

The Juicebox Assembly Tools module facilitates *de novo* assembly of mammalian genomes with chromosome-length scaffolds for under \$1000

Olga Dudchenko^{1,2,3,4}, Muhammad S. Shamim^{1,2,3,5†}, Sanjit S. Batra^{1,6†}, Neva C. Durand^{1,2,3}, Nathaniel T. Musial^{1,7}, Ragib Mostofa^{1,3}, Melanie Pham^{1,2,3}, Brian Glenn St Hilaire^{1,2,3}, Weijie Yao^{1,2,3}, Elena Stamenova^{1,8}, Marie Hoeger¹, Sarah K. Nyquist^{1,9}, Valeriya Korchina^{1,10}, Kelcie Pletch¹¹, Joseph P. Flanagan¹¹, Ania Tomaszewicz¹², Denise McAloose¹², Cynthia Pérez Estrada^{1,2,3}, Ben J. Novak¹³, Arina D. Omer^{1,2,3}, Erez Lieberman Aiden^{1,2,3,4,8*}

Affiliations:

¹The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030, USA.

²Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA.

³Department of Computer Science, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77030, USA.

⁴Center for Theoretical and Biological Physics, Rice University, Houston, TX 77030, USA.

⁵Medical Scientist Training Program, Baylor College of Medicine, Houston, TX 77030, USA.

⁶Department of Computer Science, University of California, Berkeley, CA 94720, USA.

⁷Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080, USA.

⁸Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA.

⁹Computational and Systems Biology, MIT, Cambridge, MA 02139, USA.

¹⁰DMU College of Osteopathic Medicine, Des Moines University, Des Moines, IA 50312, USA.

¹¹Houston Zoo, Houston, TX 77030, USA.

¹²Wildlife Conservation Society, New York, NY 10460, USA.

¹³Revive and Restore, Sausalito, CA 94965, USA.

*Correspondence to: erez@erez.com.

†These authors contributed equally to this work.

Hi-C contact maps are valuable for genome assembly (Lieberman-Aiden, van Berkum et al. 2009; Burton et al. 2013; Dudchenko et al. 2017). Recently, we developed Juicebox, a system for the visual exploration of Hi-C data (Durand, Robinson et al. 2016), and 3D-DNA, an automated pipeline for using Hi-C data to assemble genomes (Dudchenko et al. 2017). Here, we introduce “Assembly Tools,” a new module for Juicebox, which provides a point-and-click interface for using Hi-C heatmaps to identify and correct errors in a genome assembly. Together, 3D-DNA and the Juicebox Assembly Tools greatly reduce the cost of accurately assembling complex eukaryotic genomes. To illustrate, we generated *de novo* assemblies with chromosome-length scaffolds for three mammals: the wombat, *Vombatus ursinus* (3.3Gb), the Virginia opossum, *Didelphis virginiana* (3.3Gb), and the raccoon, *Procyon lotor* (2.5Gb). The only inputs for each assembly were Illumina reads from a short insert DNA-Seq library (300 million Illumina reads, maximum length 2x150 bases) and an *in situ* Hi-C library (100 million Illumina reads, maximum read length 2x150 bases), which cost <\$1000.

An accurate genome sequence is an essential basis for the study of any organism. To assemble a genome, a large number of DNA sequences derived from the organism of interest are overlapped with one another to create contiguous sequences, known as “contigs.” Next, linking information – derived from a wide variety of sources, such as mate-pairs, physical maps, and read-clouds – is used to order and orient these contigs into “scaffolds.” Errors often arise throughout the assembly process. Contigs may mistakenly concatenate two sequences (a ‘misjoin’). Scaffolds can contain errors in contig order (a ‘translocation’) or orientation (an ‘inversion’). Examples of such errors can be found in the best available reference genomes for many species (Robert B. Norgren 2013; Shearer et al. 2014; Tang et al. 2014; Chen et al. 2015; Davey et al. 2016; Utsunomiya et al. 2016; Schneider et al. 2017; Korlach et al. 2017). Consequently, inexpensive methods for identifying and correcting assembly errors are crucial for the generation of accurate assemblies (Salzberg and Yorke 2005; Phillippy, Schatz, and Pop 2008; Gnerre et al. 2009; Tsai, Otto, and Berriman 2010; Salzberg et al. 2012; Hunt et al. 2013; Gurevich et al. 2013; Bradnam et al. 2013; Simão et al. 2015; Fierst 2015; Muggli et al. 2015; Yuan et al. 2017; Harewood et al. 2017). Of course, improved error correction procedures can also reduce the amount of input data required, and thereby the cost of genome assembly.

Hi-C, a method for determining the 3D configuration of chromatin, is emerging as a valuable source of data for genome assembly (Lieberman-Aiden et al. 2009; Burton et al. 2013; Session et al. 2016; Peichel et al. 2016; Bickhart et al. 2017; Dudchenko et al. 2017; Mascher et al. 2017). When visualized, Hi-C data is typically represented as a heatmap. This heatmap is generated by partitioning a reference genome assembly into loci of fixed size; each heatmap entry indicates the frequency of contact between a pair of loci. When a chromosome is correctly assembled, sequences that are adjacent in the assembly are also in close physical proximity, leading to the appearance of a bright band of elevated contact frequency along the diagonal of the Hi-C heatmap. Conversely, when there are errors in a reference assembly, they are often visually obvious as anomalous patterns in the heatmap (Rao, Huntley et al. 2014; Harewood et al. 2017; Dudchenko et al. 2017; Lapp et al. 2017). Thus, in addition to its use as an input to automated assemblers, Hi-C can also facilitate the visual identification of errors in a genome assembly.

Recently, we introduced Juicebox, a set of tools that facilitate the visual exploration of Hi-C heatmaps across a wide range of scales (Durand, Robinson et al. 2016). Here, we introduce “Assembly Tools” (see Fig. 1), a new module in the Juicebox desktop application that extends the Juicebox interface in order to facilitate interactive genome assembly and reassembly using Hi-C data. Assembly Tools enables users to superimpose the positions of contigs or scaffolds in a reference assembly on top of the Hi-C heatmap, making assembly errors easier to find. When assembly errors are found, users can correct them, using a simple point-and-click interface, in a matter of seconds. Both the heatmap and the reference genome are updated in real-time to reflect these changes. Using Assembly Tools, users can improve genomes and reduce the cost of genome assembly.

To begin, a user needs to specify an assembly to be modified. Like the 3D-DNA algorithm, Assembly Tools uses a custom format, *.assembly*, that can be quickly generated from a *.fasta* file by an accompanying command-line tool. The user also needs relevant Hi-C data in the *.hic* format. In practice, this will often entail performing a Hi-C experiment in the organism of interest (Rao, Huntley et al. 2014), generating between 0.01X and 20X coverage, and running Juicer (Durand, Shamim et al. 2016).

Once the assembly and Hi-C dataset have been loaded via a pull-down menu, the user can begin to identify and correct errors. For instance, a translocation typically manifests as an extremely bright bowtie motif pointing horizontally or vertically, whose midpoint corresponds to two loci that are proximate in the genome but lie far apart in the assembly (Rao, Huntley et al. 2014; Dudchenko et al. 2017; Harewood et al. 2017). By clicking-and-dragging, a user can highlight the desired genomic interval, and – with a single click – move the interval to the desired position in the assembly (see Fig. 2a). Similarly, an inversion error – when the sequence of bases in a genomic interval is reversed – often manifests as a bowtie parallel to the diagonal (Rao, Huntley et al. 2014; Dudchenko et al. 2017). By clicking at the center of this motif, users can invert the selected interval (see Fig. 2a). Finally, a misjoin typically manifests as a point along the diagonal of the Hi-C heatmap where the upper-right and lower-left quadrants are extremely depleted, reflecting the lack of physical proximity between the erroneously concatenated loci (Dudchenko et al. 2017). Such errors can be resolved by selecting the affected scaffold and clicking on the position of the misjoin. The scaffold is then split in two in the reference genome assembly, allowing the two resulting scaffolds to be separately manipulated until they are correctly placed (see Fig. 2a). In addition, a third, short scaffold, containing the misjoined sequence itself, is excised and relocated to the end of the reference genome assembly, where anomalous scaffolds are kept for future reference. The boundaries of superscaffolds, such as chromosomes or chromosome arms, can be indicated by clicking between two scaffolds when no interval is currently selected. To simplify the above correction process, the mouse prompt changes to indicate the operation that is possible at any given moment: a circular arrow for inversion; a straight arrow for translocation; scissors for misjoin excision; and an angle to introduce a superscaffold boundary (see Fig. 2a).

After each change to the reference genome assembly, the Hi-C heatmap that is being displayed by Juicebox is updated accordingly. Crucially, the Juicebox Assembly Tools module does not recalculate the *.hic* file storing the Hi-C heatmap at each step, a process which could take many hours (Durand, Shamim

et al. 2016). Instead, the new heatmap can be thought of as a rearrangement of the pixels in the old heatmap, permuting its rows and columns. The Assembly Tools module tracks this permutation, updating it each time a change is made to the reference genome assembly.

While using the Assembly Tools module, users can also continue to employ standard Juicebox functions; for instance, they can modify the color scale, or zoom in and out (Durand, Robinson et al. 2016). A user can save the current state of the genome assembly as a new *.assembly* file. When the user is finished, a simple script can be run from the command line in order to apply this assembly file to the original reference assembly *.fasta* file, producing a corresponding assembly sequence.

To illustrate the use of the Assembly Tools module, we re-examined data from a very recent study which assembled the genome of the band-tail pigeon (*Patagioenas fasciata*), the closest living relative of the extinct passenger pigeon (Murray et al. 2017). This assembly incorporated Illumina and *in vitro* Hi-C (Chicago), yielding a scaffold N50 of 20 Mb. We generated *in situ* Hi-C data for the band-tailed pigeon (239M read pairs, 66X coverage). We then ordered the extant scaffolds from largest to smallest, loaded them into Juicebox, and performed an interactive genome assembly, resulting in chromosome-length scaffolds (scaffold N50: 76 Mb). (See Figs. 2b and S1.)

Although the Juicebox Assembly Tools module can be used independently, it can also be used as a validation and refinement system for the output of our automated 3D-DNA pipeline, which uses Hi-C data to improve genome assemblies. By adding a manual validation and refinement step, reliable genome assemblies can often be generated using less input data, reducing the cost of *de novo* genome assembly.

