nature biotechnology

High-throughput genome scaffolding from *in vivo* DNA interaction frequency

Noam Kaplan & Job Dekker

Despite advances in DNA sequencing technology, assembly of complex genomes remains a major challenge, particularly for genomes sequenced using short reads, which yield highly fragmented assemblies[1–3]. Here we show that genome-wide *in vivo* chromatin interaction frequency data, which are measurable with chromosome conformation capture–based experiments, can be used as genomic distance proxies to accurately position individual contigs without requiring any sequence overlap. We also use these data to construct approximate genome scaffolds *de novo*. Applying our approach to incomplete regions of the human genome, we predict the positions of 65 previously unplaced contigs, in agreement with alternative methods in 26/31 cases attempted in common. Our approach can theoretically bridge any gap size and should be applicable to any species for which global chromatin interaction data can be generated.

In genome assembly, massive amounts of short DNA sequencing reads can be assembled into sets of small contigs, but joining these contigs into scaffolds, a process known as scaffolding, is often difficult owing to the presence of repetitive sequences[4,5]. Currently, high-throughput scaffolding is based on <40 kb long-insert paired-end read libraries. Improving the degree of completion of genome sequences typically relies on low-throughput or laborious methods such as fluorescence immunofluorescence *in situ* hybridization (FISH)[6–9], bacterial artificial chromosome (BAC)-based sequencing[10]. Although the advancement of sequencing technology is producing longer reads and thus increasing the size of contigs, recent assessments of genome assemblers[11,12] show that complex genome assemblies, which rely only on sequencing data, are still highly ambiguous and fragmented, owing to gap sizes beyond that of long-insert molecules. In fact, even in the human genome, despite the massive effort invested in its completion, ~30 Mb of human euchromatic DNA remains unassembled[9]. Thus, high-throughput sequencing and genome assembly technology have reached a point at which an increase in the number of short reads does not substantially improve assembly quality.

Hi-C is an experimental technique that measures the *in vivo* spatial interaction frequency between chromatin segments over the whole genome, by cross-linking loci that are in close physical proximity and quantifying them with high-throughput, paired-end sequencing[13]. Every uniquely mapped paired-end read indicates an interaction between two genomic loci, so that the number of read pairs that map to distant DNA fragments can be treated as an unnormalized interaction frequency. Notably, all Hi-C experiments in eukaryotes to date have shown, in addition to species-specific and cell type–specific chromatin interactions, two canonical interaction patterns. One pattern, distance-dependent decay (DDD), is a general trend of decay in interaction frequency as a function of genomic distance. The second pattern, *cis-trans* ratio (CTR), is a significantly higher interaction frequency between loci located on the same chromosome, even when separated by tens of megabases of sequence, versus loci on different chromosomes[13–18]. These patterns may reflect general polymer dynamics, where proximal loci have a higher probability of randomly interacting[19], as well as specific nuclear organization features such as the formation of chromosome territories, the phenomenon of interphase chromosomes tending to occupy distinct volumes in the nucleus with little mixing[20]. Although the exact details of these two patterns may vary between species, cell types and cellular conditions, they are ubiquitous and prominent. In fact, these patterns are so strong and consistent that they are used to assess experiment quality and are usually normalized out of the data in order to reveal detailed interactions[14,15,21].

Here we propose that genome assembly technology can take advantage of the three-dimensional structure of genomes. We show that the features which make the canonical Hi-C interaction patterns a hindrance for the analysis of specific looping interactions, namely their ubiquity, strength and consistency, make them a powerful tool for estimating the genomic position of contigs.

We first use the CTR pattern to tackle the problem of scaffold augmentation, in which most of the genome is assumed to be correctly assembled and the challenge is to predict both the chromosome and locus of an unplaced contig, based on its pattern of interaction with the placed contigs. This is the situation for the majority of published 'finished' complex genomes, including human and mouse. Because most of the genome is assembled, it is possible to observe, quantify and computationally model the DDD and CTR interaction patterns, even if they are genome-specific or condition-specific. This model can then be used to estimate the positions of new contigs. Prior knowledge of the canonical patterns for a particular species is not needed.

As an initial test, we performed simulations on the human genome hg19 assembly[22] and a previously published Hi-C data set[16] obtained from male H1 embryonic stem cells (ESCs). To demonstrate the robustness of our approaches when using a relatively low number of reads,

Program in Systems Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. Correspondence should be addressed to N.K. (noam.kaplan2@gmail.com) or J.D. (job.dekker@umassmed.edu).

