UMI-4C for quantitative and targeted chromosomal contact profiling

Omer Schwartzman[1–4], Zohar Mukamel[3,4], Noa Oded-Elkayam[3,4], Pedro Olivares-Chauvet[3,4], Yaniv Lubling[3,4], Gilad Landan[3,4], Shai Izraeli[1,2] & Amos Tanay[3,4]

We developed a targeted chromosome conformation capture (4C) approach that uses unique molecular identifiers (UMIs) to derive high-complexity quantitative chromosome contact profiles with controlled signal-to-noise ratios. UMI-4C detects chromosomal interactions with improved sensitivity and specificity, and it can easily be multiplexed to allow robust comparison of contact distributions between loci and conditions. This approach may open the way to the incorporation of contact distributions into quantitative models of gene regulation.

Mapping chromosomal contacts using chromosome conformation capture (3C)[1] is increasingly used for assessing the physical linkage between distal genetic elements. As comprehensive catalogs of gene regulatory elements are now available[2–5], the association of regulatory elements with their targets[6–8] or, more generally, the characterization of complex relationships between multiple such elements and multiple genes have emerged as key challenges[8–12]. To address these challenges, many variants of the 3C approach were developed, offering strategies for fixation, digestion and recovery of proximity events between remote genomic elements on a massive scale[13]. Of particular interest are recent methods allowing the sequencing of 3C ligation products that can target several (4C)[14–17] to thousands (T2C, capture Hi-C) of genomic loci[18–24].

Typically, the sequencing of 3C ligation products is carried out in order to build contact matrices or contact profiles. These profiles are then analyzed statistically either to detect significant long-range contacts and loops or to compare profiles between conditions, tissues or cell types. In both cases, analysis is typically qualitative and performed through statistical hypothesis testing of contact enrichment between pairs of elements[12,25–27]. Reliable quantification of the contact intensity within defined intervals remains difficult to substantiate using current methods. Some major confounding factors prevent proper control over the technical variance of 3C experiments and quantitative analysis of the data[26–29]. In particular, in 4C, PCR-inflated ligation read counts are hard to correlate directly with the probability of proximity between elements.

Here we introduce a 4C procedure for counting ligation products in a precise manner. Experimentally, our approach uses frequent cutters, sonication and multiplexed, nested ligation-mediated PCR (LM-PCR) protocol to recover and count informative 3C ligation molecules. Computationally, we developed strategies for filtering different classes of spurious PCR products, deriving precise molecule counts and restricting experimental noise considerably while controlling technical variance in a predictable fashion. Software used to carry out these functions is freely available (https://bitbucket.org/tanaylab/umi4cpackage and **Supplementary Software**). We show that UMI-4C requires modest sequencing depth (100,000 reads per bait) and can be easily multiplexed, allowing multiple viewpoints to be selected (e.g., using a reference Hi-C map) and profiled rapidly and efficiently at high resolution. UMI-4C is a flexible solution for both the detection and quantitative comparison of chromosomal-contact distributions, with multiple applications in studies of genome regulation.

RESULTS

UMI-4C protocol

We developed a frequent-cutter 3C protocol (**Fig. 1a**) based on sonication and nested LM-PCR[7] amplification. LM-PCR is performed by ligating sequencing adapters to one end of each sonicated 3C fragment and following this by two rounds of nested amplification with bait primers directed to the other end of the targeted molecules. This strategy simultaneously facilitates the targeting of loci of interest and the construction of a sequencing-ready 4C library. The approach is simple and efficient, and it involves amplification of DNA fragments with relatively uniform (and, for the most part, locus-independent) length distributions due to sonication (**Supplementary Fig. 1a**). This uniformity helps reduce the systematic effect of PCR preferences for shorter sequences.

Viewpoint selection and optimization are also improved compared to 4C-seq, since UMI-4C primers can be designed

[1]Department of Human Molecular Genetics and Biochemistry, Tel Aviv University, Tel Aviv, Israel. [2]Functional Genomics and Childhood Leukemia Research Section, Cancer Research Center, Edmond and Lily Safra Children's Hospital, Sheba Medical Center, Ramat-Gan, Israel. [3]Department of Computer Science and Applied Mathematics, Weizmann Institute, Rehovot, Israel. [4]Department of Biological Regulation, Weizmann Institute, Rehovot, Israel. Correspondence should be addressed to A.T. (amos.tanay@weizmann.ac.il).

RECEIVED 31 AUGUST 2015; ACCEPTED 27 MAY 2016; PUBLISHED ONLINE 4 JULY 2016; DOI:10.1038/NMETH.3922

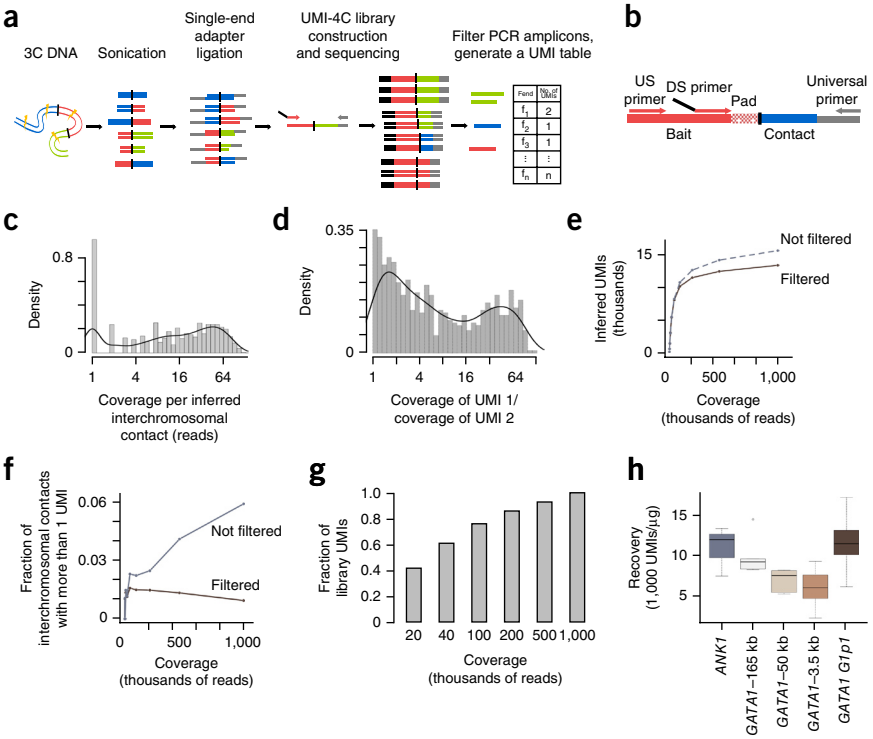

Figure 1 | Description of UMI-4C. (**a**) The UMI-4C protocol. Vertical black lines represent ligation sites, and jagged yellow lines depict sonication breakpoints. Fend, fragment end. (**b**) Bait design in UMI-4C. Note the use of a pad sequence between the restriction site and primer. US, upstream. DS, downstream. (**c**–**g**) *ANK1* bait analyzed in CMK cells. (**c**) Density plot of coverage (number of reads) per fragment end. Interchromosomal contacts with one inferred UMI (N = 3,696) are shown. (**d**) Density plot of fragment ends covered with exactly two UMIs (before further filtering) showing the ratio between the number of reads covering the first and second UMI (N = 697). (**e**) The number of inferred UMIs (with and without further filtering) as a function of sequencing depth (coverage). (**f**) The fraction of fragment ends that represent interchromosomal *ANK1* contacts and are supported by more than one inferred UMI as a function of library sequencing depth. (**g**) Fractions of UMIs detected (from the total of 13,101) as a function of down-sampled sequencing depth. (**h**) Distributions of the number of called UMIs per sample per 1 µg of starting UMI-4C template for five different baits.

flexibly relative to the restriction-site grid. The nested amplification strategy results in average target specificity of 69% (**Supplementary Table 1**), as determined by the existence of pad sequences between the sequenced bait primer and the bait restriction site (**Fig. 1b**). Most importantly, paired-end sequencing of UMI-4C libraries determines the sonication site of the ligated partners and can be used as a bona fide UMI[30] to eliminate PCR duplicates and diversify the library as discussed below. In summary, UMI-4C is a natural adaptation of current sensitive 3C strategies for applications in which focused and quantitative analysis of a locus of interest is desirable.