To illustrate the use of 3D-DNA and Juicebox Assembly Tools in tandem, we developed a procedure for assembling mammalian genomes with chromosome-length scaffolds for under \$1000 (see Fig. 3a). Our procedure involves three steps, and can be performed by a single person in roughly 10 days. First, we generate a PCR-free short insert DNA-Seq library, and sequence 300 million paired-end Illumina reads (2x150 bases). This corresponds to roughly 30X coverage for a typical (3Gb) mammal. These reads are assembled into a draft assembly using the software package w2rap (B. Clavijo et al. 2017; B. J. Clavijo et al. 2017). Second, we generate an *in situ* Hi-C library (Rao, Huntley et al. 2014), and sequence 100 million paired-end Illumina reads, corresponding to roughly 10X coverage for a typical mammal, which are used to improve the draft assembly by providing both as inputs to 3D-DNA (Dudchenko et al. 2017). Finally, we validate and refine the improved assembly using Juicebox Assembly Tools. Note that this procedure does not require advance knowledge of the exact size of the mammalian genome, or of the number of chromosomes.

To confirm the accuracy of the resulting genomes, we used our procedure to generate a *de novo* assembly of a human genome, see Fig. 3b and S2. We took 300 million raw reads from the NA12878

dataset shared by the Genome in a Bottle Consortium (NIST NA12878 HG001 HiSeq 300x) and added 100 million reads from the GM12878 Hi-C library published in (Rao, Huntley et al. 2014, HIC001). The resulting assembly, hs-1k, contains 23 chromosome-length scaffolds which together span 88,735 contigs (contig N50: 36,914) and 2,399,853,403 sequenced bases, comprising 85.2% of the genome assembly. It also contains 1,169 small scaffolds, spanning 1,652 contigs (contig N50: 21,506) and 397,814,093 bases, comprising the remaining 14.1% of the assembly. These small scaffolds contain contigs that could not be positioned reliably using Hi-C data, typically because they were very short.

Comparison of hs-1k with the human genome reference, hg38, showed that the 23 chromosome-length scaffolds in hs-1k correctly corresponded to the 23 human chromosomes. Of the 37,074 scaffolds that were incorporated into chromosome-length scaffolds in hs-1k and that could be uniquely placed in hg38, 99.97% (comprising 99.99% of the sequenced bases) were assigned to the correct chromosome. Together, the chromosome-length scaffolds in hs-1k spanned 99.34% of the length and 82.43% of the sequence in the chromosome-length scaffolds of hg38.

Next we examined the accuracy of the ordering of these chromosome-length scaffolds. When pairs of draft scaffolds assigned to the same chromosome were examined, the order in hs-1k matched the order in hg38 in 99.86% of cases. For scaffolds that were adjacent in hs-1k, the order matched hg38 96.03% of the time, reflecting the fact that Hi-C data is less effective at determining fine structure order; when the two scaffolds were longer than 100kb, the rate increased to 99%. Similarly, the orientation of scaffolds in hs-1k matched the orientation in hg38 91.64% of the time, with the errors again arising mostly from short scaffolds.

It is interesting to compare hs-1k to the draft genome reported by the International Human Genome Sequencing Consortium (hg5) (Lander et al. 2001). The hs-1k genome assembly contains ~10% less sequence in chromosome-length scaffolds than hg5. By contrast, the fine structure order is considerably more accurate in hs-1k. For instance, 23.1% of 1-kilobase intervals are in the wrong orientation in hg5; for hs-1k, the value is 5.2%, a 4.5-fold decrease. Note, however, that hg5 was a draft genome, and subsequent finishing steps on each chromosome greatly improved its fine structure accuracy. (See Fig. 3b.)

It is also interesting to examine the effect of replacing 3D-DNA in the above assembly strategy with two other automated Hi-C-based assembly algorithms, Lachesis (Burton et al. 2013) and SALSA (Ghurye et al. 2017). When provided the same inputs that were used for hs-1k, SALSA, which is designed to work with long-read assemblies, did not meaningfully improve upon the input. Lachesis successfully anchored many of the contigs but did not provide an accurate chromosome-scale ordering. Consequently, subsequent refinement with Juicebox Assembly Tools proved unrealistic in both cases. (See Fig. 3b.)

Having validated the \$1000 genome assembly procedure, we implemented it in order to generate *de novo* assemblies of three mammals for which no assembly has been published to date: the common wombat, *Vombatus ursinus*, the Virginia opossum, *Didelphis virginiana*, and the common raccoon, *Procyon lotor* (see Fig. 4a). In each case, the result was a set of chromosome length scaffolds: for wombat, the procedure generated 7 chromosome-length scaffolds (vu-1k), spanning 83.7% of the sequenced bases, with a total contig length of 2.74Gb; for Virginia opossum, the procedure generated 11 chromosome-length scaffolds (dv-1k), spanning 79.9% of the sequenced bases, with a total contig length of 2.67Gb; and for raccoon, 19 chromosome-length scaffolds (pl-1k), spanning 77.6% of sequenced bases, with a total contig length of 1.94Gb. The new assemblies facilitate the study of karyotype evolution in marsupials and carnivores (see Figs. 4b, S3 and Supplementary table S2).

Taken together, these findings demonstrate that the procedure we describe can be reliably employed in order to generate *de novo* assemblies of mammalian genomes with chromosome-length scaffolds.

Strikingly, the cost of the *de novo* genome assembly strategy described above is comparable to the present cost of human genome resequencing, in which short insert size DNA-Seq reads from an individual are compared to the existing human reference genome. To achieve human genome resequencing for \$1000, Illumina introduced a strategy that generates up to 400 million paired-end DNA-Seq reads (2x150 bases) on a HiSeq X instrument (Illumina, Inc. 2016). The *de novo* genome assembly strategy described above uses extremely similar inputs, simply replacing 100 million of the 400 million paired-end DNA-Seq reads with *in situ* Hi-C reads.

Of course, the genome assemblies generated using the strategy we describe can be further improved. For example, the genomes are not “finished” (Consortium 2004): it would be valuable to incorporate additional sequence into the chromosome-length scaffolds to fill gaps, and to correct errors in the fine-scale ordering of small adjacent contigs. Finally, although we did not encounter this issue with the hs-1k assembly, it can sometimes be difficult to correctly orient genomic intervals separated by extremely large gaps, such as chromosome arms separated by very large centromeres. These issues can be partially alleviated by additional short read Illumina data. For instance, doubling the number of PE150 reads included in our marsupial assemblies led to a larger number of sequenced bases in chromosome length scaffolds (common wombat, vu-2k: 2.72Gb→2.87Gb; Virginia opossum, dv-2k: 2.67Gb→2.85Gb). More expensive data types, such as long-read DNA sequences, can be employed to further improve the genome assembly (see Fig. 3b). The above methods are compatible with all data types of which we are aware, and we provide examples for a variety of such use cases in Table 1 and Table S3.

Finally, we note that this methodology is not restricted to mammals, and can be applied successfully to many other clades. Depending on the size of the genome of interest, more or less input data may be required. Similarly, the approach could be used to generate personalized genomes in a clinical setting.

SOFTWARE AVAILABILITY AND DOCUMENTATION OF TOOL REVIEW. The Assembly Tools module is available as a part of the Juicebox data visualization system for Hi-C, which can be downloaded at aidenlab.org/juicebox. The code, which is available at <https://github.com/theaidenlab/juicebox> is open source, and is licensed under the MIT license. Genomes, datasets, tutorials and other procedures associated with this publication are available at aidenlab.org/assembly.

ACKNOWLEDGMENTS. We acknowledge the McDonnell Genome Institute at the Washington University School of Medicine for sharing the NA12878_prelim.2.1 contigs, which were used for one of the assemblies in Table 1. We thank André Soares for providing the data from the UCSC Paleogenomics Lab on behalf of the Murray et al. group. We acknowledge David Oehler and Jean Paré of the Wildlife Conservation Society and Andrea Lee, Jess Jimerson, Katie Plaeger and Erin Neer of the Houston Zoo for their help with sample collection. We thank Christine Molter, Maryanne Tociłowski, Lauren Howard and Judilee Marrow for veterinary work with the mammalian subjects. We also thank Chad Nusbaum and Andreas Gnirke for comments on the manuscript, and David Weisz, Alyssa Blackburn, Sheikh Russell for computational assistance. Finally, we thank Terry Leatherland, Grace Liu, Loic Fura and Victoria Nwobodo for access to a high RAM IBM E880 server. This work was supported by a Center for Theoretical Biological Physics postdoctoral fellowship to O.D., an NIH New Innovator Award (1DP2OD008540-01), an NSF Physics Frontiers Center Award (PHY-1427654, Center for Theoretical Biological Physics), the Welch Foundation (Q-1866), an NVIDIA Research Center Award, an IBM University Challenge Award, a Google Research Award, a Cancer Prevention Research Institute of Texas Scholar Award (R1304), a McNair Medical Institute Scholar Award, an NIH 4D Nucleome Grant U01HL130010, an NIH Encyclopedia of DNA Elements (ENCODE) Mapping Center Award UM1HG009375, the President's Early Career Award in Science and Engineering to E.L.A.

REFERENCES.