Received 3 August; accepted 8 November; published online 24 November 2013; doi:10.1038/nbt.2768

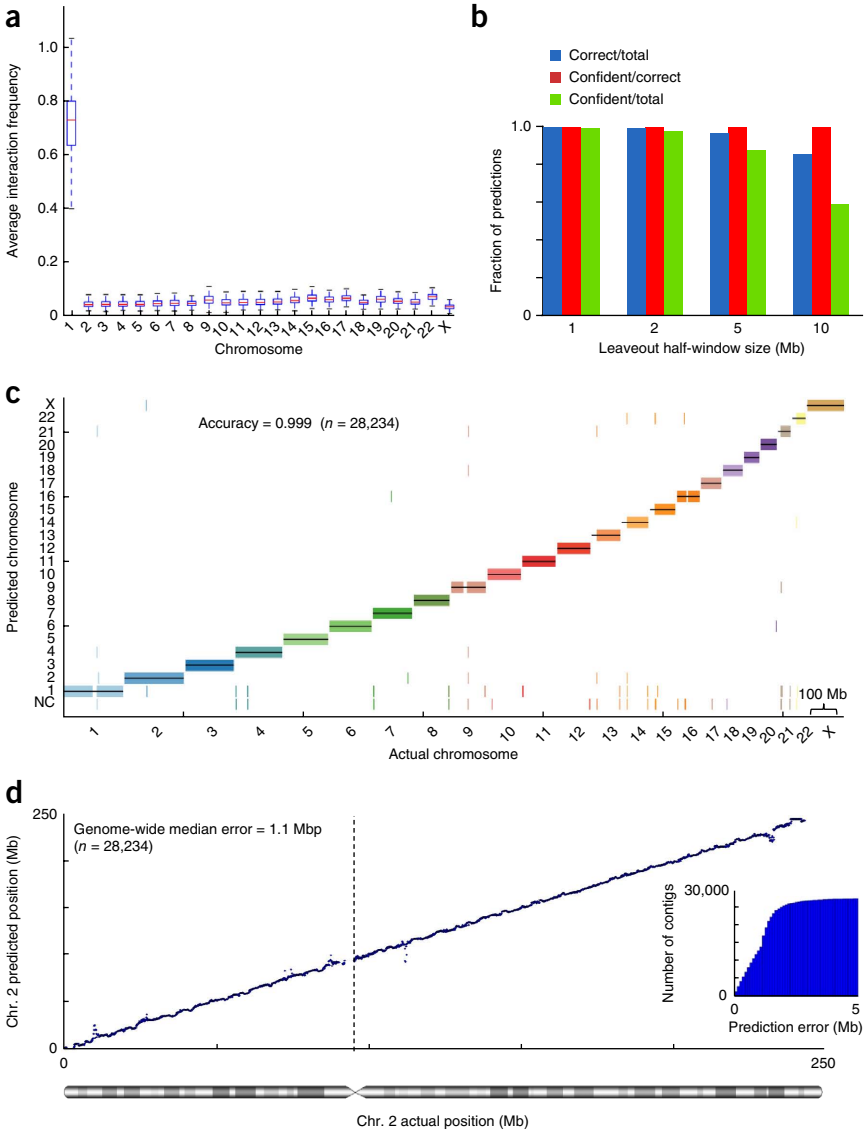

Figure 1 Interaction frequency accurately predicts chromosome and locus for scaffold augmentation. (**a**) Average interaction frequency strongly separates interchromosomal from intrachromosomal interactions. For each 100-kb contig in chromosome 1, we calculate its average interaction frequency with each chromosome. We exclude interaction data from the contig's 1-Mb regions on each side, where the strongest interaction frequencies are typically found. The box plot shows the distribution of average interaction frequencies of all contigs over all chromosomes and demonstrates that the distribution of interchromosomal interaction frequencies is separated from intrachromosomal interaction frequencies. Whiskers represent minimal and maximal points within 1.5 of the interquartile range. (**b**) Naive Bayes predictive performance at various gap sizes. We trained a naive Bayes classifier and predicted the chromosome of each contig, leaving out a 1-, 2-, 5- or 10-Mb flanking region on each side of the contig. Confident predictions are predictions with a posterior probability of at least 0.2. (**c**) Genome-wide view of naive Bayes predictive performance. The prediction for each contig is marked by a short vertical line, colored according to its true chromosome. Predictions showed were performed leaving out a 1-Mb flanking region on each side of the contig. Predictions that did not pass the confidence threshold are marked as "NC". (**d**) Interaction frequencies accurately predict chromosomal locus. For every contig, we exclude interaction data from the contig's 1-Mb flanking regions on each side and then predict its location in cross-validation. The inset shows the cumulative distribution of the absolute prediction error. All statistics are genome-wide.

we chose to use only a third of the Hi-C reads available for this cell type in the data set. We first quantified the CTR pattern by partitioning the human genome into 100-kb bins, each representing a large virtual contig, and calculated for each placed contig its average interaction frequency with each chromosome. To simulate a more difficult scenario and evaluate localization over long ranges, we omitted from this statistic the interaction data of the contig with its flanking 1 mb on each side, where the strongest Hi-C interaction signals are present. Then, we asked how well this statistic separates interchromosomal interactions from intrachromsomal interactions (**Fig. 1a**). We found that the average interaction frequency strongly separates inter- from intrachromosomal interactions, with an average area under the curve (AUC) of 0.9998, suggesting this statistic is highly predictive of which chromosome a contig belongs to. Next, we trained a simple multiclass model, a naive Bayes classifier, to predict the chromosome of each contig based on its average interaction frequency with each chromosome (Online Methods). To test the classifier, for each contig in the genome, we removed the interaction data for the contig and a flanking region of 1, 2, 5 or 10 Mb on each side, and used the classifier to predict the position of the contig solely from Hi-C data (**Fig. 1b**,**c**), achieving a genome-wide accuracy of 0.998 when leaving out 1 Mb on each side. By thresholding the associated posterior probabilities for each prediction output by the classifier to identify high-confidence predictions, we find that at a threshold of $P > 0.2$ the classifier can achieve a near-constant error rate of <0.005 even when leaving 10-Mb gaps on each side of the contig (100 times the size of the contig). We conclude that the CTR interaction pattern can be used to accurately predict to which chromosome an unplaced contig belongs, even if it is flanked by large gaps.