Molecule counting and filtering in UMI-4C

To evaluate UMI-4C performance, we applied the technique to five loci (located around the *GATA1* and *ANK1* genes) in five cell lines (**Supplementary Table 1**). We developed a computational pipeline for mapping ligated products to the genomic reference and counting raw UMIs based on the coordinate of the mapped sonication site (see Online Methods). When sequencing the library at high depth, we expected the number of reads mapping to each raw UMI to increase consistently for all molecules, with rapidly decreasing detection of new UMIs once the median read count per called molecule was sufficiently high (e.g., ten or more). We also expected 4C profiles to maintain a constant fraction of *trans*-chromosomal contacts as coverage increased, since *cis*- and *trans*-contacting molecules, once formed by 3C, are expected to be amplified with similar efficiency.

Empirically, however, we observed several sources of UMI skews that should be accounted for by further computational filtering. First, we observed that a subset of the raw UMIs was covered by one or only a few reads, even when sequencing depth was high (**Fig. 1c** and **Supplementary Fig. 2a**). These products may represent low-frequency spurious PCR events that link nonspecific genomic sequences (present in excess during the initial PCR cycles) with sequences sampled from specific templates. Second, we observed raw UMIs showing low read coverage in restriction fragments that were also mapped by deeply sequenced UMIs (**Fig. 1d** and **Supplementary Fig. 2b**). These products likely resulted from synthesis or sequencing errors that occur at a low rate but diversify valid ligation products and create spurious UMIs.

A computational strategy for filtering potential artifacts using read-coverage statistics can eliminate a significant fraction of the noise resulting from the two effects described above (see Online Methods). Filtering raw UMIs ensures convergence of the total number of called molecules when sequencing depth increases (**Fig. 1e** and **Supplementary Fig. 2c**) and prevents a continuous increase in the fraction of *trans*-chromosomal contacts covered by more than one UMI, which are unlikely to represent true double-proximity events (**Fig. 1f**). The UMI strategy thus successfully decouples sequencing depth from library complexity and quality. Inference of molecules from raw UMI statistics must be approached carefully, however, since even a low rate of PCR and sequencing errors with low representation in read space can contaminate the molecule counts and introduce noise.

UMI-4C profile complexity

After filtering, UMI-4C profile statistics can be used to estimate the required sequencing depth for 4C libraries. A UMI-4C library that provides support for 13,101 UMIs at high sequencing depth allows 60% of the UMIs to be detected when sequencing as few as 50,000 reads and 93% of the UMIs to be detected when sequencing 500,000 reads (**Fig. 1g**). This scaling represents typical UMI-4C behavior, suggesting that sequencing 10× or 20× the number of molecules in the library can be adequate for most applications. We calculated the number of UMIs (or complexity) extracted

per µg of 4C starting material for five different baits in different cell lines (**Fig. 1h**). Yields varied but were not dramatically different between baits, suggesting that 5,000–10,000 UMIs can be extracted per µg of starting material. This represents an effective yield of 1.5–3.3% of the starting material (assuming 300,000 genomic copies per µg), with lost molecules attributable to the LM-PCR procedure and to undigested or religated restriction sites (see Online Methods). Our data indicate that the characteristic UMI-4C yield and sequencing depth allow contact profiles to be built from thousands of UMIs when starting from a few million cells and using a few hundred thousand reads per bait.

We compared UMI-4C to 4C-seq[15,17], which utilizes two rounds of restriction digestion and inverse PCR to generate sequencing-ready libraries. We generated 4C-seq[15] libraries from CMK cells for four baits that target the same loci used for constructing UMI-4C libraries. Using a similar (or somewhat higher for 4C-seq) amount of starting material, we observed that the fraction of the fragment ends covered by at least one read in the 1-Mb window around the viewpoint was similar in the two approaches (**Supplementary Fig. 3a**). The general correlation between the coverage per locus in the two methods was high (**Supplementary Fig. 3b**). Nevertheless, the wide distribution of 4C-seq read coverage per called interchromosomal contact, which is *a priori* expected to represent a single ligation event (**Supplementary Fig. 3c**), reconfirmed the degree of PCR amplification bias in this approach. The high variability in the number