- Bickhart, Derek M., Benjamin D. Rosen, Sergey Koren, Brian L. Sayre, Alex R. Hastie, Saki Chan, Joyce Lee, et al. 2017. "Single-Molecule Sequencing and Chromatin Conformation Capture Enable *de Novo* Reference Assembly of the Domestic Goat Genome." *Nature Genetics* 49 (4):643–50. <https://doi.org/10.1038/ng.3802>.
- Bradnam, Keith R., Joseph N. Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, et al. 2013. "Assemblathon 2: Evaluating *de Novo* Methods of Genome Assembly in Three Vertebrate Species." *GigaScience* 2 (July):10. <https://doi.org/10.1186/2047-217X-2-10>.
- Burton, Joshua N., Andrew Adey, Rupali P. Patwardhan, Ruolan Qiu, Jacob O. Kitzman, and Jay Shendure. 2013. "Chromosome-Scale Scaffolding of *de Novo* Genome Assemblies Based on Chromatin Interactions." *Nature Biotechnology* 31 (12):1119–25. <https://doi.org/10.1038/nbt.2727>.
- Chen, Meili, Yibo Hu, Jingxing Liu, Qi Wu, Chenglin Zhang, Jun Yu, Jingfa Xiao, Fuwen Wei, and Jiayan Wu. 2015. "Improvement of Genome Assembly Completeness and Identification of Novel Full-Length Protein-Coding Genes by RNA-Seq in the Giant Panda Genome." *Scientific Reports* 5 (December):18019. <https://doi.org/10.1038/srep18019>.
- Clavijo, Bernardo, Gonzalo Garcia Accinelli, Jonathan Wright, Darren Heavens, Katie Barr, Luis Yanes, and Federica Di Palma. 2017. "W2RAP: A Pipeline for High Quality, Robust Assemblies of Large

- Complex Genomes from Short Read Data." *BioRxiv*, February, 110999.
<https://doi.org/10.1101/110999>.
- Clavijo, Bernardo J., Luca Venturini, Christian Schudoma, Gonzalo Garcia Accinelli, Gemy Kaithakottil, Jonathan Wright, Philippa Borrill, et al. 2017. "An Improved Assembly and Annotation of the Allohexaploid Wheat Genome Identifies Complete Families of Agronomic Genes and Provides Genomic Evidence for Chromosomal Translocations." *Genome Research* 27 (5):885–96.
<https://doi.org/10.1101/gr.217117.116>.
- Consortium, International Human Genome Sequencing. 2004. "Finishing the Euchromatic Sequence of the Human Genome." *Nature* 431 (7011):931. <https://doi.org/10.1038/nature03001>.
- Davey, John W., Mathieu Chouteau, Sarah L. Barker, Luana Maroja, Simon W. Baxter, Fraser Simpson, Richard M. Merrill, et al. 2016. "Major Improvements to the *Heliconius Melpomene* Genome Assembly Used to Confirm 10 Chromosome Fusion Events in 6 Million Years of Butterfly Evolution." *G3 (Bethesda, Md.)* 6 (3):695–708. <https://doi.org/10.1534/g3.115.023655>.
- Derjushcheva, Svetlana, Anna Kurganova, Felix Habermann, and Elena Gaginskaya. 2004. "High Chromosome Conservation Detected by Comparative Chromosome Painting in Chicken, Pigeon and Passerine Birds." *Chromosome Research* 12 (7):715.
<https://doi.org/10.1023/B:CHRO.0000045779.50641.00>.
- Dudchenko, Olga, Sanjit S. Batra, Arina D. Omer, Sarah K. Nyquist, Marie Hoeger, Neva C. Durand, Muhammad S. Shamim, et al. 2017. "De Novo Assembly of the *Aedes Aegypti* Genome Using Hi-C Yields Chromosome-Length Scaffolds." *Science (New York, N.Y.)* 356 (6333):92–95.
<https://doi.org/10.1126/science.aal3327>.
- Durand, Neva C., James T. Robinson, Muhammad S. Shamim, Ido Machol, Jill P. Mesirov, Eric S. Lander, and Erez Lieberman Aiden. 2016. "Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom." *Cell Systems* 3 (1):99–101.
<https://doi.org/10.1016/j.cels.2015.07.012>.
- Durand, Neva C., Muhammad S. Shamim, Ido Machol, Suhas S. P. Rao, Miriam H. Huntley, Eric S. Lander, and Erez Lieberman Aiden. 2016. "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments." *Cell Systems* 3 (1):95–98.
<https://doi.org/10.1016/j.cels.2016.07.002>.
- Fierst, Janna L. 2015. "Using Linkage Maps to Correct and Scaffold *de Novo* Genome Assemblies: Methods, Challenges, and Computational Tools." *Frontiers in Genetics* 6 (June).
<https://doi.org/10.3389/fgene.2015.00220>.
- Ghurye, Jay, Mihai Pop, Sergey Koren, Derek Bickhart, and Chen-Shan Chin. 2017. "Scaffolding of Long Read Assemblies Using Long Range Contact Information." *BMC Genomics* 18 (July).
<https://doi.org/10.1186/s12864-017-3879-z>.
- Gnerre, Sante, Eric S. Lander, Kerstin Lindblad-Toh, and David B. Jaffe. 2009. "Assisted Assembly: How to Improve a *de Novo* Genome Assembly by Using Related Species." *Genome Biology* 10 (August):R88. <https://doi.org/10.1186/gb-2009-10-8-r88>.
- Gnerre, Sante, Iain MacCallum, Dariusz Przybylski, Filipe J. Ribeiro, Joshua N. Burton, Bruce J. Walker, Ted Sharpe, et al. 2011. "High-Quality Draft Assemblies of Mammalian Genomes from Massively Parallel Sequence Data." *Proceedings of the National Academy of Sciences* 108 (4):1513–18.
<https://doi.org/10.1073/pnas.1017351108>.
- Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. "QUAST: Quality Assessment Tool for Genome Assemblies." *Bioinformatics (Oxford, England)* 29 (8):1072–75.
<https://doi.org/10.1093/bioinformatics/btt086>.
- Harewood, Louise, Kamal Kishore, Matthew D. Eldridge, Steven Wingett, Danita Pearson, Stefan Schoenfelder, V. Peter Collins, and Peter Fraser. 2017. "Hi-C as a Tool for Precise Detection and

- Characterisation of Chromosomal Rearrangements and Copy Number Variation in Human Tumours." *Genome Biology* 18 (June):125. <https://doi.org/10.1186/s13059-017-1253-8>.
- Hunt, Martin, Taisei Kikuchi, Mandy Sanders, Chris Newbold, Matthew Berriman, and Thomas D. Otto. 2013. "REAPR: A Universal Tool for Genome Assembly Evaluation." *Genome Biology* 14 (5):R47. <https://doi.org/10.1186/gb-2013-14-5-r47>.
- Jain, Miten, Sergey Koren, Josh Quick, Arthur Rand, Thomas Sasani, John Tyson, Andrew Beggs, et al. 2017. "Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads." *BioRxiv*, April, 128835. <https://doi.org/10.1101/128835>.
- Korlach, Jonas, Gregory Gedman, Sarah B. Kingan, Chen-Shan Chin, Jason T. Howard, Jean-Nicolas Audet, Lindsey Cantin, and Erich D. Jarvis. 2017. "De Novo PacBio Long-Read and Phased Avian Genome Assemblies Correct and Add to Reference Genes Generated with Intermediate and Short Reads." *GigaScience* 6 (10):1–16. <https://doi.org/10.1093/gigascience/gix085>.
- Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822):860–921. <https://doi.org/10.1038/35057062>.
- Lapp, S. A., J. A. Geraldo, J.-T. Chien, F. Ay, S. B. Pakala, G. Batugedara, J. Humphrey, et al. 2017. "PacBio Assembly of a Plasmodium Knowlesi Genome Sequence with Hi-C Correction and Manual Annotation of the SICAvir Gene Family." *Parasitology*, July, 1–14. <https://doi.org/10.1017/S0031182017001329>.
- Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 25 (14):1754–60. <https://doi.org/10.1093/bioinformatics/btp324>.
- Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science (New York, N.Y.)* 326 (5950):289–93. <https://doi.org/10.1126/science.1181369>.
- Illumina, Inc. 2016. "HiSeq X™ Series of Sequencing Systems. Maximum Throughput and Lowest Cost for Population-Scale Whole-Genome Sequencing." <https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf>.
- Love, R. Rebecca, Neil I. Weisenfeld, David B. Jaffe, Nora J. Besansky, and Daniel E. Neafsey. 2016. "Evaluation of DISCOVAR de Novo Using a Mosquito Sample for Cost-Effective Short-Read Genome Assembly." *BMC Genomics* 17:187. <https://doi.org/10.1186/s12864-016-2531-7>.
- Mascher, Martin, Heidrun Gundlach, Axel Himmelbach, Sebastian Beier, Sven O. Twardziok, Thomas Wicker, Volodymyr Radchuk, et al. 2017. "A Chromosome Conformation Capture Ordered Sequence of the Barley Genome." *Nature* 544 (7651):427. <https://doi.org/10.1038/nature22043>.
- Muggli, Martin D., Simon J. Puglisi, Roy Ronen, and Christina Boucher. 2015. "Misassembly Detection Using Paired-End Sequence Reads and Optical Mapping Data." *Bioinformatics* 31 (12):i80–88. <https://doi.org/10.1093/bioinformatics/btv262>.
- Murray, Gemma G. R., André E. R. Soares, Ben J. Novak, Nathan K. Schaefer, James A. Cahill, Allan J. Baker, John R. Demboski, et al. 2017. "Natural Selection Shaped the Rise and Fall of Passenger Pigeon Genomic Diversity." *Science* 358 (6365):951–54. <https://doi.org/10.1126/science.aao0960>.
- Nie, W, J Wang, W Su, D Wang, A Tanomtung, P L Perelman, A S Graphodatsky, and F Yang. 2012. "Chromosomal Rearrangements and Karyotype Evolution in Carnivores Revealed by Chromosome Painting." *Heredity* 108 (1):17–27. <https://doi.org/10.1038/hdy.2011.107>.
- Peichel, Catherine L., Shawn T. Sullivan, Ivan Liachko, and Michael A. White. 2016. "Improvement of the Threespine Stickleback (*Gasterosteus aculeatus*) Genome Using a Hi-C-Based Proximity-Guided Assembly Method." *BioRxiv*, August, 068528. <https://doi.org/10.1101/068528>.

