (Promega), whereas Oct4–13kb (1,400 bp), Oct4–15kb (1,504 bp), Satb1–648kb (1,097 bp), Satb1–470kb (1,119 bp) and Satb1–253kb (1,108 bp) were directly cloned into TATA-box–containing pGL4.10(luc2) plasmid using the In-Fusion HD cloning kit (Clontech).

Mouse ES (mES) cells were grown in buffalo rat liver cell–conditioned medium combined with DMEM GlutaMAX (Gibco) containing 15% FBS (Invitrogen), leukemia inhibitory factor (LIF), 2-β-mercaptoethanol and nonessential amino acids (Invitrogen), as previously described[24]. Each experiment was performed with 2 ng renilla luciferase plasmid, 100 ng (or amounts corrected for plasmid size) pGL4.10 plasmid, and an unrelated 'stuffer' plasmid up to 800 ng per well of a 24-well plate (Gibco). mES cells were transfected using Lipofectamine 2000 Transfection Reagent according to the manufacturer's instructions. Because primary mouse lymphocytes are difficult to culture and transfect, we used the human lymphoid Jurkat cell line, which expresses *SATB1* at reasonably high levels, as a surrogate system to test enhancer activity of the selected sites around the *Satb1* gene. The Jurkat cells were grown in RPMI 1640 medium (GIBCO) with 10% FCS and 1% penicillin and streptomycin at 37 °C and 5% CO_2. Each experiment was performed with 500 ng pGL4.10 plasmid and 10 ng renilla plasmid. Jurkat cells were transfected through electroporation. At 24 h after transfection, dual luciferase reporter assays (Promega) were carried out with a Centro XS3 Microplate Luminometer LB960 (Berthold Technologies) according to the DLR kit protocol (Promega). All experiments were conducted under the approval of the animal care committee of the KNAW (Netherlands Royal Academy of Arts and Sciences).

Chromatin immunoprecipitation. ChIP analysis on fetal liver cells was performed according to standard procedures, as previously described[25]. The primers used for this experiment are included in **Supplementary Table 4**.

20. Yaffe, E. & Tanay, A. *Nat. Genet.* **43**, 1059–1065 (2011).
21. Simonis, M., Kooren, J. & de Laat, W. *Nat. Methods* **4**, 895–901 (2007).
22. Rozen, S. & Skaletsky, H. *Methods Mol. Biol.* **132**, 365–386 (2000).
23. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. *J. Comput. Biol.* **7**, 203–214 (2000).
24. Di Stefano, B. *et al. PLoS ONE* **5**, e16092 (2010).
25. Kooren, J. *et al. J. Biol. Chem.* **282**, 16544–16552 (2007).

doi:10.1038/nmeth.2173