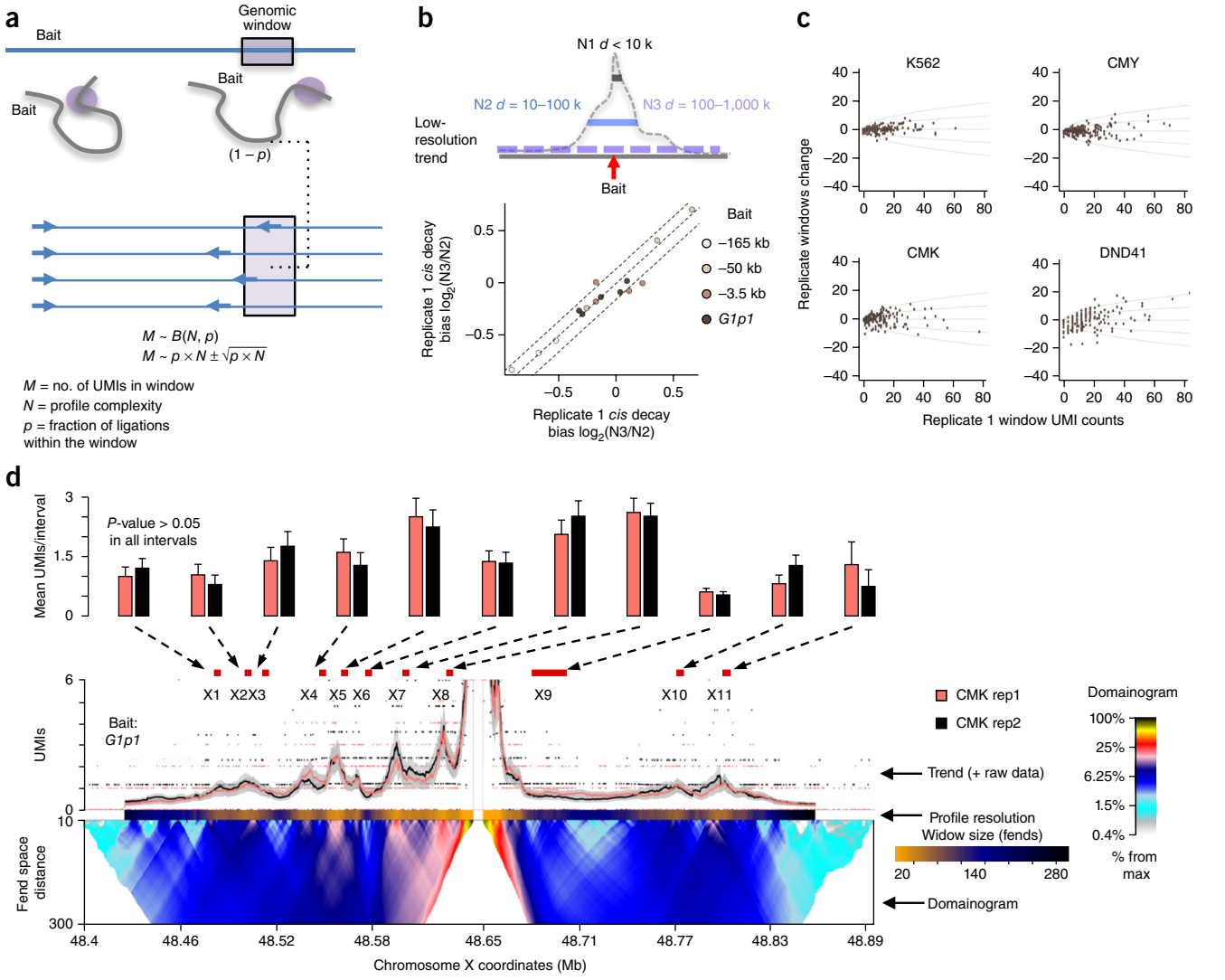

Figure 2 | UMI-4C reproducibility. (**a**) Schematic of molecule counting and sampling variance in UMI-4C. (**b**) UMI-4C normalization uses scaling based on three parameters (N1–N3) representing the total contacts shown in windows of distances (*d*) of 10 kb, 100 kb and 1 Mb around the bait (schematically shown on upper panel). Lower panel, ratio between 100 kb and 1 Mb statistics. (**c**) UMI counts in nonoverlapping windows of 20 fragment ends compared with the difference in UMI counts between two normalized replicate experiments. Dashed lines represent one and two theoretical binomial standard deviations. The analysis presented was performed on data from bait *G1p1*. (**d**) Replicated UMI-4C profiles for a bait located on the *GATA1* promoter assayed in CMK cells. The top bar graphs show normalized UMI counts in selected intervals (**Supplementary Table 3**). The middle panel describes the contact profile with the its components: smoothed trend line and raw counts (points) (top), adjusted profile resolution (middle color-coded band), and domainogram (bottom heat map). The trend reports mean UMIs per restriction-fragment end for overlapping windows defined such that each include at least 30 observed molecules. Resulting window sizes are reported in the profile resolution band. The domainogram reports mean contact per fragment end for a series of window sizes (from 10 to 300 fragment ends). Error bars, estimated binomial s.d. Fend, fragment end. Rep, replicate.

of 4C-seq reads per called UMI-4C contact (**Supplementary Fig. 3d**) demonstrated the difficulty in interpreting read-coverage statistics quantitatively. While 4C-seq and UMI-4C profiles and domainograms are globally similar (**Supplementary Fig. 3e**), the use of UMI counts provides improved quantitative resolution and statistical control while easing design and allowing flexibility for multiplexing, as discussed below.

Normalization and reproducibility of UMI-4C profiles

Using precise counts makes it possible to predict and control the technical variance of UMI-4C data. When a fixed locus of interest (defined by a genomic window) is assayed for contacts with a bait, the number of UMIs observed within the window, given a known profile complexity, is distributed binomially (**Fig. 2a**). This distribution reflects the sampling of molecules within the window at some constant probability per molecule (**Fig. 2a**). Assuming full control of all sources of systematic variation, two technically identical experiments should reproduce the profile within a s.d. that is predicted by this distribution. To control for the potential of different 3C batches to introduce global profile changes, we fit a low-resolution coverage profile to each experiment (**Fig. 2b**). This low-resolution trend quantified and normalized differences in total ligation frequencies in the 10-kb, 100-kb, and 1,000-kb distance bands—therefore not affecting significantly localized contacts. We found that replicate experiments using the same bait and cell line typically varied less than 20% in their ratios of total contacts in the 1-Mb and 100-kb bins,

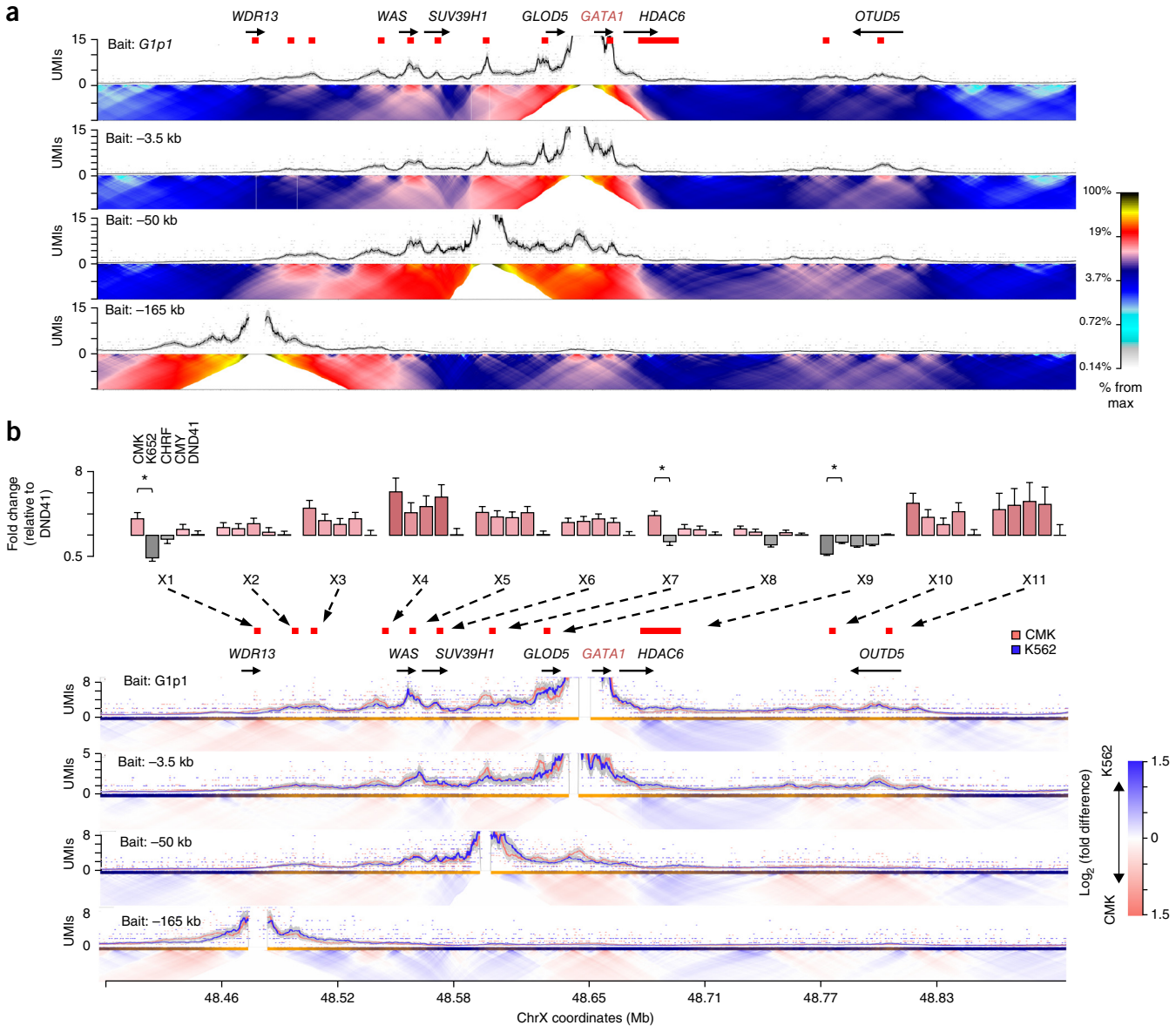

Figure 3 | UMI-4C analysis of *GATA1*. (**a**) UMI-4C trends and domainograms for the *GATA1* promoter (*G1p1*), two putative enhancers (–3.5 kb and –50 kb) and a domain border upstream of the gene (–165 kb). (**b**) Upper panel, fold change of contact intensities in 5 cell lines over 11 genomic intervals (X1–X11) along the *GATA1* region (**Supplementary Table 3**), quantifying the color-coded visualization shown in the lower panels. Lower panel, overlay of CMK and K562 UMI-4C trends and a domainogram showing the fold change in mean contact intensity. *P*-values <0.01 are denoted by an asterisk (χ^2 test; see Online Methods). Error bars, estimated binomial s.d. Domainogram color values are relative to the maximum profile value in the genomic window.

but larger differences could occasionally be observed (**Fig. 2b**). Upon normalization, however, we observed reproducibility that was largely compatible with our expectations (**Fig. 2c**), meaning that differences in the number of contacts per window (across all nonoverlapping windows) were within the theoretically predicted deviation.