- Phillippy, Adam M., Michael C. Schatz, and Mihai Pop. 2008. "Genome Assembly Forensics: Finding the Elusive Mis-Assembly." *Genome Biology* 9 (March):R55. <https://doi.org/10.1186/gb-2008-9-3-r55>.
- Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. "A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping." *Cell* 159 (7):1665–80. <https://doi.org/10.1016/j.cell.2014.11.021>.
- Robert B. Norgren, Jr. 2013. "Improving Genome Assemblies and Annotations for Nonhuman Primates." *ILAR Journal* 54 (2):144. <https://doi.org/10.1093/ilar/ilt037>.
- Robert S. Harris. 2007. "Improved Pairwise Alignment of Genomic DNA." PhD, The Pennsylvania State University. http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf.
- Salzberg, Steven L., Adam M. Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen, et al. 2012. "GAGE: A Critical Evaluation of Genome Assemblies and Assembly Algorithms." *Genome Research* 22 (3):557–67. <https://doi.org/10.1101/gr.131383.111>.
- Salzberg, Steven L., and James A. Yorke. 2005. "Beware of Mis-Assembled Genomes." *Bioinformatics* 21 (24):4320–21. <https://doi.org/10.1093/bioinformatics/bti769>.
- Schneider, Valerie A., Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, et al. 2017. "Evaluation of GRCh38 and *de Novo* Haploid Genome Assemblies Demonstrates the Enduring Quality of the Reference Assembly." *Genome Research* 27 (5):849–64. <https://doi.org/10.1101/gr.213611.116>.
- Session, Adam M., Yoshinobu Uno, Taejoon Kwon, Jarrod A. Chapman, Atsushi Toyoda, Shuji Takahashi, Akimasa Fukui, et al. 2016. "Genome Evolution in the Allotetraploid Frog *Xenopus Laevis*." *Nature* 538 (7625):336–43. <https://doi.org/10.1038/nature19840>.
- Shearer, Lindsay A., Lorinda K. Anderson, Hans de Jong, Sandra Smit, José Luis Goicoechea, Bruce A. Roe, Axin Hua, James J. Giovannoni, and Stephen M. Stack. 2014. "Fluorescence *In Situ* Hybridization and Optical Mapping to Correct Scaffold Arrangement in the Tomato Genome." *G3: Genes/Genomes/Genetics* 4 (8):1395–1405. <https://doi.org/10.1534/g3.114.011197>.
- Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31 (19):3210–12. <https://doi.org/10.1093/bioinformatics/btv351>.
- Tang, Haibao, Vivek Krishnakumar, Shelby Bidwell, Benjamin Rosen, Agnes Chan, Shiguo Zhou, Laurent Gentzbittel, et al. 2014. "An Improved Genome Release (Version Mt4.0) for the Model Legume *Medicago Truncatula*." *BMC Genomics* 15 (April):312. <https://doi.org/10.1186/1471-2164-15-312>.
- Tsai, Isheng J., Thomas D. Otto, and Matthew Berriman. 2010. "Improving Draft Assemblies by Iterative Mapping and Assembly of Short Reads to Eliminate Gaps." *Genome Biology* 11 (April):R41. <https://doi.org/10.1186/gb-2010-11-4-r41>.
- Utsunomiya, Adam T. H., Daniel J. A. Santos, Solomon A. Boison, Yuri T. Utsunomiya, Marco Milanesi, Derek M. Bickhart, Paolo Ajmone-Marsan, et al. 2016. "Revealing Misassembled Segments in the Bovine Reference Genome by High Resolution Linkage Disequilibrium Scan." *BMC Genomics* 17 (1). <https://doi.org/10.1186/s12864-016-3049-8>.
- Weisenfeld, Neil I., Shuangye Yin, Ted Sharpe, Bayo Lau, Ryan Hegarty, Laurie Holmes, Brian Sogoloff, et al. 2014. "Comprehensive Variation Discovery in Single Human Genomes." *Nature Genetics* 46 (12):1350–55. <https://doi.org/10.1038/ng.3121>.
- Yuan, Yuxuan, Philipp E. Bayer, Armin Scheben, Chon-Kit Kenneth Chan, and David Edwards. 2017. "BioNanoAnalyst: A Visualisation Tool to Assess Genome Assembly Quality Using BioNano Data." *BMC Bioinformatics* 18 (June):323. <https://doi.org/10.1186/s12859-017-1735-4>.

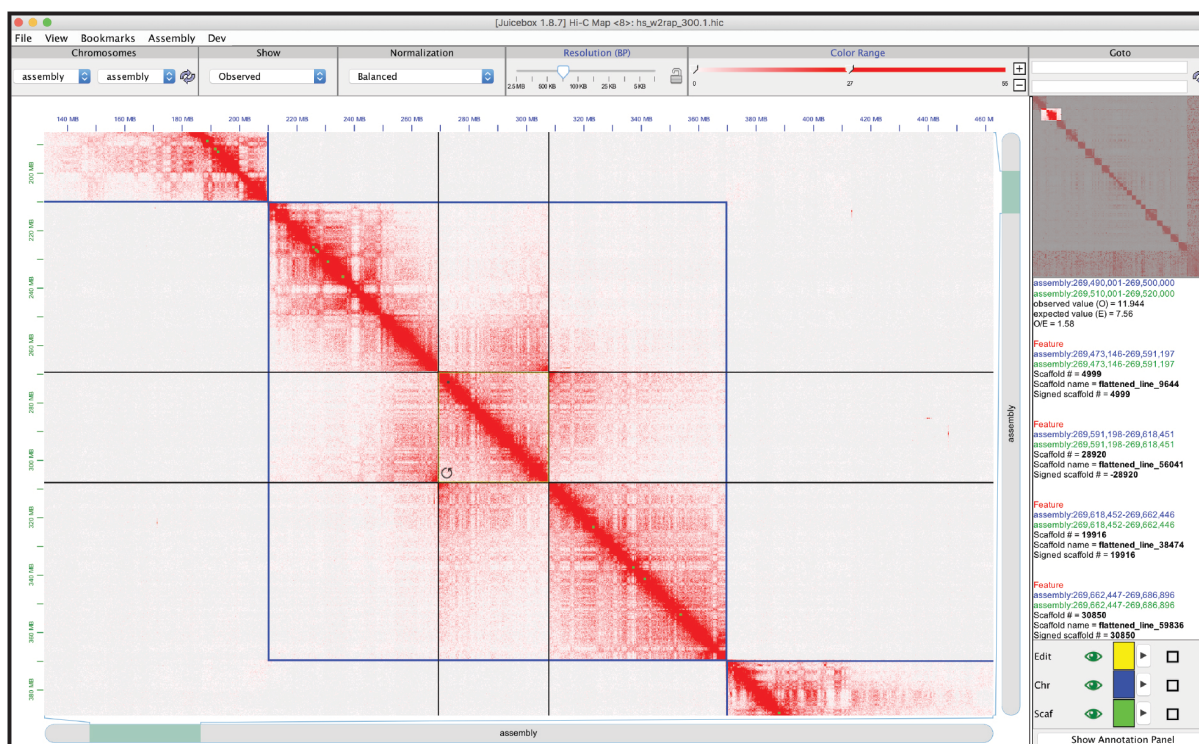
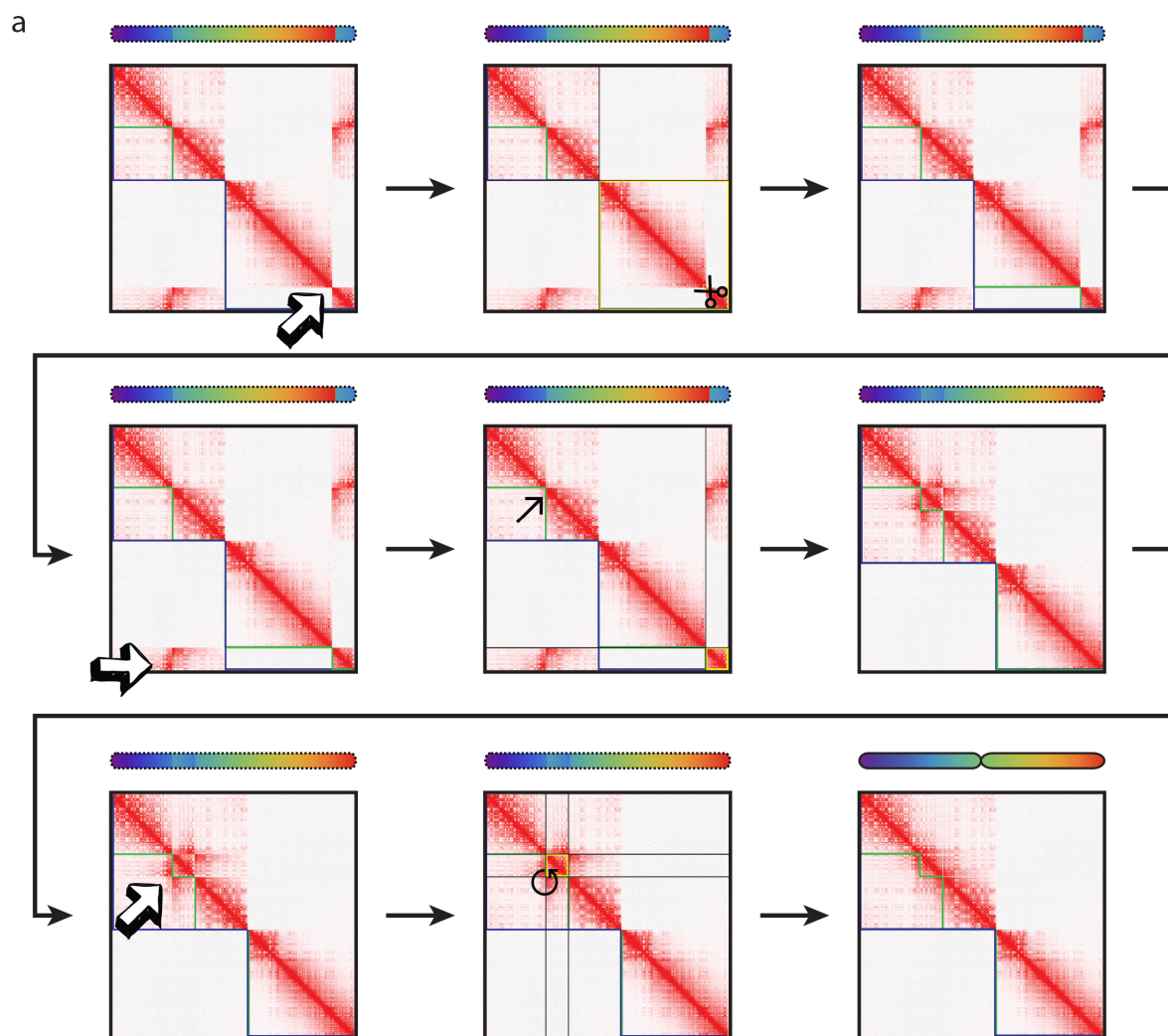


Figure 1: Juicebox Assembly Tools module enables visualization and interactive manipulation of genome assemblies. A screenshot of Juicebox Assembly Tools module zoomed in to 100-kb resolution on a region of a human genome assembly. A mouse prompt for inversion is shown appearing in the lower left of the selected genomic interval. While using the Assembly Tools, users can continue to employ standard Juicebox functions. The toolbar at the top allows users to quickly navigate between different views, normalizations, and resolutions as well as to load and save assembly files. At the top right, a mini-map shows the whole chromosome at low resolution. Below, hover text shows data for scaffolds in the selected genomic interval (yellow and black highlight). Two-dimensional features representing scaffold and chromosome boundaries are superimposed on the main map. Their appearance can be modified using Annotation panel in the lower right.