Next we sought to predict the genomic locus along a chromosome of an unplaced contig, given its chromosome and interaction pattern with placed contigs on the chromosome. We used the assembled portion of the genome to fit a probabilistic single-parameter exponential decay model describing the relationship between Hi-C interaction frequency and genomic distance (the DDD pattern). We removed in turn each contig from the chromosome, along with a flanking region of 1 Mb on each side, for the reasons mentioned previously, and estimated its most likely position by given its interaction profile and the decay model (**Fig. 1d**). We quantified the prediction error as the absolute value of the distance between the predicted position and the actual position. Our results show a cross-validated, genome-wide median error of 1.1 Mb. Additionally, 89.5% of the contigs are placed within 2 Mb of their actual position and 24.0% are within 0.5 Mb of their actual position (**Fig. 1d**, inset). We conclude that the DDD interaction pattern can be used to accurately predict the position of an unlocalized contig.

To show the utility of our approach for improving finished genomes, we collected two sets of contigs from hg19 (ref. 22) and HuRef[7],

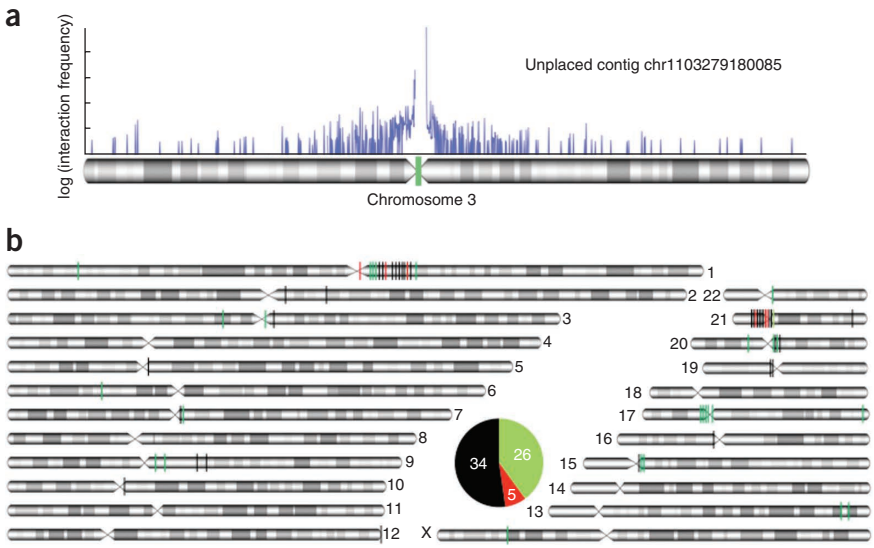

Figure 2 Scaffold augmentation of the human genome. (**a**) Interaction frequency data of an unplaced contig (chr1103279180085) with its predicted chromosome. Green bar marks the predicted contig position. (**b**) Predicted positions of unplaced contigs. Vertical lines indicate contigs. Green and red colors indicate agreement and disagreement, respectively, with previous predictions[9].

totaling 65 contigs (13.6 mb in total) that had sufficient Hi-C interaction data for further analysis and predicted their locations (**Fig. 2** and **Supplementary Table 1**). As validation, we compared our predictions to a recent study[9] that predicted the location of some of these contigs using extensive population SNP data to perform admixture mapping. Our predictions agree with the previous results for 26/31 (84%) of the contigs placed by both methods (Online Methods and **Supplementary Table 1**). In addition, 24/30 (80%) of our predictions were consistent with FISH localization measurements compiled in the same study. We conclude that our method can be used to increase the level of completion of complex genome assemblies by placing contigs that have proven difficult to assemble despite years of efforts, as in the case of the human genome.

Next, we explored whether Hi-C data could be used for *de novo* genome scaffolding. The challenge is to determine the karyotype (i.e., the number of chromosomes and the chromosomal assignment) and position of all contigs simultaneously based on their mutual interaction frequencies. *De novo* scaffolding is markedly more difficult than scaffold augmentation for two main reasons. First, as we have no knowledge of any contig positions, we cannot observe or fit the CTR and DDD functions. Instead, we must make assumptions regarding how interaction frequencies relate to genomic distance and hope that these crude approximations produce useful results. Second, instead of resolving only the distances of a single unplaced contig from an array of placed contigs, all distances between all contigs must be resolved jointly. Under most formulations of this problem, calculation of a global optimal solution cannot be guaranteed.