The well-defined behavior of the technical variance in UMI-4C opens the way to simpler and more powerful statistical analyses. For example, to visualize the contact intensity profile at the maximal resolution allowed by the data, we select for each genomic coordinate the minimal window that is supported by a threshold number of UMI-4C molecules, and we compute the average contacts per restriction fragment ends in this window. In this approach, the variance of the estimated contact intensity is not affected by the distance from the bait (**Fig. 2d**, top and middle). We can also generate domainograms to visualize the contact

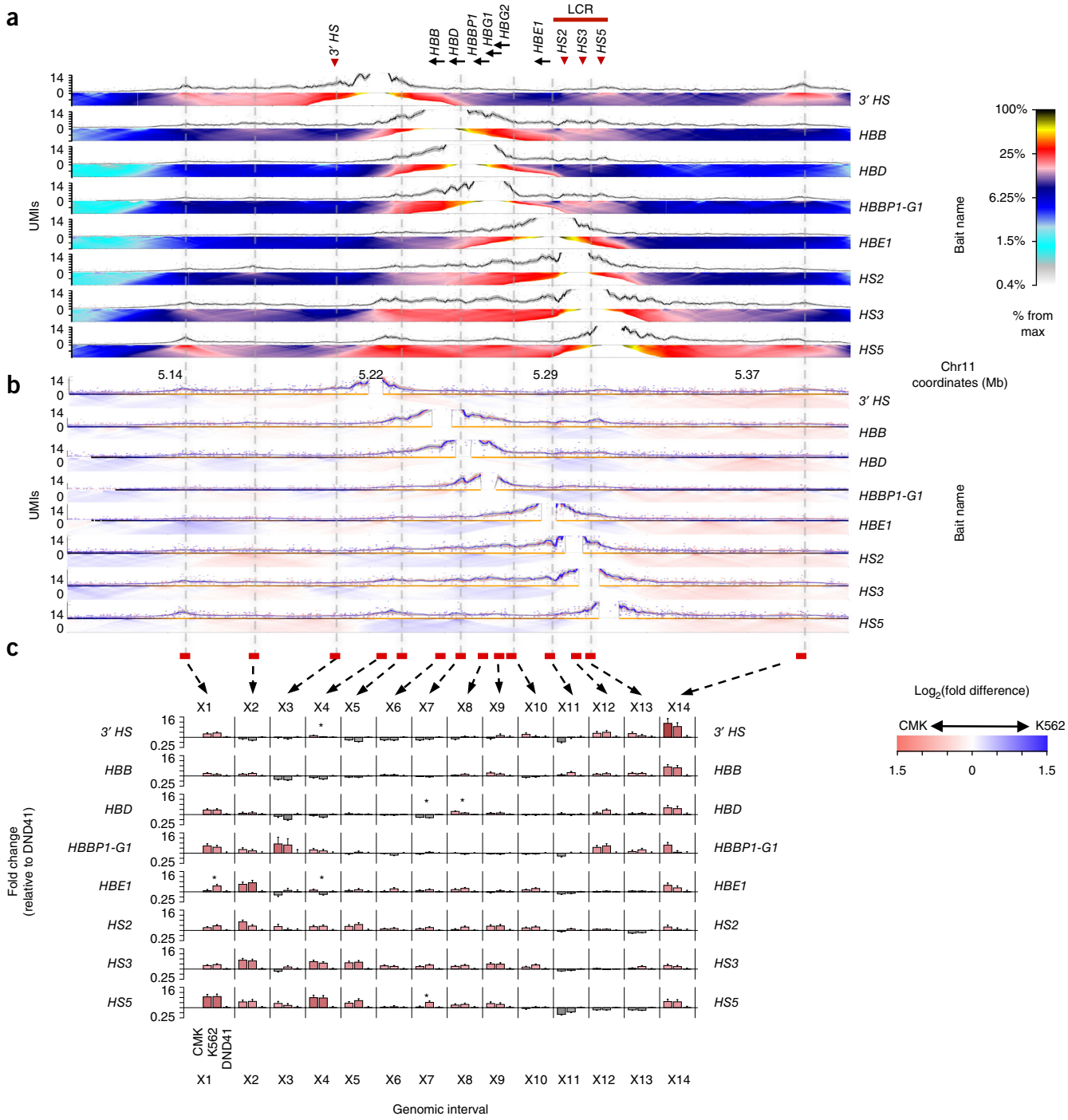

Figure 4 | Multiplexing UMI-4C in the *HBB* locus. (**a**) Contact profiles of eight baits around the human beta-globin locus and its locus-control region (LCR) in CMK cells. (**b**) Profile comparison of CMK and K562 for the eight baits shown in **a**. (**c**) Fold change of the contact intensities of CMK, K562 and DND41 cells over 14 genomic intervals (X1–X14, red boxes, **Supplementary Table 3**) along the *HBB* region. Each profile is normalized to DND41 and the fold change is relative to DND41. Asterisk denotes $P < 0.01$ (χ^2 test) for comparison between CMK and K562. Error bars, estimated binomial s.d.

profiles at multiple scales in one figure as done previously[10,15,31] (**Fig. 2d**, bottom), but UMI-4C allows the number of contacts per restriction fragment end to be used directly, without having to compute *P*-values and perform indirect hypothesis testing to account for PCR skews.

In conclusion, UMI-4C requires minimal normalization using a three-parameter low-resolution trend to heuristically account for the bulk of experimental 3C variance. Following this normalization, the data is defined by a predictable technical variance and is ready for direct and quantitative analysis. This can also be used for estimating the overall power of the assay (**Supplementary Fig. 4a**,**b**; see Online Methods).

Multiplexing UMI-4C

UMI-4C can be multiplexed by introducing multiple forward-target primers and a constant reverse primer (i.e., universal primer; see **Fig. 1b**) into the first and second LM-PCR stages of the protocol. We tested this approach by comparing multiplexed and independent UMI-4C on a set of four baits covering the *GATA1* and *ANK1* loci. Our protocol scaled well with respect to the overall number of UMIs extracted (**Supplementary Fig. 5a**). A high-complexity profile was derived for each of the multiplexed primers simultaneously and with minimal evidence for cross-primer contamination (**Supplementary Table 2**) or bias (**Supplementary Fig. 5b**,**c**). We noted that, even when variable amplification efficiency between bait primers led to an uneven read distribution between viewpoints (a challenge for 4C multiplexing in general), the total yield (measured by the number of UMIs, not reads) for each bait could be balanced (**Supplementary Table 2**). The amplification strategy and UMI-based analysis in UMI-4C thus streamlines multiplexing, allowing efficient interrogation of promoters or enhancers at high resolution from multiple genomic positions.

Quantitative UMI-4C analysis

We demonstrated the typical use of UMI-4C by prescreening a genomic region around *GATA1* using reference functional elements and K562 Hi-C maps[32] (**Supplementary Fig. 6a**). We designed a set of viewpoints for interrogating contact structure at the *GATA1* promoter and three upstream loci. The −3.5-kb and −50-kb viewpoints target regions of putative enhancers[33,34], and the −165-kb viewpoint was designed upstream of the *GATA1* domain border. UMI-4C was performed using 3C template from the *GATA1*-expressing leukemia cell lines K562, CMK, CMY and CHRF and from the *GATA1*-nonexpressing cell lines DND41 (T-cell leukemia) and WI-38 (immortalized fibroblast). We applied an adaptive-smoothing approach (see Online Methods) to generate the trend line for each profile (**Supplementary Fig. 6b**). As expected, the trend lines showed high correlation between different cells from the same lineage (**Supplementary Fig. 6c**). As demonstrated for CMK (**Fig. 3a**), we observed several long-range contacts upstream and downstream from the *GATA1* promoter at genomic distances from a few kilobases to hundreds of kilobases. Contacts observed from the promoter viewpoint were validated by contacts from the −3.5-kb viewpoint and by reciprocal contacts from the −50-kb viewpoint.

As UMI-4C is designed to quantify contact intensity across a wide range of genomic distances rather than to detect statistically significant 'interactions', we implemented a more quantitative approach that defines a set of potential genomic elements and assesses their contact intensities across cell types. By pooling replicate profiles and comparing CMK and K562 profiles (**Fig. 3b**), we discovered that both cell lines showed robust *GATA1* expression (**Supplementary Fig. 7a**), but the intensity of contacts between the promoter and its strongest long-range target (located at 50 kb upstream) was altered significantly (interval ×7, fold change 2.4, $P < 10^{-6}$, χ^2 test). A similar effect was observed when comparing the adjacent −3.5-kb bait (**Supplementary Fig. 7b**) and the reciprocal profile (**Supplementary Fig. 7c**, interval *G1p1*, fold change 1.61, $P < 9.1 \times 10^{-3}$, χ^2 test). Loss of contacts for elements downstream of the *GATA1* locus was observed in CMK (interval ×9, fold change 0.67, $P < 2 \times 10^{-4}$, χ^2 test). The *ANK1* locus provided another example for mapping differential contacts using UMI-4C (**Supplementary Fig. 8a**).