b



Band-tailed pigeon (*Patagioenas fasciata*)

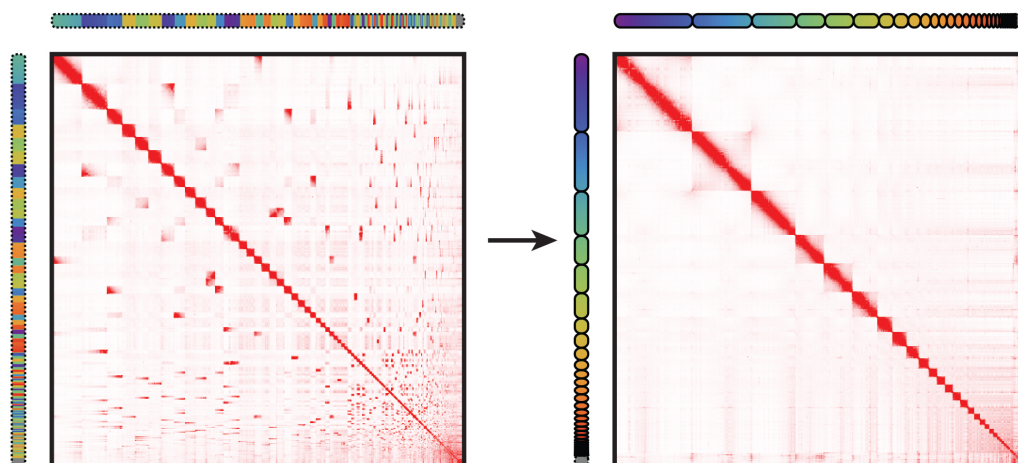
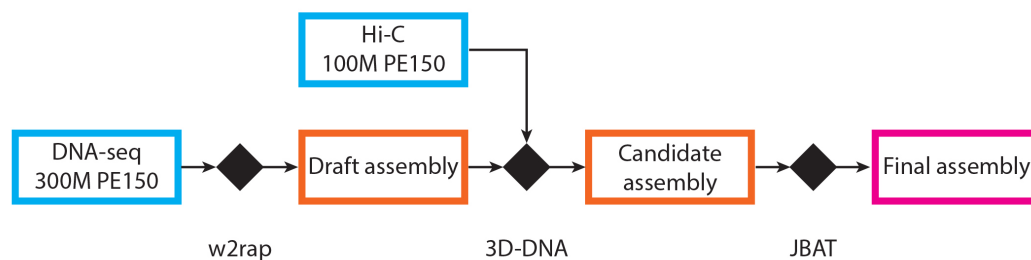


Figure 2: Hi-C maps allow for visual identification and interactive correction of errors in genome assemblies. (a) Here we show a contact matrix generated by aligning a GM12878 Hi-C data set (HIC001 library from (Rao, Huntley et al. 2014)) to a simulated genome assembly containing several errors. Each ‘pixel’ in the map indicates the frequency of contact between a pair of loci in the assembly. As the original assembly is being modified interactively using Juicebox Assembly Tools, the changes are reflected in the heatmap. This process continues until no more anomalous signals can be found on the heatmap. The simulated assembly was created by deliberately introducing errors into the sequence of two chromosome-length scaffolds from hg19 (chromosomes 2 and 4). The position of the loci according to hg19 is shown using chromograms. For the purpose of illustration, gaps have been removed from hg19 sequence. Anomalies in the Hi-C heatmap associated with 3 types of misassemblies (misjoin, translocation and inversion) are indicated by hand-drawn errors on the left-side panels. The interaction with Juicebox Assembly Tools (cut, paste and invert) and the accompanying mouse prompt are indicated in the middle panels. The resulting heatmaps are shown on the right-side panel. The simulated assembly consists of two chromosomes (boundaries outlined with blue annotations). The first chromosome comprises two pieces (boundaries outlined with green), while the second one comprises one. (b) Interactive genome assembly using Juicebox Assembly Tools results in chromosome-length scaffolds for the band-tailed pigeon. The left-side panel shows the draft genome assembly generated by aligning a Hi-C data set to the draft genome assembly GCA_002029285.1 (Murray et al. 2017), which was used as input into Assembly Tools. The right-side panel illustrates the contact map for resulting assembly. Corresponding loci are indicated using chromograms.

a



b

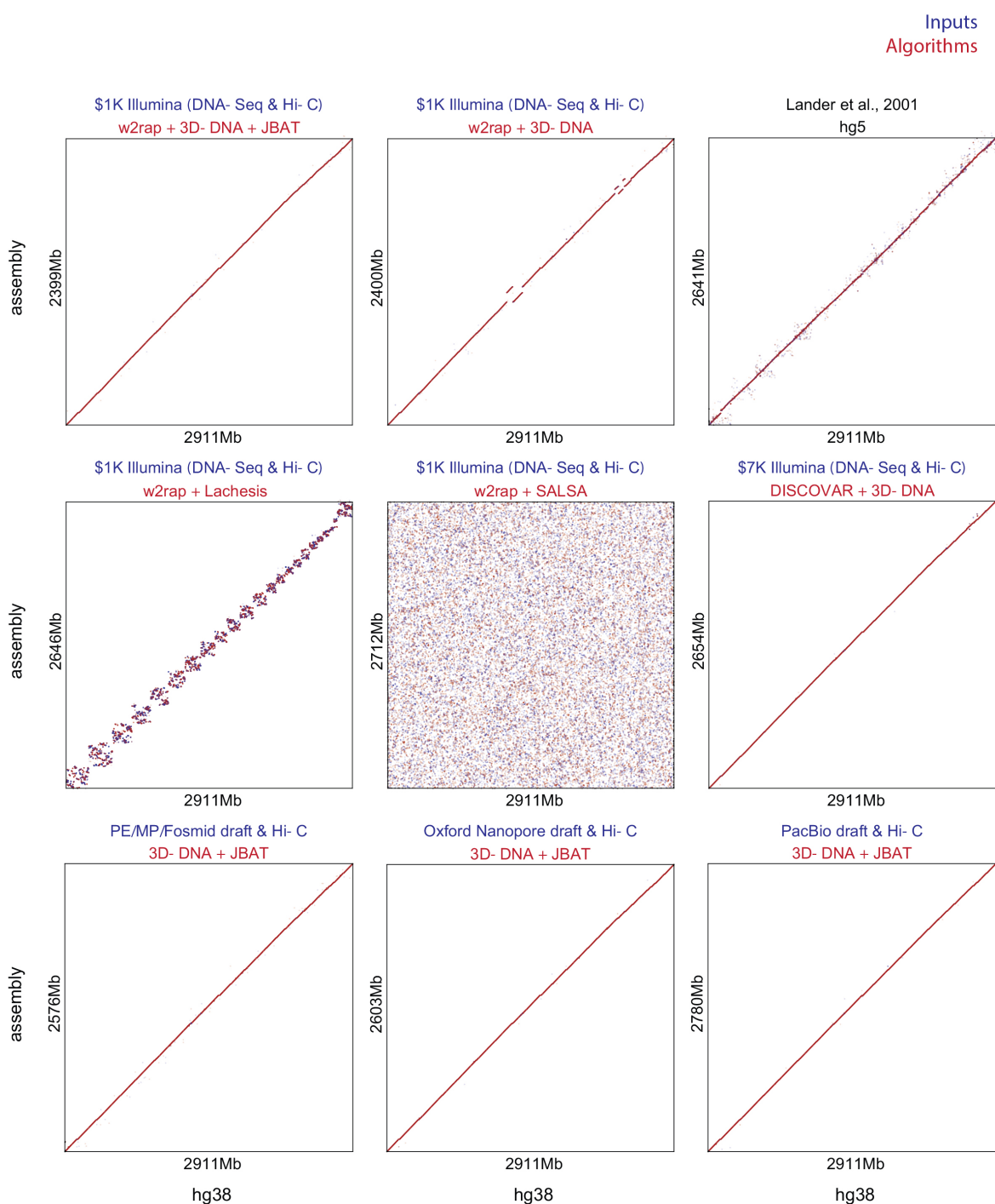
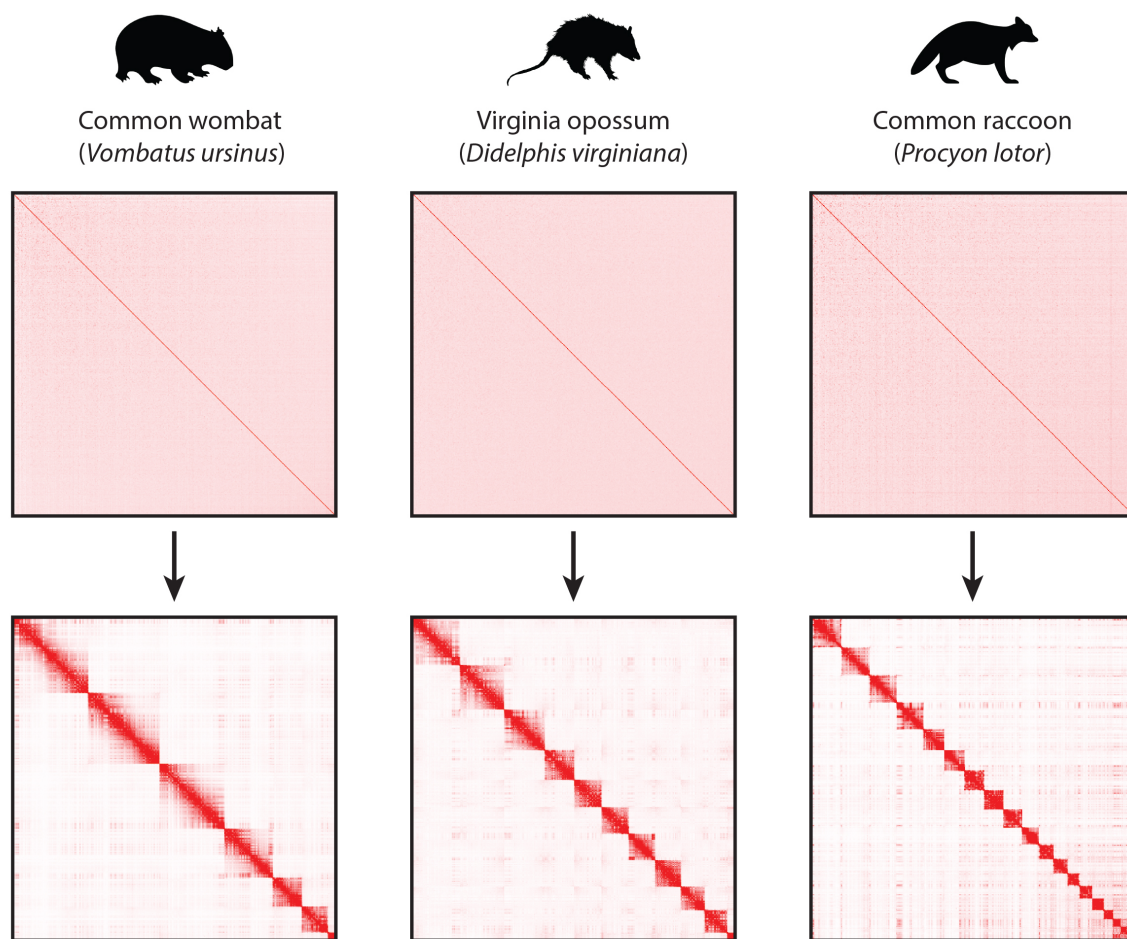