To examine scaffolding over long ranges, we simulated a large-gap scenario where we retained every tenth contig in the human genome so that we were left with an array of 100-kb virtual contigs separated by 900-kb gaps, thus omitting the bulk of the Hi-C signal. First, we asked whether it was possible to group all the contigs into their respective chromosomes *de novo* (*de novo* karyotyping). Assuming the DDD was approximately exponential, we transformed the matrix of interaction frequencies into approximate unscaled genomic distances (Online Methods). These distances are very crude approximations because at far distances the Hi-C interaction frequency, given as a discrete read number, will approach zero and thus will not distinguish between vastly different far distances. We applied standard average-linkage hierarchical clustering to the approximate distance matrix, and found that the maximal average cluster step (Online Methods) occurs at the point where 23 clusters are formed (**Fig. 3a**), demonstrating that the number of chromosomes is predicted

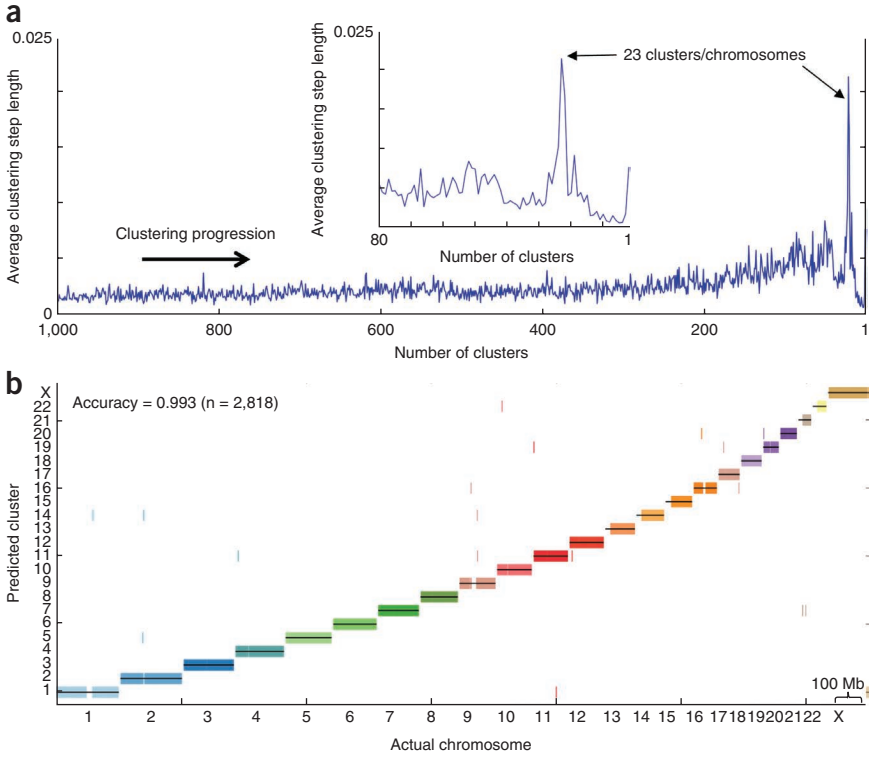

Figure 3 *De novo* karyotyping (chromosome assignment). We retained every tenth 100-kb contig in the genome, leaving 0.9-Mb gaps between contigs. We then transformed the interaction frequencies into approximate distances and applied standard average linkage hierarchical clustering to the approximate distance matrix, without using any prior knowledge regarding the positions of the contigs. (**a**) Average clustering step length along the final 1,000 clustering steps (inset shows final 80 steps). We find the maximum located at the point where 23 clusters remain. (**b**) The cluster assignment for each contig is marked by a short vertical line, colored according to its true chromosome.