Next, we revisited the mapping of contacts in the human *HBB* locus using multiplexed UMI-4C on eight baits and three cell lines. Using only 1–3 million reads per sample, we constructed profiles that are likely of higher complexity than previously described (**Supplementary Table 1**), while deriving predictable sensitivity and specificity for detecting differential contacts. Analysis of the loci in lymphoid DND41 cells (**Supplementary Fig. 9a**) recovered the known symmetric and unperturbed contact distribution of the globin cluster in its inactive state. Nevertheless, the data also suggest weak contacts between a distal element in the locus control region (LCR) (baits HS2 and HS3) and the *HBB* locus (bait HBB), even in the inactive state. In addition, we observed a general increase in the intensity of long-range contacts between the locus and elements outside of the classic globin loop (**Supplementary Fig. 9a**). Analysis of the loci in CMK cells demonstrated strong long-range contacts with two flanking elements (**Fig. 4a**, marked as intervals X1 and X14) in the most upstream and most downstream baits tested (3HS and HS5). Comparative analysis of the region in K562 and CMK cells showed that the classic nested looping structure is present in both, but suggested the presence of a consistent conformation change that causes the LCR-globin minidomain to be more intensively contacting in K562 cells (**Fig. 4b**,**c**). It is possible that these conformation changes are correlated with the K562-specific expression of the most LCR-proximal globin gene, *HBE1* (**Supplementary Fig. 9b**).

DISCUSSION

UMI-4C is an efficient and accurate method for analyzing targeted loci, providing experimental flexibility and maximal complexity for the derived contact profiles with very low sequencing burden and experimental prerequisites. A key part of our approach is a new software package that facilitates UMI-4C analysis. We provide tools for the quantitative design of UMI-4C experiments as well as for *de novo* and comparative profile analysis with defined sensitivity to detect contact intensity changes at given fold-change enrichment levels. The use of precise molecule counting ensures that experiments are quantitatively comparable.

The UMI-4C method does not substitute for global approaches for chromosome conformation capture such as Hi-C and its many variants[13]. Also, it may not be the method of choice when large groups of promoters, single-nucleotide polymorphisms (SNPs), or large domains are to be analyzed routinely, in which case targeted approaches involving sequence capture may be more effective[18,19,21,23]. However, these approaches require significant sequencing and/or investments in capture libraries which can be

avoided by using UMI-4C. The threshold number of viewpoints that merit transition from UMI-4C to a broader targeted approach is hard to define precisely, but we suggest that experiments that focus on 20–50 viewpoints would be most economical and precise when using the multiplexed UMI-4C technique presented here.

An important aspect of UMI-4C is that it opens the way to quantitative interpretation of 3C data. Current Hi-C profiles are limited by sequence coverage and library complexity, making it difficult to accurately characterize the contact-distribution landscape around genomic windows on the scale of one functional element. Current 4C approaches, on the other hand, are difficult to interpret quantitatively because of PCR and other biases, even following extensive normalization[7,12,15,25,31]. More fundamentally, 3C data measures contact enrichment over an extremely heterogeneous background, which spans some five orders of magnitude in intensity on account of linear chromosome structure[26–28]. Together, these effects force statistical modeling of contact maps and contact profiles and prevent quantitative assessment of the intensity (rather than the significance) of contact enrichment for specific pairs of elements. We suggest that, by applying molecule counting and simple modeling of the background signal, UMI-4C can be used to estimate contact-intensity fold-enrichment levels over the background for a well-defined range of genomic linear distances and with statistical power that can be explicitly estimated.

METHODS

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

Accession codes. Gene Expression Omnibus: GSE76763. The UMI-4C pipeline is available at https://bitbucket.org/tanaylab/umi4cpackage.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

Research in the Tanay Lab was supported by the European Research Council, the MODHEP project and the Israeli Science Foundation. Research in the Izraeli Lab was supported by the Israel Science Foundation, the Waxman Cancer Research Foundation and the Dotan Center for Hematological Malignancies at Tel Aviv University. This work was performed in partial fulfillment of the requirements for a PhD degree by O.S.

AUTHOR CONTRIBUTIONS

O.S., S.I. and A.T. designed the study. O.S. and Z.M. developed the experimental protocol with help and reagents from N.O.-E., P.O.-C., Y.L. and G.L. O.S. performed experiments. O.S. and A.T. analyzed the data and developed the pipeline. O.S. and A.T. wrote the paper. S.I. and A.T. supervised research.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
2. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
3. Roadmap Epigenomics Consortium. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
4. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
5. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
6. Hakim, O. *et al.* Diverse gene reprogramming events occur in the same spatial clusters of distal regulatory elements. *Genome Res.* **21**, 697–706 (2011).
7. Apostolou, E. *et al.* Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. *Cell Stem Cell* **12**, 699–712 (2013).
8. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
9. Nora, E.P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
10. Splinter, E. *et al.* The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes Dev.* **25**, 1371–1383 (2011).
11. Dixon, J.R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
12. de Wit, E. *et al.* The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* **501**, 227–231 (2013).
13. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* **26**, 11–24 (2012).
14. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).
15. van de Werken, H.J.G. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat. Methods* **9**, 969–972 (2012).
16. Stadhouders, R. *et al.* Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nat. Protoc.* **8**, 509–524 (2013).
17. Splinter, E., de Wit, E., van de Werken, H.J.G., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods* **58**, 221–230 (2012).
18. Jäger, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.* **6**, 6178 (2015).
19. Hughes, J.R. *et al.* Analysis of hundreds of *cis*-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46**, 205–212 (2014).
20. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
21. Kolovos, P. *et al.* Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics Chromatin* **7**, 10 (2014).
22. Davies, J.O.J. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat. Methods* **13**, 74–80 (2016).
23. Sahlén, P. *et al.* Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biol.* **16**, 156 (2015).
24. Sanborn, A.L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. USA* **112**, E6456–E6465 (2015).
25. Ghavi-Helm, Y. *et al.* Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **512**, 96–100 (2014).
26. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* **43**, 1059–1065 (2011).
27. Ay, F., Bailey, T.L. & Noble, W.S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* **24**, 999–1011 (2014).
28. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
29. Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**, 3131–3133 (2012).
30. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
31. Bantignies, F. *et al.* Polycomb-dependent regulatory contacts between distant Hox loci in *Drosophila*. *Cell* **144**, 214–226 (2011).
32. Rao, S.S.P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
33. Valverde-Garduno, V. *et al.* Differences in the chromatin structure and *cis*-element organization of the human and mouse GATA1 loci: implications for *cis*-element identification. *Blood* **104**, 3106–3116 (2004).
34. Xu, J. *et al.* Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Dev. Cell* **23**, 796–811 (2012).

ONLINE METHODS

Cell culture. Human leukemia cell lines CMK, CMY, CHRF, K562 and DND41 were grown in RPMI 1640 (GIBCO) supplemented with 10% fetal bovine serum (FBS), 1% glutamine (GIBCO), penicillin (100 U/mL) and streptomycin (100 µg/mL). WI38 cells were grown in DMEM supplemented with 10% FBS, 1 mM sodium pyruvate, 1% glutamine (GIBCO), penicillin (100 U/mL) and streptomycin (100 µg/mL) (GIBCO). All cell lines were grown at 37 °C and 5% CO_2. Mycoplasma tests were routinely performed on all the cell lines. CMK, K562 and DND41 were acquired from the DSMZ. WI38 was acquired from ATCC. CMY and CHRF were kindly provided byJ. Crispino (Northwestern University, Chicago, IL). CMK and CMY were validated by sequencing of exon 2 of *GATA1* confirming the presence of mutations in the gene, distinctive of Down syndrome acute megakaryocytic leukemia, and by western blot analysis demonstrating exclusively the truncated GATA1 form (GATA1s) (data not shown). K562 and CHRF were validated by expression of the full-length *GATA1* in western blot. K562 cells were further validated by fluorescence *in situ* hybridization (FISH) for the genomic translocation t(9;22) *BCR-ABL*. DND41 cells were validated by the presence of the *IL7R* mutation (p.L242_L243insLSRC)[35]. CMK, CMY, CHRF and K562 cells were also validated by flow cytometry immunophenotyping for expression of the expected lineage markers.