Figure 3: *De novo* assembly of mammalian genomes with chromosome-length scaffolds for \$1000. (a)

A schematic representation of a \$1000 mammalian sequencing procedure. First, we generate a PCR-free short insert DNA-Seq library, and sequence 300 million paired-end reads (2x150). These reads are assembled into a draft assembly using the software package w2rap (B. Clavijo et al. 2017; B. J. Clavijo et al. 2017). Second, we generate an *in situ* Hi-C library (Rao, Huntley et al. 2014), and sequence 100 million paired-end reads (2x150). Note that the Hi-C library can often be sequenced with shorter reads. When using PE150 all necessary data, in principle, can be obtained from a single lane of an Illumina HiSeq X instrument. The data are used to improve the draft assembly using 3D-DNA (Dudchenko et al. 2017). Finally, we validate and refine the improved assembly using Juicebox Assembly Tools (JBAT). (b) Dotplots showing alignment of several different human genome assemblies to hg38 chromosome-length scaffolds, genome-wide view. The hg38 reference (NCBI accession number GCA_000001405.23) is shown on the X-axis. Each dot represents the position of a 1kb sequence chunk aligned to hg38. The dotplots are subsampled such that every 50th chunk is displayed. The color of the dots reflects the orientation of individual alignments with respect to hg38 (red indicates a match, whereas blue indicates disagreement). Alignment was performed using BWA (Li and Durbin 2009). The assemblies shown are, left to right and top to bottom: (1) hs-1k genome assembly presented in this study; (2) 3D-DNA assembly algorithm (Dudchenko et al. 2017) applied to \$1000 data, without Juicebox Assembly Tools review; (3) hg5 genome assembly produced in 2001 by the International Human Genome Consortium (Lander et al. 2001); (4) Lachesis algorithm for Hi-C scaffolding (Burton et al. 2013) applied to \$1000 data; (5) SALSA algorithm for scaffolding long-read assemblies with Hi-C (Ghurye et al. 2017) applied to \$1000 data; (6) Hs2-HiC genome assembly, produced with PE250 short insert DNA-Seq data and Hi-C, scaffolded with 3D-DNA, see (Dudchenko et al. 2017); (7-9) 3D-DNA with Juicebox Assembly Tools procedure applied to more expensive DNA-Seq input data types: a collection of Illumina libraries with varying insert sizes including paired-end, mate-pair and fosmid libraries (Gnerre et al. 2011); Oxford Nanopore reads (Jain et al. 2017) and Pacific Biosciences long reads. Note that the automatic 3D-DNA chromosome splitter failed to split the assembly at the coverage associated with hs-1k input. As such we have split the 3D-DNA output into 23 chromosomes manually. The order and orientation of scaffolds inside chromosomes was not changed. All assemblies except for hg5 are NA12878 genome assemblies that have been scaffolded using the same Hi-C data: the first 100 million reads from the HIC001 library, whose generation and initial analysis was reported in (Rao, Huntley et al. 2014).

a



b

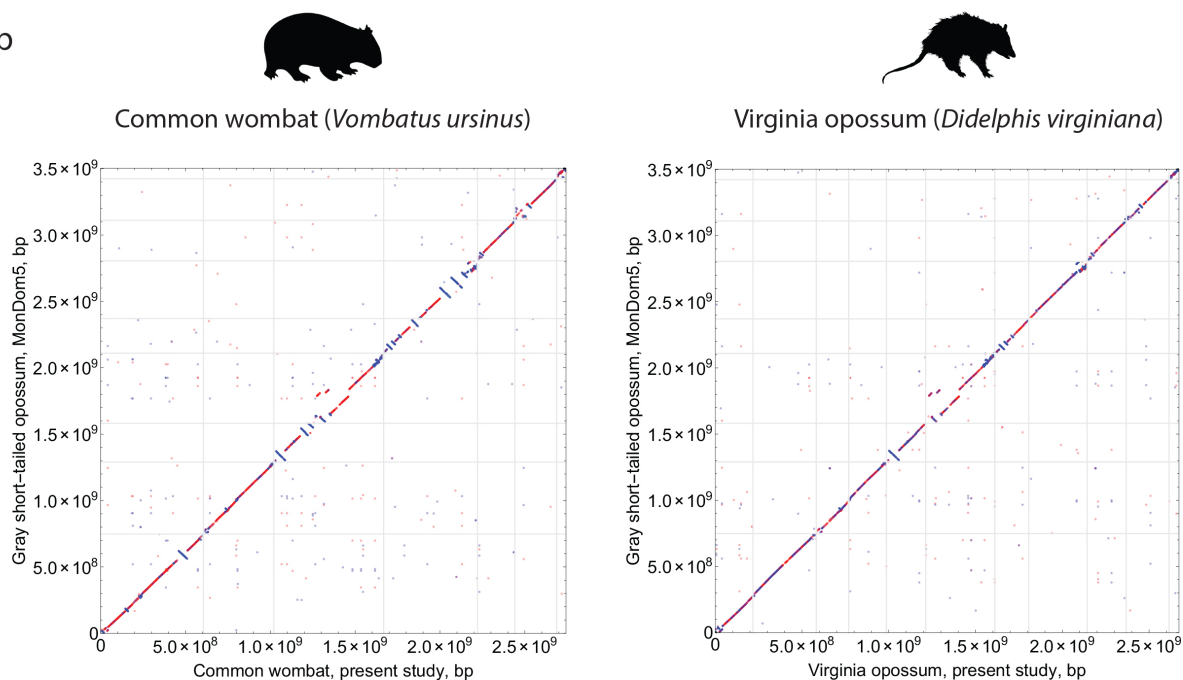


Figure 4: *De novo* genome assembly of three mammalian species – common wombat, Virginia opossum, and common raccoon – each using 400 million Illumina reads (2x150). (a) Draft assemblies are visualized on the top panel while final, chromosome-length assemblies are shown in the bottom. Only scaffolds larger than 15kb are displayed. Both the draft and the final assemblies are visualized using the Hi-C data employed for 3D-DNA genome assembly. (b) Chromosome-length *de novo* assemblies for the common wombat and Virginia opossum facilitate the analysis of karyotype evolution in marsupials. For this analysis, the gray-tailed opossum genome assembly (GCF_000002295.2) and the common wombat (vu-1k) and Virginia opossum (dv-1k) *de novo* assemblies were aligned using the LastZ alignment algorithm (Robert S. Harris 2007) using “--notransition --step=20 --nogapped” command options; the gray-tailed opossum genome assembly was used as a target. Here, we show alignment blocks with scores larger than 50,000 for the common wombat and larger than 65,000 for the Virginia opossum (Robert S. Harris 2007), with direct syntenic blocks colored red, and inverted blocks colored blue. Chromosome order and orientation has been modified in order to facilitate the comparison.

Table 1: The results of 3D-DNA and Juicebox Assembly Tools using various input data types as compared to the reference genomes produced by the International Human Genome Consortium in 2001 (hg5) and the most recent reference hg38 (GRCh38.p12). The Oxford Nanopore NA12878 assembly is based on draft shared by (Jain et al. 2017); the Pacific Biosciences NA12878 assembly is based on a high-quality contigs GCA_002077035.2 shared by the McDonnell Genome Institute. Assumed genome size for NG50 estimates is 3031.04 Mb. See also Supplementary table S3.

Input Data: 7X Hi-C &		300M PE150 Illumina	Oxford Nanopore	Pacific Biosciences	HGP (hg5)	hg38 (GRCh38.p12)
Total bases, Mb		2,400	2,604	2,780	2,641	2,911
chromosome-length scaffolds	%hg38	82.43%	89.45%	95.50%	90.72%	100.00%
Contig N50 (chromosome-length scaffolds), kb		37	3,517	14,519	80	57,879
Contig NG50 (chromosome-length scaffolds), kb		28	2,936	14,519	53	56,413
Anchoring errors (% of 1kb chunks)		0.12%	0.07%	0.09%	1.77%	n/a
Ordering errors (% of random 1 kb chunk pairs)		0.27%	0.17%	0.24%	3.98%	n/a
Orientation errors (% of 1 kb chunks)		5.18%	0.94%	0.67%	23.07%	n/a