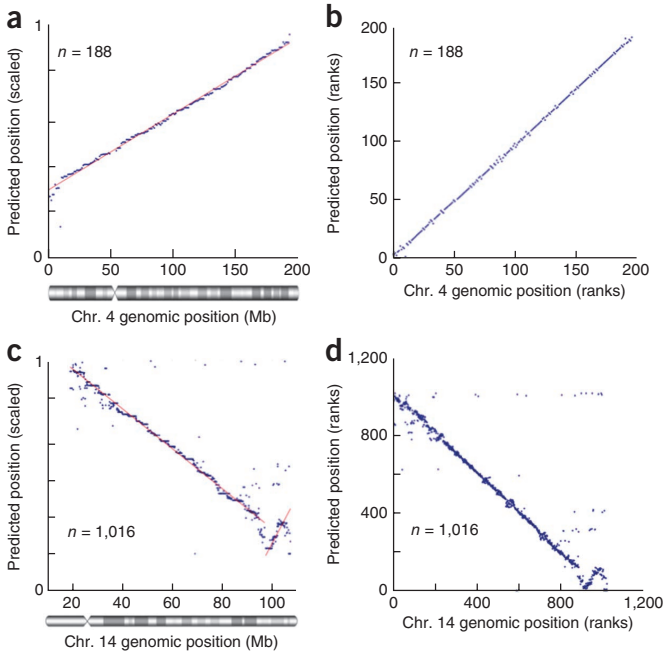

Figure 4 Accurate *de novo* chromosome scaffolding with interaction frequencies. (**a**) We retained every tenth 100-kb contig in the genome, leaving 0.9-Mb gaps between contigs. We then estimated the positions of all contigs, without using any prior knowledge regarding their positions. We arbitrarily scaled the predicted positions to the interval [0,1]. Note that the slope, which reflects scaling and orientation, is arbitrary. Scaled predicted contig positions versus actual contig positions on chromosome 4. (**b**) Ranks of predicted contig positions versus rank of actual contig positions. (**c**) *De novo* scaffolding applied to a real set of contigs from chromosome 14 (median contig size 20 kb). Shown are the scaled predicted contig positions versus actual contig positions. (**d**) Ranks of predicted contig positions versus rank of actual contig positions.

correctly. There was a high correspondence between the clusters and chromosomes; 99.3% of all contigs were placed on the correct chromosome (**Fig. 3b**).

Finally, we asked whether we could use interaction frequencies between unlocalized contigs to estimate their positions along a chromosome. We used a probabilistic model that assumes the DDD is approximately exponential, and attempted to find a set of likely contig positions for our simulated 100-kb virtual contigs (Online Methods). We arbitrarily scaled the predicted contig positions to range from 0 to 1. The predicted positions were highly consistent with their actual positions along most of the chromosomes (**Fig. 4a** and **Supplementary Fig. 1**). We estimate a median error rate of ~2 Mb and an error <10 Mb in ~93% of the predictions (**Supplementary Table 2**). As an alternative method of evaluation, we compared the ranks (contig order) of the predicted and actual positions (**Fig. 4b**). The ranked predictions seemed slightly more accurate than the predicted positions, with an estimated median rank error of 1 (**Supplementary Table 2**), possibly suggesting that the distances between neighboring contigs were distorted because of local variations in the DDD function. This was expected owing to the presence of locus-specific structures such as chromatin loops and structural domains[16,19,23]. Notably, our approach lays out an entire contiguous scaffold for each chromosome, rather than the highly fragmented scaffolds resulting from long-insert scaffolding. Most chromosomes contain no significant translocation or inversion errors, with a minority of chromosomes containing 1–2 major inversion errors.

We next applied *de novo* scaffolding to a previously published set of contigs from human chromosome 14 produced by the ALLPATHS-LG assembler[24] from actual sequencing libraries as part of the GAGE assembly[12] evaluation. We mapped Hi-C data to the assembled set of contigs (median contig size 20 kb), and estimated their chromosomal positions using our approach for *de novo* chromosome scaffolding (Online Methods). We then compared the predicted positions to the actual positions of the contigs when aligned to hg19 (**Fig. 4c**,**d**). The contigs were assembled into one large segment, containing one major inversion. Within each segment, the predicted positions were consistent with the actual positions. We estimated a median error of 976 kb with less than 10 Mb error in 96.4% of the predictions, and a median rank error of 6 (**Supplementary Table 3**). We conclude that the DDD pattern can be used to achieve accurate *de novo* chromosome scaffolding in various assembly scenarios, even without precise knowledge of the decay function and without the use of long-insert paired-end data.

In conclusion, we show that high-throughput *in vivo* genome-wide chromatin interaction data can be used to infer genomic location. We provide a conceptual framework with which additional experimental and computational strategies may be applied for further improvement (**Supplementary Discussion**). Although genome assembly methods have reached a point where increased sequencing depth does not improve assembly quality, our method importantly resuscitates the usefulness of additional sequence reads, as its resolution is largely a factor of the number of Hi-C reads. The power of our method may be attributed not to the sophistication of the computational tools, which are purposefully simple, but rather to two canonical interaction patterns. The fact that these patterns are strong, consistent across the genome, and ubiquitous in all species, cell types and conditions observed to date, suggests that this method is widely applicable. Finally, we have addressed only two out of several possible applications of Hi-C data for genome assembly, which include targeted assembly (for example, by using 4C[25,26] or 5C[27]), detection of assembly errors, resolution of nonunique genomic sequences and detection of chromosomal aberrations.

METHODS

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

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

ACKNOWLEDGMENTS

We thank B.R. Lajoie for help with processing of the Hi-C data. We thank the members of the Dekker Lab and G. Fudenberg for helpful discussions. This study is supported by the National Human Genome Research Institute (HG003143 to J.D.). N.K. is supported by a Long-Term Fellowship from the Human Frontier Science Program.

AUTHOR CONTRIBUTIONS

N.K. and J.D. conceived the strategy for genome assembly. N.K. performed all analyses and developed all computational approaches. N.K. and J.D. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nat. Rev. Genet.* **14**, 157–167 (2013).
2. Alkan, C., Sajjadian, S. & Eichler, E.E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).