3C. 10^7 cells were processed to generate the primary 3C template as described in ref. 17. In brief, cells were collected and centrifuged at 1,200 r.p.m., the supernatant was removed and the pellet was resuspended in PBS/10% FBS. The cells were cross-linked with formaldehyde to a final concentration of 2% for 10 min. The reaction was quenched by addition of glycine to a final concentration of 0.125 M. The cells were then centrifuged and resuspended in cold lysis buffer (50 mM Tris–HCl, pH 7.5, 150 mM NaCl, 5 mM EDTA, 0.5% NP-40, 1% TX-100), supplemented with 1 tablet protease inhibitor (cOmplete, Mini, EDTA-free, Roche; for 10 mL buffer volume). The nuclei were extracted and resuspended in 450 µl UPW, 60 µl DpnII buffer, 15 µl 10% SDS, and they were incubated on a Thermomixer (Eppendorf) 37 °C, 60 min, 900 r.p.m. 75 µl 20% TX-100 was added to the solution and incubated 37 °C, 60 min, 900 r.p.m. The chromatin was then digested using HC DpnII (NEB R0543M) at 37 °C, 900 r.p.m., in three rounds as follows: 200 U for 2 h, 200 U overnight and 200 U for 2 h. An aliquot was taken from the tube and analyzed by gel electrophoresis to inspect digestion efficiency. If digestion was successful, the restriction enzyme was inactivated by incubating at 65 °C for 20 min. Next, the chromatin was ligated with 15 µl of HC T4 ligase (NEB M0202M) with 10× ligase buffer to a final volume of 7 mL and incubated overnight at 16 °C. Ligation efficiency was inspected by gel electrophoresis on an aliquot of the sample. The chromatin was then reverse cross linked overnight at 65 °C with 30 µl proteinase K (Sigma-Aldrich, 10 mg/mL). 30 µl of RNase A (Sigma-Aldrich, R4642) were added and incubated for 45 min, 37 °C. The 3C DNA was then extracted with phenol-chloroform, ethanol precipitated, and resuspended in 10 mM Tris-HCl, pH 8. 3C aliquots were stored at −20 °C for later use.

UMI-4C primer design. UMI-4C primers can, in theory, be designed to interrogate ligations partners of any restriction fragment end in the vicinity of a given genomic point of interest. Our nested approach involved designing a set of two primers for each locus, thereby improving specificity. The primer for the second PCR reaction (referred to as *DS*) should ideally be located 5–15 bps from the interrogated restriction site. The pad sequence is used to report on amplification specificity, and it is relatively short to minimize the number of noninformative bases before the first ligation junction in the sequence. However, in problematic loci it is possible to locate the second primer further upstream of the restriction site. The primer used in the first PCR reaction (*US*) is designed upstream of the *DS* primer ideally with no overlap (but some overlap could exist if needed). We find that *US* and *DS* primer combinations with minimal number of hits in Primer-BLAST[36] result in increased on-target specificity. For multiplexing, we designed primers independently, aiming for a melting temperature (T_m) of 55–56 °C. A list of the primers used in this study is provided in **Supplementary Table 4**.

UMI-4C. 5–10 µg DNA in 200 µl elution buffer (EB, 10 mM Tris–HCl, pH 8) of 3C template was sonicated at 5 °C for five cycles of 30 s on, 60 s off on a Bioruptor (Diagenode). The sonicated material was inspected by gel electrophoresis of 10 µl aliquots. We found a size distribution of 450–550 bp ideal for the assay (**Supplementary Fig. 1a**). The sonicated DNA was subjected to an end-repair reaction—20 µl 10× end-repair buffer and 10 µl end-repair mix (NEB E6050L) were added to the sonicated DNA and incubated at 20 °C for 30 min. The DNA was cleaned with 2.2× AmpureXP beads (Beckman Coulter) and eluted in 76 µl EB. A-tailing was performed by adding 10 µl NEBbuffer 2, 4 µl Klenow fragment (NEB M0212M) and 10 µl 10 nM dATP. Following A-tailing, the 5′ ends of the DNA were dephosphorylated by adding 2 µl calf intestinal alkaline phosphatase (NEB M0290S) and incubated at 50 °C for 60 min. The DNA was cleaned with 2× AmpureXP beads and eluted in 67 µl EB. The DNA template was ligated with 5 µl of 15 µM Illumina-compatible forked indexed adapters (final concentration 0.4 µM; see **Supplementary Table 4**), 10 µl ligase mix and 80 µl 10× buffer (NEB M2200) for 15 min at 25 °C. To release the nonligated strand of the adaptor, the DNA was denatured at 95 °C for 2 min and cleaned with 1× AmpureXP beads. DNA concentration was measured as following: 1 µl aliquot was taken from the prepared template, diluted 1:5, denatured at 95 °C for 2 min and then placed on ice. DNA concentration was immediately measured with Qubit ssDNA kit (Thermo-Fisher Scientific).

Two nested PCR reactions were used for UMI-4C library construction. The PCR reactions were performed in 50 µl volume with the following reagents: 10 µl GoTaq Flexi buffer (Promga M792A), 3 µl MgCl 25 mM (Promega A3511), 0.2 mM dNTPs, 2 µl 10 mM *US* bait primer (0.4 mM final concentration), 2 µl 10 mM Illumina enrichment primer 2 (0.4 mM final concentration), 1 µl GoTaq Hot Start (Promega M5005) and 200 ng UMI-4C DNA template. PCR program: 2 min 95 °C, 20 cycles of 30 s 95 °C, 30 s 56 °C and 60 s 72 °C and final extension of 5 min 72 °C. Following the PCR reaction, the amplified DNA was cleaned with 1× AmpureXP beads, and the eluate was used as a template in the second PCR reaction. Conditions for the second PCR reaction were identical to the first, and the only differences were the use of the bait primer (*DS*) with the Illumina dangling adaptor (as shown in **Fig. 1b**) and the use of 15–18 amplification cycles instead of 20.

doi:10.1038/nmeth.3922

The amplified DNA was cleaned with AmpureXP beads 0.7×. The size distribution was inspected by Tapestation D1000 (Agilent Technologies), and the DNA concentration was measured using Qubit HS kit (Life Technologies). Typically, UMI-4C libraries have size distribution of around 500 bp (**Supplementary Fig. 1b**). The libraries were then pooled and diluted to 4–10 nM concentration and multiplexed with other libraries and sequenced on the Illumina platform.

The UMI-4C profiles described in this work were generated by pooling 5–10 200 ng PCR reactions in order to achieve high complexity. Moreover, 2–3 different UMI-4C templates, which were generated independently with different barcodes, were typically used. This allowed for the templates to serve as technical replicates, and it also increased the final complexity of the profiles. Mixing in the same PCR-tube templates with different barcodes was avoided in order to prevent barcode switching.

For multiplexed UMI-4C, the library preparation was done similarly to single-bait reactions as described above. Two primer pools were prepared , one for the first PCR reaction pooling the *US* primers and a second pool with the *DS* primers for the second PCR reaction. The pools were diluted such that each primer (including the Illumina universal primer) would have a final concentration of 5–10 mM (note that, when performing heavier multiplexing of >30 baits, reducing the primer concentrations to 0.33–1 mM per primer improves the yield significantly).

4C-seq. 20 µg 3C template DNA was processed as described in ref. 15. Briefly, the 3C template was incubated with 25 U of CviQI (NEB, R0639L) for 2 h in 37 °C, 900 r.p.m., in a Thermomixer (Eppendorf). Digestion efficiency was examined using Genomic TapeStation (Agilent). Next, the digested DNA was ligated with high-concentration T4 DNA Ligase (NEB M0202M) in 8 mL total volume at 16 °C overnight. The DNA was then purified with phenol-chloroform and eluted in ultrapure water (UPW).