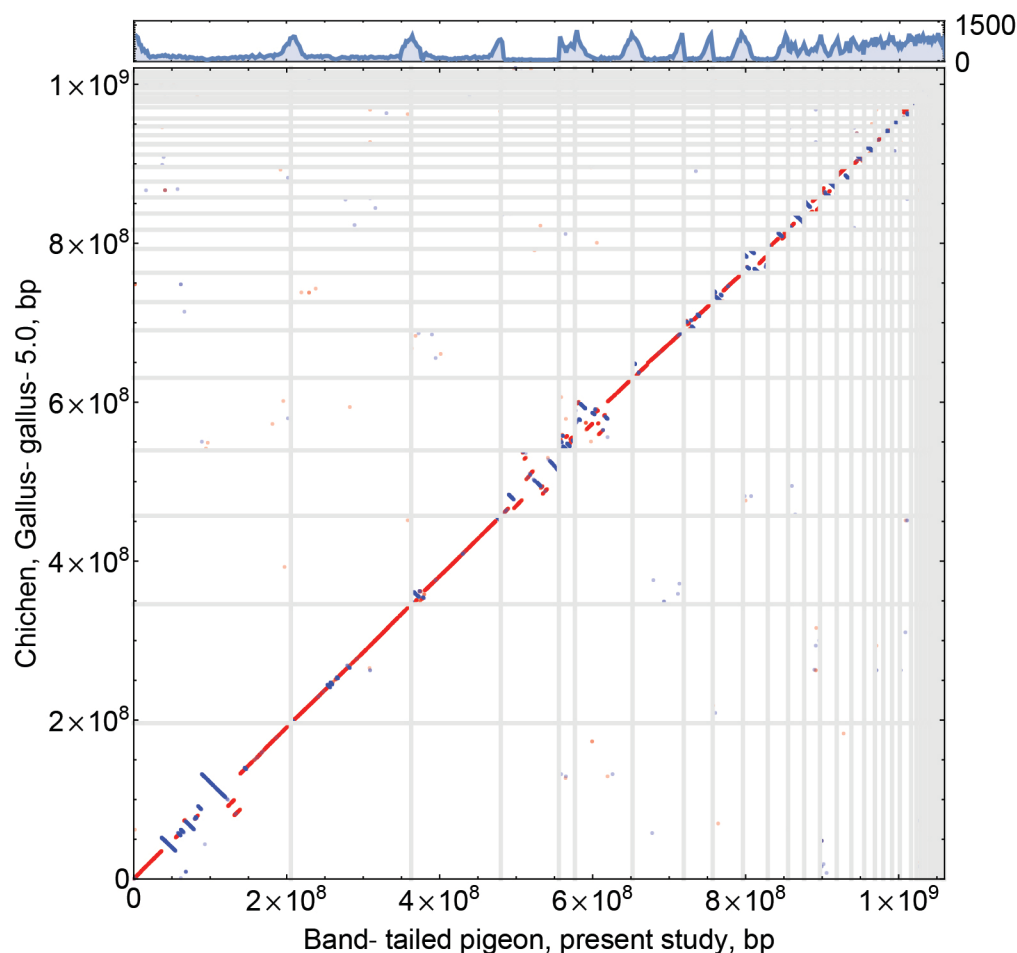


Figure S1: Strong conservation of synteny between the band-tailed pigeon (reassembled in this study using Juicebox Assembly Tools) and the chicken. We note very low levels of interchromosomal rearrangements (Derjushcheva et al. 2004). For this analysis, the chicken (GCF_000002315.4) and the band-tailed pigeon assemblies were aligned using the LastZ alignment algorithm (Robert S. Harris 2007) using “--notransition --step=20 --nogapped” command options; the chicken assembly was used as a target. Here, we show alignment blocks with scores larger than 25,000 (Robert S. Harris 2007), with direct syntenic blocks colored red, and inverted blocks colored blue. Chromosome order and orientation has been modified in order to facilitate the comparison. We also use the new assembly to revisit the question of regional variation in nucleotide diversity in passenger pigeons (Murray et al. 2017). The track on top of the synteny plot shows a total number of multiallelic sites in non-overlapping 2-Mb windows calculated from a .vcf file that describes the results of aligning passenger pigeon data to the draft band-tailed pigeon assembly.

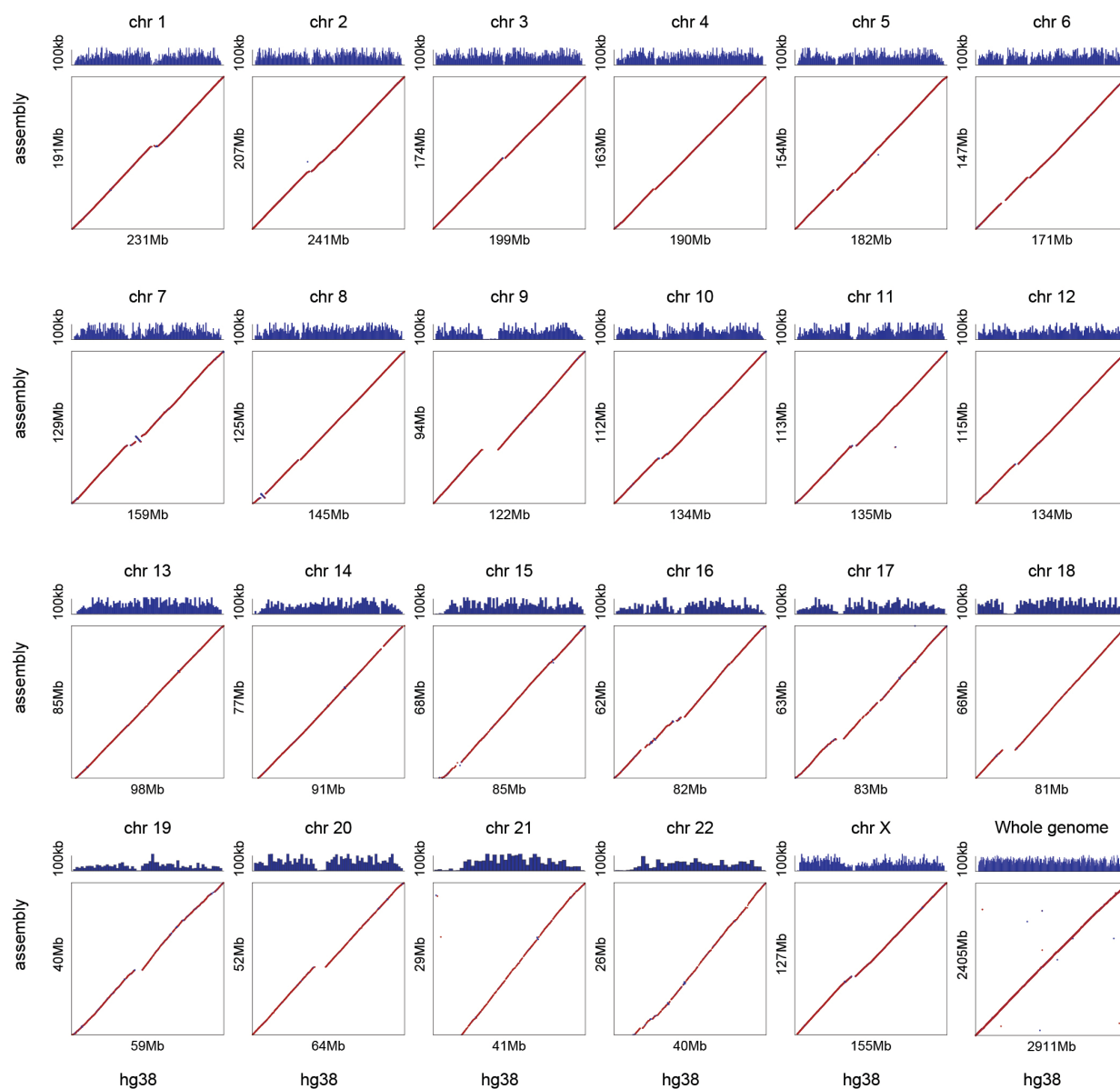


Figure S2: Dotplots showing alignment of chromosome-length scaffolds from hs-1k and hg38. The hg38 reference (NCBI accession number GCA_000001405.23) is shown on the X axis. The Y axis shows the 23 largest scaffolds of the hs-1k assembly; they have been ordered and oriented to match the chromosomes as defined in hg38 in order to facilitate comparison. (For the same reason, all gaps are removed in both assemblies.) Each dot represents the position of an individual resolved scaffold aligned to hg38. The color of the dots reflects the orientation of individual alignments with respect to hg38 (red indicates a match, whereas blue indicates disagreement). The track on top illustrates the scaffold N50 of the draft w2rap *de novo* assembly as a function of position (calculated in windows of 1Mb for individual chromosomes and 10Mb for the whole-genome graph). Alignment was performed using BWA (Li and Durbin 2009). The dotplots illustrate excellent correspondence between hg38 and hs-1k, with the exception of a few low-complexity regions of the human genome.

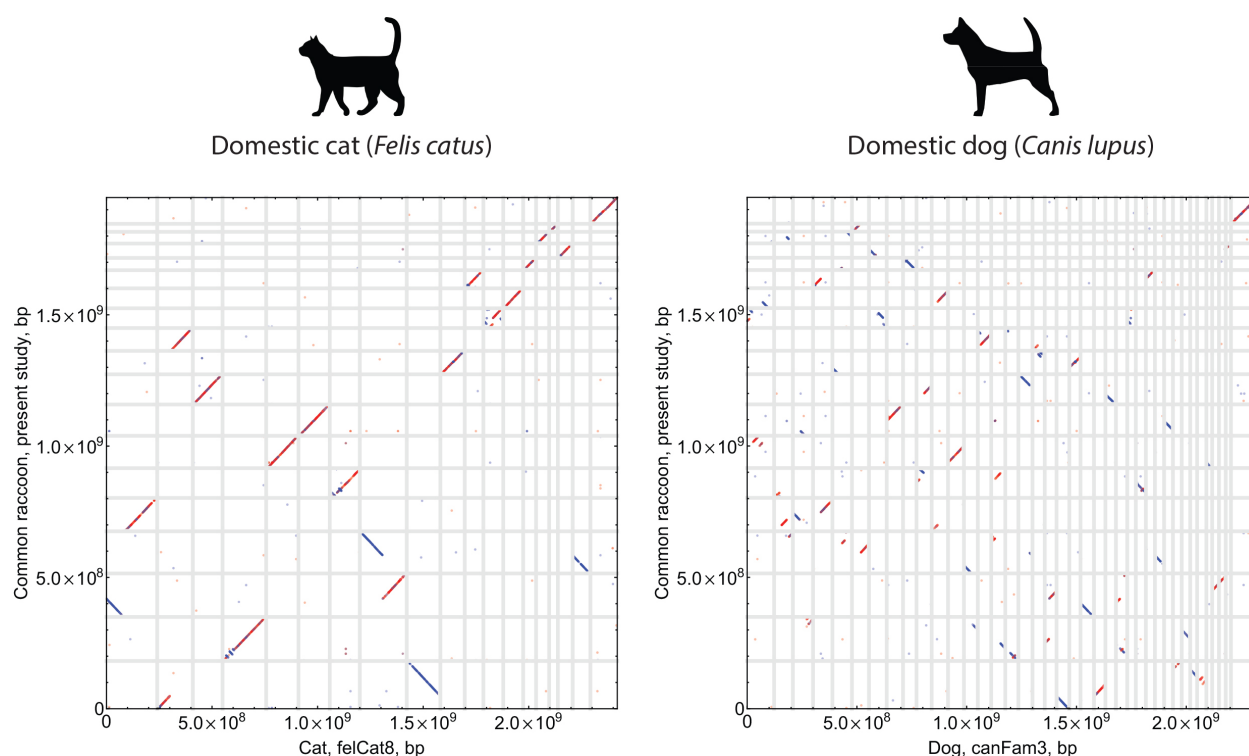


Figure S3: Conservation of synteny between cats, dogs, and the common raccoon (pl-1k). The analysis demonstrates a high degree of karyotype conservation between the raccoon and the cat, as well as a highly rearranged structure of the dog chromosomes. The results are in agreement with prior studies (Nie et al. 2012). For this analysis, the cat (GCF_000181335.2), the dog (GCF_000002285.3) and the common raccoon (pl-1k) genome assemblies were aligned using the LastZ alignment algorithm (Robert S. Harris 2007) using “--notransition --step=20 --nogapped” command options; the cat and dog assemblies were used as targets. Here, we show alignment blocks with scores larger than 50,000 (Robert S. Harris 2007), with direct synteny blocks colored red, and inverted blocks colored blue. Chromosome order and orientation of the common raccoon chromosomes has been modified in order to facilitate the comparison with (Nie et al. 2012).