3. Birney, E. Assemblies: the good, the bad, the ugly. *Nat. Methods* **8**, 59–60 (2011).
4. Baker, M. *De novo* genome assembly: what every biologist should know. *Nat. Methods* **9**, 333–337 (2012).
5. Schatz, M.C., Delcher, A.L. & Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20**, 1165–1173 (2010).
6. van den Engh, G., Sachs, R. & Trask, B.J. Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. *Science* **257**, 1410–1412 (1992).
7. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
8. Cheung, V.G. *et al.* Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–958 (2001).
9. Genovese, G. *et al.* Using population admixture to help complete maps of the human genome. *Nat. Genet.* **45**, 406–414 (2013).
10. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
11. Bradnam, K.R. *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**, 10 (2013).
12. Salzberg, S.L. *et al.* GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **22**, 557–567 (2012).
13. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
14. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).
15. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010).
16. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
17. Zhang, Y. *et al.* Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**, 908–921 (2012).
18. Moissiard, G. *et al.* MORC family ATPases required for heterochromatin condensation and gene silencing. *Science* **336**, 1448–1451 (2012).
19. Dekker, J., Marti-Renom, M.A. & Mirny, L.A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14**, 390–403 (2013).
20. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harb. Perspect. Biol.* **2**, a003889 (2010).
21. Sanyal, A., Lajoie, B., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
22. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
23. Nora, E.P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
24. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**, 1513–1518 (2011).
25. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38**, 1341–1347 (2006).
26. 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).
27. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).

ONLINE METHODS

Data sets. Throughout the paper we use human genome assembly hg19 (ref. 22). Unplaced contigs were taken from hg19 and HuRef[7].

For Hi-C data, we used a previously published data set[16]. Specifically, we used a third of the reads available for human H1 ESC, totaling ~248 M unique mapped reads, except when stated otherwise. As described previously[28], Hi-C reads were mapped using a customized pipeline, filtered for restriction fragment PCR amplifications, dangling ends and self-circles, and summed in 100-kb nonoverlapping bins. Oversequenced bins were removed and Sinkhorn-Knopp balancing was applied to the interaction matrices to correct for bin-specific biases. We note, however, that the matrix correction did not have a significant effect on the results. Hi-C matrices are available at: http://my5c.umassmed.edu/triangulation/.

All analysis was implemented in Python, mainly using Scipy and scikit-learn[29] modules. Code was deposited in a public repository at: https://github.com/NoamKaplan/dna-triangulation/ and is available as **Supplementary Data**.

Scaffold augmentation: chromosome prediction. For each contig, we calculated its mean interaction frequency with each chromosome and used this statistic to quantify the CTR pattern. Each contig can then be associated with a vector $(a_1, \ldots, a_{23})$ representing its mean interaction frequency with each chromosome. This naturally fits into a multiclass classification framework, where we fit a function that maps such vectors to one of 23 chromosomes.

We chose to train a simple naive Bayes classifier on the data. For the observed variables $a_1, \ldots, a_{23}$, the naive Bayes classifier assumes these variables are conditionally independent given the chromosome variable c. To make predictions, the naive Bayes classifier then calculates the posterior probability:

$$p(c \mid a_1,\ldots,a_{23}) = \frac{p(c)\Pi_{i=1}^{23} p(a_i \mid c)}{Z}$$

where $p(c)$ is a prior probability proportional to the number of contigs in each chromosome, $p(a_i|c)$ is a conditional multinomial distribution and Z is a normalization factor. The chromosome with the highest posterior probability is selected.

All predictions were performed using a cross-validation scheme. Each contig was left out along with a flanking region, a model was trained on the remaining data and predictions were made on the left out contig. This process was repeated for all contigs.

We use the associated posterior probabilities as a measure of confidence. We define all predictions with posterior probability ≥0.2 to be confident predictions.

To quantify the separation between the *cis* and *trans* mean interaction frequencies, we used the standard AUC (area under ROC curve) metric. An AUC of 1 for a given chromosome means that when ranking all contigs by their mean interaction frequencies, the ranks of all contigs from that chromosome are higher than those of contigs from other chromosomes, whereas an AUC of 0.5 is the value expected by randomly shuffling the ranks.

Scaffold augmentation: locus prediction. One can view the pairwise interaction matrix resulting from a Hi-C experiment as the result of a random process where n interactions are sampled, with repetition, from the set of all m^2 possible pairwise interactions. This process is described exactly by a multinomial distribution, where each possible interaction is associated with its own probability. The probability mass function of the multinomial distribution in this context is given by:

$$p(x_{1,1},\ldots,x_{m,m}) = \frac{n!}{\Pi_{i=1}^{m}\Pi_{j=1}^{m}(x_{i,j}!)}\prod_{i=1}^{m}\prod_{j=1}^{m} p_{i,j}^{\,x_{i,j}}$$

where $x_{i,j}$ is the number of times an interaction between contigs i and j is observed, n is the total number of interactions sampled and $p_{i,j}$ is the probability associated with the interaction between contigs i and j.