The 4C-seq libraries in the paper were generated using the same forward bait as the UMI-4C baits for *G1p1*, G1-165 kb and *ANK1* and a different reverse primer targeting the CviQI fragment end. For the G1-50 kb bait, a DpnII fragment adjacent to the UMI-4C bait was selected due to the reverse primer PCR considerations (primer sequences are supplied in **Supplementary Table 4**). Each 4C-seq library was constructed from a total of 1.2 µg template (six 200 ng pooled PCR reactions). For bait G1-165 kb, 1 µg of 4C template was used. The libraries were sequenced on the MiSeq platform. We used the read count of the 4C-seq for the analysis described in the paper.

Sequencing, filtering and mapping. A flow chart outlining our UMI-4C processing pipeline is shown in **Supplementary Figure 10**. The main steps are described below. A software package implementing the process is also provided for complete details.

The sequencing UMI-4C libraries was done on Illumina machines (Miseq, Nextseq or HiSeq), typically multiplexed with other another sequencing application on the same flow cell. We found that the actual yield of sequencing varies and depends on the composition of the entire sequencing run. Since UMI-4C libraries are typically longer than average sequencing libraries (e.g., ChIP and RRBS) and also possibly contain a fraction of unsequenceable DNA (**Supplementary Fig. 1c**), loading 10–15% more DNA than the calculated molarity can compensate for this gap (e.g., if mixing 1:1 ChIP libraries and UMI-4C, the actual mixing will be 0.8:1 ChIP:UMI-4C). The minimal sequencing requirement is a paired-end run with read length of at least 60–70 bases from read 1 (depending on the length of the bait + pad sequence) and at least 40–50 bases from read 2. Each paired-read covers the 3′ and 5′ end of one library product (**Fig. 1b**). When read lengths are not very long (less than 150 cycles for each end), these tags are typically nonoverlapping, since the size-selected sequencing library is longer than 300 bp (**Supplementary Fig. 1b**). FASTQ files for the 5′ end (read 1) are used to perform initial filtering and to demultiplex raw reads by searching for prefixes matching primer sequences. Sequences containing one of the primer sequences are tested for a match with the designed pad sequence, separating the primer site and the interrogated restriction site. Reads lacking this match are considered nonspecific and are filtered from additional analysis.

In cases of deteriorating sequencing quality in later cycles of the sequencing process (e.g., deteriorating quality caused by multiplexed sequencing runs with low-complexity material or other technical issues), we found that it is important to prune reads suffixes such that the average quality in the last retained base is better than 20 (Phred score) for at least 80% of the bases. This approach decreases the effects of inconclusive mapping and improves the accuracy of the UMIs that are attached to each read. Following pruning, read pairs are split by occurrence of a perfect match of the restriction-enzyme sequence (GATC for DpnII), creating a new FASTQ file with an entry for each segment (including the flanking restriction sequences if these exist). The prefix of read 2 is extracted and attached to the segments derived from each read pair (to be used as *UMI1*). We used Bowtie 2 (ref. 37) to map the segment sequences to the genome (independently for each segment).

Mapped segments (one or more) for each read pair with MAPQ greater than 30 are further analyzed to test for compatible (reverse-complementing) segments in read 1 and read 2, based on the mapping coordinate and the read length. Each read pair is then transformed to a series of (one or more) mapped coordinates within distinct restriction fragments, where the fragments identified in read 1 are followed (in reverse order) by those identified in read 2.

At this stage, all informative read pairs (involving more than one restriction fragment) are represented by a fragment-chain table (denoted fendchain) defined as follows (i is an index running for $1 \ldots n$ for each read pair, n being the number of mapped fragment for that read pair).

TABLE FendChain. *RID*: Read-pair identifier
UMI1: sequence of last 10 bases of the 3′ end of the read pair
For $i = 1 \ldots n$:
$Frag_i$: restriction fragment identifier (determined based on the mapped coordinate)
$Strand_i$: strand
$Off5_i$: 5′ offset within fragment
$Off3_i$: 3′ offset within fragment

Each of these series includes one of the bait fragments as the first element (because of the filtering performed in the initial step of the pipeline). The second restriction fragment, if it exists, defines a potentially informative ligation partner for the captured target or a PCR duplicate of such a ligation molecule.

UMIs and molecule calling. Every informative read pair defines a potential UMI-4C ligation that is determined by the two pairs ($Frag_1$, $Strand_1$) and ($Frag_2$, $Strand_2$). The reason for considering the strand is that each restriction fragment can be ligated on either the 3′ or the 5′ end.

Naively, UMI-4C PCR duplicates are expected to be easily identifiable as sequencing products specifying the same ligation that share a *UMI1* sequence (5′ end of the read-pair). However, sequencing and synthesis errors can diversify the *UMI1* tags of UMI-4C duplicates, and additional PCR effects can introduce spurious ligations that must be filtered out to ensure perfectly reproducible profiles. Our UMI filtering algorithm therefore uses both *UMI1* sequences and information on the entire series of mapped restriction segments for each read pair, as well as coverage statistics, to infer the number of molecules present for each ligation in a given library.

Algorithm UMI filtering (for a given ligation).

1. Collect all read-pairs specifying a ligation ($Frag_1$, $Strand_1$) and ($Frag_2$, $Strand_2$).

Extract and summarize UMIs.

2. For each read pair, extract the *UMI1* sequence.
3. Create a *UMI2* identifier by concatenating all ($Frag_i$, $Strand_i$, $Off5_i$, $Off3_i$). Create a *UMI3* identifier using ($Frag_n$,$Off3_n$) (*n* being the 3′-most fragment for the read pair).
4. Find all keys *umi* = *UMI1.UMI2.UMI3* for the ligation to generate a set *raw_key_set*.
5. Summarize the number of read pairs for each key in *key_cov*{*umi*}.

Greedy filtering of potentially overlapping UMIs.

6. Initialize *key_set* to an empty set.

While *raw_key_set* is not empty repeat 7–10:

7. Find *cur_umi* as *cu1.cu2.cu3* = the key (or one of the keys) in *raw_key_set* with maximum *key_cov*.
8. *dominated_keys* = all keys *UMI1.UMI2.UMI3* in *raw_key_set* such that at least one the following holds:
   a. number of mismatches(*cu1*, *UMI1*) < *max_filter_hamming* (= 3).
   b. *cu2* equals *UMI2*.
   c. *cu3* is within *max_sonic_d* (= 1) bp from *UMI3*.
9. Add *cur_umi* to *key_set*, and add to *key_cov*{*cur_umi*} the sum of *key_cov* on all keys within *dominated_keys* (i.e. since we assume all *dominated_keys* are duplicates of *cur_key*).
10. Remove *dominated_keys* and *cur_key* from *raw_key_set* and *key_cov*.

Filter UMIs with low relative coverage.

11. max_reads = max(key_cov{u})
12. Remove from *cur_key* all *u′* such that:
    key_cov{*u′*} < *max_reads* × *switch_ratio* (switch_ratio = 0.1)
13. Report the number of retained umis in *key_cov* as the number of inferred molecules.

After the algorithm for all detectable ligations has been applied, the UMI-4C experiment is summarized in an adjacency table (denoted Table Adj) specifying molecule counts per ligations:

Table Adj. Table with the following columns:
Fragment end identifier 1
Fragment end identifier 2
Number of inferred molecules

Calling nondigested products and avoiding UMI saturation in high-complexity libraries. Our analysis of UMI-4C, and more generally of 3C sequencing products, suggests that fragment pairs mapped to genomic coordinates within less than 1,500 bp likely represent nondigested products (e.g., amplification of endogenous sequences). Therefore, all UMIs representing such products are excluded from further processing and their relative frequency is reported as a QC parameter. Note that sonication diversifies UMI-4C products in a limited way, since the theoretical maximum number of molecules we can infer is bound by the number of sonication sites in the fragment (actually half this number since we filter adjacent sonication sites as described above). Therefore, it is problematic to count the frequency of nondigested products (which are representing 30–50% of the library-sequenced reads) using UMIs, and the total number of reads representing nondigested UMIs for is used for QC.

As a general guideline, several independent libraries should be used when aiming to achieve high-complexity profiles in which valid long-range contacts (>1,500 bp) can be covered by many UMIs (over 20). This strategy helps avoid underestimation of the contact intensity on proximal or other loci with high contact intensity due to UMI saturation. Independent libraries should be processed separately to infer UMIs, and the molecule counts from these libraries are then summed up without further filtering.