Supplementary table S1: Material for the mammalian assemblies described in this study. The material has been collected in the Houston Zoo by performing three opportunistic blood draws, one draw per each mammal, secondary to veterinary and/or husbandry procedures scheduled to maintain health and welfare of the animals. Each blood draw (~1ml) was split in two to prepare DNA-Seq and Hi-C libraries. For the DNA-Seq library, we extracted DNA using QIAGEN DNeasy Blood & Tissue Kit, following the manufacturer's protocols. The DNA was sheared and prepared for Illumina sequencing using the TruSeq DNA PCR-Free kit, following the manufacturer's protocols. For Hi-C, peripheral blood mononuclear cells were separated using a Percoll gradient. The cells were crosslinked, and *in situ* Hi-C libraries were prepared in accordance with (Rao, Huntley et al. 2014). The band-tailed pigeon sample (frozen liver) was provided by the Wildlife Conservation Society. Tissue was crosslinked and dounce homogenized. Nuclei were purified on a sucrose gradient and processed to prepare *in situ* Hi-C libraries as previously described (Rao, Huntley et al. 2014). *This is the isolate that was used to generate the Hi-C data. The draft assembly has been created from a different individual (Murray et al. 2017).

	Common wombat	Virginia opossum	Common raccoon	Band-tailed pigeon
Binomial:	<i>Vombatus ursinus</i>	<i>Didelphis virginiana</i>	<i>Procyon lotor</i>	<i>Patagioenas fasciata</i>
Isolate:	Lilly	Luna	Gracie	N2016-0142*
Sex:	Female	Female	Female	Male
Source:	Houston Zoo	Houston Zoo	Houston Zoo	WCS
Tissue type:	Blood	Blood	Blood	Liver
Amount:	2ml	0.75ml	0.75ml	10mg
Notes:	-	leucistic, wild caught	-	hatchling

Supplementary table S2: Assembly statistics for *de novo* mammalian genomes produced in the current study.

	vu-1k	dv-1k	pl-1k
Binomial:	<i>Vombatus ursinus</i>	<i>Didelphis virginiana</i>	<i>Procyon lotor</i>
Total bases:	3,273,318,797	3,339,913,245	2,505,094,826
Total bases in chr-length scaffolds:	2,739,633,857	2,669,132,684	1,944,533,151
Total bases in chr-length scaffolds (% total):	83.70%	79.92%	77.62%
Contig N50, bp:	53,404	29,622	34,230
Scaffold N50, bp:	557,133,179	227,394,598	114,539,748

Supplementary table S3: Detailed statistics for human genomes assembled with 3D-DNA and Juicebox Assembly Tools as compared to several reference genomes. Here we present some more detail on the assemblies shared in Table 1 and add a comparison to two more NA12878 genome assemblies: one generated using short insert size Illumina PE250 data and assembled using DISCOVAR *de novo* (Weisenfeld et al. 2014; Love et al. 2016) and 3D-DNA (Dudchenko et al. 2017); and another one generated with 3D-DNA and Juicebox Assembly Tools (JBAT) from a collection of short read Illumina libraries with varying insert sizes from (Gnerre et al. 2011). Accuracy statistics listed include: (1) the percentage of 1kb sequences that are placed in chromosome-length scaffolds and corresponds to the “correct” chromosome (identified by whole-genome alignment) in hg38; (2) the percentage of randomly selected pairs of 1kb sequences assigned to the same chromosome-length scaffold in the assembly that are ordered in agreement with hg38; (3) the percentage of consecutive pairs of 1kb sequences that are ordered in agreement with hg38; (4) the percentage of 1kb sequences that are oriented in agreement with hg38. Only sequences uniquely aligning to hg38 (mapq \geq 60) are considered in all of the analyses.

	Input: w2rap + 3D-DNA + JBAT	5K Illumina Seq & Hi-C	DNA 5K Illumina Seq & Hi-C	PE/MP/Fosmid draft & Hi-C	Oxford Nanopore draft & Hi-C	Pacific Biosciences draft & Hi-C	hg38 (GRCh38.p12)
Algorithms:	JBAT	DISCOVAR + 3D-DNA	ALLPATHS-LG + 3D-DNA + JBAT	Canu + 3D-DNA + JBAT	Falcon + 3D-DNA + JBAT		
draft total seq bases (>1K)	2,712,354,371	2,819,306,710	2,614,901,442	2,646,010,004	2,858,827,405	n/a	n/a
draft number of contigs	164,715	80,223	231,190	2,886	3,628	n/a	n/a
draft contig N50	32,574	102,922	23,924	3,620,647	14,520,880	n/a	n/a
draft contig NG50*	28,415	94,232	19,559	3,024,148	13,176,815	n/a	n/a
draft longest contig	333,426	768,671	145,971	27,160,256	52,422,359	n/a	n/a
draft total length of scaffolds	2,717,536,071	2,819,952,110	2,786,254,569	2,646,010,004	2,858,827,405	n/a	n/a
draft number of scaffolds	112,925	73,770	11,389	2,886	3,628	n/a	n/a
draft scaffold N50	77,994	125,775	12,084,118	3,620,647	14,520,880	n/a	n/a
draft scaffold NG50*	67,475	115,268	10,404,037	3,024,148	13,176,815	n/a	n/a
draft longest scaffold	839,367	1,261,627	48,689,161	27,160,256	52,422,359	n/a	n/a
chr-length total seq bases	2,399,853,403	2,654,127,695	2,576,009,768	2,603,945,898	2,780,087,239	2,641,174,512	2,911,225,061
chr-length number of contigs	88,735	36,616	213,278	2,441	1,200	148,041	665
chr-length contig N50	36,914	108,937	24,318	3,517,161	14,518,630	80,184	57,879,411
chr-length contig NG50*	28,200	93,928	19,509	2,935,826	14,518,630	53,330	56,413,054
chr-length longest contig	333,426	768,671	260,490	27,039,813	47,391,326	28,515,322	141,414,041
chr-length total length of scaffolds	2,423,876,603	2,669,861,395	2,735,224,778	2,605,154,898	2,780,675,739	3,258,491,231	3,031,042,417
chr-length number of scaffolds	23	23	23	23	23	23	23
chr-length scaffolds N50**	125,170,150	141,244,516	146,426,750	138,911,586	149,363,931	162,599,930	156,040,895
chr-length scaffold NG50*	114,962,098	136,507,704	140,094,183	133,352,303	141,190,470	166,623,906	156,040,895
chr-length longest scaffold	207,780,125	225,222,252	229,098,500	219,596,205	232,958,391	282,193,664	248,956,422
full output total seq bases	2,712,354,371	2,819,306,710	2,614,901,442	2,646,010,004	2,858,827,405	2,692,861,938	3,095,979,984
full output number of contigs	165,597	80,725	231,751	3,749	4,668	150,750	1,502
full output contig N50	32,499	102,793	23,907	3,490,673	13,912,961	80,911	56,413,054
full output contig NG50*	28,350	93,976	19,531	2,935,826	13,912,961	57,055	56,413,054
full output longest contig	333,426	768,671	260,490	27,039,813	47,391,326	28,515,322	141,414,041
full output total length of scaffolds	2,736,505,371	2,835,055,510	2,784,512,041	2,647,366,504	2,859,645,405	3,400,374,049	3,257,347,282
full output number of scaffolds	75,863	44,065	11,071	1,036	3,032	42	595
full output scaffolds N50**	125,170,150	141,244,516	146,426,750	138,911,586	149,363,931	152,776,421	145,138,636
full output scaffold NG50*	114,962,098	136,507,704	140,094,183	133,352,303	141,190,470	166,623,906	156,040,895
full output longest scaffold	207,780,125	225,222,252	229,098,500	219,596,205	232,958,391	282,193,664	248,956,422
chr-length, % of draft assembly seq bases	88.48%	94.14%	98.51%	98.41%	97.25%	98.08%	94.03%
chr-length, % of hg38 chr spanned	99.34%	99.10%	99.33%	99.49%	99.42%	99.47%	100.00%
chr-length, % of hg38 chr-length seq bases	82.43%	91.17%	88.49%	89.45%	95.50%	90.72%	100.00%
% of IHGSC 2001 chr-length seq bases (hs5)	90.86%	100.49%	97.53%	98.59%	105.25%	100.00%	110.22%
1kb chunks, % of assigned to correct chr	99.88%	99.83%	99.84%	99.93%	99.91%	98.23%	n/a
1kb chunks, % of correctly ordered for randomly selected pairs	99.73%	99.13%	99.60%	99.83%	99.76%	96.02%	n/a
1kb chunks, % of correctly ordered for adjacent pairs	94.95%	95.73%	99.40%	99.10%	99.40%	77.55%	n/a
1kb chunks, % of correctly oriented	94.82%	95.58%	99.07%	99.06%	99.33%	76.93%	n/a

*Assumed genome size: 3031.04 Mb

**The scaffold N50 and NG50 for the output assemblies are determined almost entirely by the lengths of chromosomes