However, in the case of Hi-C this would require m^2 parameters $p_{i,j}$ to model the data. Since m is equal to the total number of contigs, this number will be very large. Thus, it is useful to introduce constraints on these probabilities. Specifically, we want these constraints to formally represent an assumption regarding how the interaction probability is associated with the genomic distance between loci, with a small number of parameters. We assume that the probability of two loci interacting decays approximately exponentially with their distance. So, formally, we would like to define the probability of sampling an interaction between two contigs at positions s_i and s_j as:

$$p_{s_i,s_j} = e^{\alpha|s_i - s_j|}$$

where $\alpha < 0$ is a scale parameter. However, we must add a normalization term to ensure that the sum of probabilities over all possible interactions is 1, so we define the final probabilities as:

$$p_{s_i,s_j} = \frac{e^{\alpha|s_i-s_j|}}{Z} = \frac{e^{\alpha|s_i-s_j|}}{\sum_{k=1}^{m}\sum_{l=1}^{m} e^{\alpha|s_k-s_l|}}$$

resulting in the probabilistic model:

$$p(x_{1,1},\ldots,x_{m,m}) = \frac{n!}{\Pi_{i=1}^{m}\Pi_{j=1}^{m}(x_{i,j}!)}\prod_{i=1}^{m}\prod_{j=1}^{m}\left(\frac{e^{\alpha|s_i-s_j|}}{\sum_{k=1}^{m}\sum_{l=1}^{m} e^{\alpha|s_k-s_l|}}\right)^{x_{i,j}}$$

Thus, we parameterize the probabilities of the multinomial function with a much smaller number of parameters, namely the position of each contig and the single scale parameter (a total of $m + 1$ parameters). This provides us with a precise probabilistic model that formalizes our knowledge and assumptions regarding how the experimental data are generated.

Next, we can estimate the parameters of the model to using the maximum-likelihood approach. The log-likelihood for a multinomial distribution is given by:

$$L(p_{1,1},\ldots,p_{m,m} \mid x_{1,1},\ldots,x_{m,m}) = \log(n!) - \sum_{i=1}^{m}\sum_{j=1}^{m}\log(x_{i,j}!) + \sum_{i=1}^{m}\sum_{j=1}^{m} x_{i,j}\log p_{i,j}$$

Given a data set, the first two terms are constant and can thus be omitted as they will not affect the solution.

We then obtain the final function for optimization:

$$Q(x_{1,1},\ldots,x_{m,m}) = \sum_{i=1}^{m}\sum_{j=1}^{m} x_{i,j}\log p_{i,j} = \sum_{i=1}^{m}\sum_{j=1}^{m} x_{i,j}\alpha\,|s_i - s_j| - n\log\left(\sum_{i=1}^{m}\sum_{j=1}^{m} e^{\alpha|s_i-s_j|}\right)$$

To estimate the position of an unplaced contig μ, we first use the positions of the known contigs to estimate the negative scaling parameter α that maximizes Q given the known positions and observed reads. Next, we find the position by estimating s_μ that maximizes:

$$Q_\mu(x_{1,\mu},\ldots,x_{m,\mu}) = \sum_{i=1}^{m} x_{i,\mu}\alpha\,|s_i - s_\mu| - n_\mu\log\left(\sum_{i=1}^{m} e^{\alpha|s_i-s_\mu|}\right)$$

where n_μ is the sum of interactions with μ.

All predictions were performed using a cross-validation scheme as explained previously.

Augmenting the human genome. We compiled a set of 65 unplaced human contigs from hg19 (ref. 22) and HuRef[7], totaling 13.6 Mb. We then binned each contig into 100-kb bins. To predict a contig's position, we separately predicted the position of each of its bins. Bin predictions were generally in good agreement, but if the bins were mapped to multiple chromosomes, we chose the chromosome assignment to be that of the bin with the highest posterior probability and also chose this bin for locus prediction.

We compared our predictions to those made by population admixture mapping[9] and to FISH measurements[7–9]. As these data were generally limited to chromosomal cytoband resolution or lower, we counted prediction agreements as those that are on the same chromosome and within the same region (<= 2 cytobands for 24/26 contigs). We did not observe any cases in which the

doi:10.1038/nbt.2768

chromosome prediction agreed but the location within the chromosome was in strong disagreement.

***De novo* karyotyping.** We transformed Hi-C interaction frequency data into approximate distances by adding 1 and taking the log, and then flipping by subtracting the data from its maximum. Next, we applied average linkage hierarchical clustering to the approximate distance matrix. This clustering scheme starts with singleton clusters, and at each clustering step the two closest clusters are merged to form a new cluster, where distance between clusters is defined as the average distance between all intercluster pairs:

$$d(C_1, C_2) = \frac{\sum_{i \in C_1, j \in C_2} D_{ij}}{|C_1||C_2|}$$

where C_1 and C_2 are clusters, $|C_i|$ is the cardinality of cluster C_i and D_{ij} is the approximate distance between contigs i and j.