Contact intensity profiles. To construct contact intensity profiles around a UMI-4C viewpoint (*Fragment ID*, *Strand*), data from libraries that amplified material from the same 3C experiment are merged by summation of the molecule counts per ligated fragment. Then, a spatial profile is generated by extracting the number of ligations for all restriction fragment ends within a genomic window around the viewpoint. This statistic includes restriction fragment ends that are not covered by any molecule. Restriction fragments of less than 20 bp in length or less than 60% mappability are removed. The data is then organized into two lists:

*Raw_prof*_ds = (X^{ds}_i, mol^{ds}_i), where *X* is a genomic coordinate and *mol* is the molecule count for that coordinate, and all fragments downstream of the viewpoint are included in increasing order.

*Raw_prof*_us = (X^{us}_i, mol^{us}_i), similar to *Raw_prof_ds*, but for fragments upstream of the bait in decreasing order.

This strategy of splitting the profile into upstream and downstream regions is applied in order to control better for the effects of highly covered contacting fragments within the first few kb around the bait. The effective resolution of the contact intensity profile is determined by requiring a minimum molecule count (*win_cov*) per window:

$Res\_ds[i] = min(h \text{ s.t. } sum(mol^{ds}_{i-h}..mol^{ds}_{i+h}) >= win\_cov)$
$Res\_us[i] = min(h \text{ s.t. } sum(mol^{us}_{i-h}..mol^{us}_{i+h}) >= win\_cov)$

Then, the average contact intensity per fragment is computed as follows:

$U4c\_ds[i] = mean(mol^{ds}_{i-h}..mol^{ds}_{i+h})$, where $h = Res\_ds[i]$
$U4c\_us[i] = mean(mol^{us}_{i-h}..mol^{us}_{i+h})$, where $h = Res\_us[i]$

win_cov = 70 is used for the profiles in the paper (unless otherwise specified). Given that UMI count within a window can be modeled as a binomial sampling of ligations with a fixed probability of hitting the window of interest, the margins of the trend are computed as ±2 of the s.d. of the mean estimator, which is

doi:10.1038/nmeth.3922

approximately the square root of the window sum of molecules divided by the size of the window.

The domainogram is a color-coded visualization of a sliding window mean statistics derived from the raw profiles. It is a stacked representation of contact intensity values in increasing fixed-window sizes from 10 to 300 fragment ends. The values are normalized by dividing all the mean window values by the maximal profile value. Note that window sizes are defined by the number of (mappable, >20-bp-long) interrogated fragment ends and not by the size of the genomic interval in base pairs. This approach regularizes the domainogram (e.g., in repetitive regions), but should be carefully evaluated if fragment density is changing significantly between regions around the 4C viewpoint.

Normalizing and comparing two profiles. To compare two UMI-4C profiles, large-scale symmetric biases are first corrected as follows (**Fig. 2b**). For each profile, the total molecule count (raw_prof_us, raw_prof_ds) for restriction fragment ends within 1–10 kb, 10–100 kb and 100–1,000 kb of the viewpoint coordinate are computed, generating *N11*, *N12* and *N13* for profile 1 and *N21*, *N22* and *N23* for profile 2. Note that this is assuming a symmetric background model (both downstream and upstream statistics are considered). Then, one profile is selected as the reference (e.g., the first), and all statistics *U4c_ds* and *U4c_us* of the second profile are scaled by the factors *N11*/*N21*, *N12*/*N22* and *N13*/*N23* (selecting the factor according to the distance from the viewpoint). The correction factors in the 50 fragment ends around the 10 kb and 100 kb distances are smoothed linearly.

When plotting a comparison of two profiles, the trend resolution is set as described above, but both profiles must have at least *win_cov* molecules in the plotted window.

The domainogram in comparison plots is based on $\log_2$ of the ratio between the domainogram values of the compared profiles.

Estimation of UMI-4C power. We developed a simple model for power analysis and estimated the minimum number of molecules required for a UMI-4C profile for a given level of detection sensitivity. Our model is based on the understanding that chromosomal contact intensity first and foremost depends on the linear genomic distance. For example, the expected contact intensity per a constant genomic window at the 10 kb range is 100 times larger than the intensity at the 1 Mb range and 1,000 times larger than the intensity at the 10 Mb range[26,38,39]. This implies strikingly different background levels when analyzing medium-range contacts within topological domains (at a distance scale of 100 kb) or when searching for long-range interdomain contacts (at a distance scale of 1 Mb). Using deep Hi-C data[32] for calibration (**Supplementary Fig. 4a**), we can estimate the expected (background) number of derived 4C UMIs per (noninteracting) genomic window size and genomic linear distance from an idealized bait. We can then compute the minimum UMI-4C profile complexity required for detecting two-fold change in the locus contact intensity at a desired confidence level. Our model involves a simple correction for the multiple testing implied by screening all windows up to a maximum linear distance. The background 3C contact probability distribution is assumed to be a simple function of linear genomic distance $F(d)$, estimated directly from high-coverage Hi-C data using the frequency of contacts in 19 $\log_2$(distance) bins. According to the background model, sampling UMI-4C ligations is expected to generate the background $F(d)$ distribution, and so in a UMI-4C profile of N molecules, the number of molecules expected in a genomic window of size W and mean linear distance d is approximately:

$$E = N \times W \times F(d)$$

A contact hotspot can be defined based on its fold-change enrichment compared to the background, denoted r. In this case, we expect more UMI-4C contacts in the window:

$$E_1 = r \times E = r \times N \times W \times F(d)$$

We can estimate the detection power of UMI-4C by assuming a binomial distribution of the background with N samples and success probability E/N. We can then compute the exact binomial P-value $Pr(B(N, E/N) > E_1)$ to determine the false-positive rate of detecting fold change r in the given coverage depth, linear distance and window size. To correct for multiple testing, we heuristically multiply the P-value by the number of nonoverlapping tests that can be performed for a window W up to distance d, deriving:

$$P' = Pr(B(N, E/N) > E1) \times (W/d)$$

For the power estimation plot in **Supplementary Figure 4b**, we simulated different conditions of fold-change difference between two profiles, effect-window size and distance from bait. For each condition, we simulated increasing coverage and asked what should be the minimum molecule coverage in order to achieve $P' < 0.05$. For example, the analysis demonstrated that detecting two-fold enrichment over the background for a 3 kb window requires a UMI-4C profile with as little as 7,000 molecules for the 100 kb range, but around 100,000 molecules if the element is located 1 Mb from the bait (**Supplementary Fig. 4b**).

Quantitative analysis of genomic intervals between profiles. First, the coverage of the profiles is normalized, as described above, to a single reference profile (DND41 in **Figs. 3** and **4**; **Supplementary Figs. 7** and **8**). Then, molecule counts for manually selected genomic intervals of sizes 3–5 kb are aquired (**Supplementary Table 3**). The mean interval count is calculated as the total number of molecules divided by the total number of potential fragment ends in the interval. The error bars report the s.d. of the mean estimator which is, as described above, approximately the square root of the window sum of molecules divided by the size of the window. Prior to visualization, the mean values of each interval are divided by the mean of the reference interval. For each sample the molecule count inside the interval against the number of molecules outside the interval (but within 1 Mb from the bait) is compared to derive a χ^2 P-values.

Validation of multiplex UMI-4C. Two independent UMI-4C templates (200 ng DNA from each) were amplified for multiplexed library construction with baits −165 kb, −50 kb, −3.5 kb and *ANK1*. The two multiplexed profiles were pooled and compared to the single-bait corresponding profiles.

Gene expression data. Microarray gene expression data was downloaded from the CCLE project[40]. The values are the mean intensities for all probes that map to each gene.

Code availability. Software package used is this study is available at: https://bitbucket.org/tanaylab/umi4cpackage.

35. Porcu, M. *et al.* Mutation of the receptor tyrosine phosphatase PTPRC (CD45) in T-cell acute lymphoblastic leukemia. *Blood* **119**, 4476–4479 (2012).
36. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134 (2012).
37. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
38. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).
39. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
40. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

doi:10.1038/nmeth.3922