To estimate the number of chromosomes, we sought an intrinsic measure that would indicate what a likely number of chromosomes may be. A natural measure of progression of the hierarchical clustering process is the average distance between the clusters being merged at each step. In this context, the clustering step length, defined as the difference between average distances associated with consecutive clustering steps, can be indicative of a stable partitioning[30]. As the clustering step length may be sensitive to noise, we derived a robust version of this metric by repetitively ($n = 20$) randomly sampling and clustering a random 80% of the data, and finally averaging the clustering step length. We thus examine the final 1,000 clustering steps, under the assumption that the actual number of chromosomes is less than 1,000 and find the maximal average cluster step. We estimate the number of chromosomes to be the number of clusters remaining at this point.

It is important to note that estimation of the correct number of clusters is not strictly required for the accuracy of the clusters themselves. As long as there is some point in the clustering tree where the clusters are highly accurate, any clustering up to that stage would also be accurate, so even if we overestimate the number of chromosomes it should not affect cluster quality.

***De novo* chromosome scaffolding.** For *de novo* chromosome scaffolding we use the same probabilistic model developed for augmentation locus prediction. Here we estimate both the parameters $s_1,\ldots,s_m$ (contig positions) and α (scaling of our probabilistic model) from the observed data:

$$\underset{\alpha, s_1, \ldots, s_m}{\operatorname{argmax}} Q(x_{1,1}, \ldots, x_{m,m})$$

Because we only assume that the interaction frequencies are inversely proportional to genomic distances, we arbitrarily restrict the positions $s_1,\ldots,s_m$ to the interval [0,1] and α to be negative.

To solve the optimization problem, we randomly initialize the parameters $s_1,\ldots,s_m$ in the interval [0,1] and α in the interval [0,−10], and then apply the L-BFGS numerical optimization algorithm[31], supplying it with the gradient of Q for speed. As the optimization problem is nonconvex, it can have many local optima and we cannot guarantee a globally optimal solution. Thus, each problem is run multiple times with different initializations. We found that solving this problem for a set of ~1,000 contigs produced by the ALLPATHS-LG assembler typically takes 1–5 min on a single CPU. For this scale of problem, ~2,000 iterations are typically sufficient to achieve a good solution. However, we note that further runs are expected to improve the solution (inferred from a simulation of a near-optimal solution, as estimated by initialization with the actual positions of the contigs).

Finally, we note that we defined $x_{i,j}$ not as the measured number of interactions but rather as the logarithm of the measured number of interactions plus 1. We found this transformation to be helpful, most likely due to suppression of multiplicative errors (we do not explicitly model noise) in the data and better numerical properties.

For *de novo* scaffolding of an actual set of contigs, we obtained the set of human chromosome 14 contigs assembled by ALLPATH-LG[24] in the GAGE evaluation[12]. Again we used human H1 ESC Hi-C data from Dixon *et al.*[16], but here we used all available reads (~750 M total, ~15 M unique mapped reads for this set of contigs). We next filtered contigs that do not match (95% identity for 95% of the sequence) a position on hg19 chromosome 14 and contigs with insufficient mapped data. The remaining set of 1,016 contigs has a median contig length of 20 kb. Because Hi-C signal grows, approximately, in direct linear proportion to the size of the contig, we normalized the Hi-C signal on each contig by its size.

Evaluation of *de novo* scaffolding is nontrivial, as several different metrics exist, each with its own strengths and weaknesses. We thus decided on multiple evaluation metrics, including a strict comparison of absolute positions. To compare positions, we needed to appropriately scale the predicted positions. We fit a regression line between the actual and predicted positions. If there was a major breakpoint (in most chromosomes there were none), such as an inversion, we fit the segments separately. In order to minimize the effect of outliers on the linear fit, we used an iterative fitting scheme where first the top and bottom 5% of the data are excluded from regression, and in following iterations data with residuals of more than 10 Mb are excluded. The resulting linear regression essentially provides a formula for transforming the predicted positions so that they can be compared with the actual positions. We proceeded to compare the rescaled predicted positions with the actual positions by calculating the median error and the fraction of errors larger than 10 Mb. It is important to note that these measures are likely overestimating the actual error, due to a couple of reasons. First, whereas visually the fits seem reasonable, they are clearly not optimal. Incorrect scaling can increase the error in the majority of contigs. Second, gap errors have a cumulative effect on these metrics. An incorrect gap size at a single point will effectively introduce errors in the absolute positions of all following points even though their relative distances and ordering are correct, so a few incorrectly predicted gap sizes could significantly increase the apparent error. Visual inspection of our results suggests that this may indeed be the case. As an alternative measure of performance, we compared the ranks (ordering) of predicted and actual positions. As here also a single mispredicted rank can distort all following ranks, we shifted the predicted ranks by an integer so that the median signed rank error would be zero. We then evaluated the results by calculating the median rank error, the fraction of rank errors greater than 10 and the fraction of correct neighbors (where actually neighboring contigs are predicted to be neighboring).

28. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
29. Pedregosa, F., Weiss, R. & Brucher, M. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
30. Kaplan, N., Friedlich, M., Fromer, M. & Linial, M. A functional hierarchical organization of the protein sequence space. *BMC Bioinformatics* **5**, 196 (2004).
31. Nocedal, J. Updating Quasi-Newton Matrices with Limited Storage. *Math. Comput.* **35**, 773–782 (1980).