Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions

Joshua N. Burton[1], Andrew Adey[1], Rupali P. Patwardhan[1], Ruolan Qiu[1], Jacob O. Kitzman[1], Jay Shendure[1]

[1] Department of Genome Sciences, University of Washington, Seattle, WA 98115, USA

Correspondence to:
Jay Shendure (shendure@uw.edu)
Joshua Burton (jnburton@uw.edu)

Supplementary Material

Supplementary Information

TABLE OF CONTENTS

Metric		*De novo* assemblies			
		Human	Human hi-contiguity, from ref. 6	Mouse	*Drosophila*
Assembly metrics	Total assembly length, gapped (Mb)	2,739	2,773	2,370	127.2
	Length / reference length	93.0%	94.1%	87.0%	75.4%
	N. contigs or scaffolds	18,921	3,811	25,964	7,109
	N50 contig/scaffold, ungapped (Kb)	437	11,547	224	68
% sequence (% contigs) clustered into groups		98.2% (71.5%)	99.0% (53.3%)	98.0% (87.8%)	81.2% (64.3%)
% clustered sequence (% contigs) mis-clustered		0.14% (1.4%)	4.7% (7.9%)	0.24% (0.5%)	3.4% (10.5%)
Full orders	% clustered sequence (% contigs) ordered	94.4% (55.3%)	99.4% (28.6%)	86.7% (42.7%)	82.0% (24.5%)
	% ordered sequence (% contigs) w/ ordering errors	0.5% (0.8%)	8.4% (9.5%)	0.5% (1.1%)	4.6% (5.2%)
	% ordered sequence (% contigs) w/ orientation errors	1.2% (2.5%)	6.4% (10.3%)	1.9% (4.6%)	4.1% (6.1%)
	% ordered sequence (% contigs) w/ high quality	92.8% (79.0%)	88.4% (51.6%)	93.3% (82.9%)	94.1% (88.1%)
	% high-quality sequence (% contigs) w/ ordering errors	0.3% (0.4%)	4.7% (3.7%)	0.3% (0.7%)	3.3% (3.4%)
	% high-quality sequence (% contigs) w/ orientation errors	0.4% (0.5%)	3.4% (3.3%)	0.5% (1.0%)	2.5% (2.7%)
Trunks	% ordered sequence (% contigs) in trunks	88.4% (88.5%)	82.4% (76.2%)	90.4% (88.4%)	70.7% (70.6%)
	% sequence in trunks (% contigs) w/ ordering errors	0.2% (0.4%)	5.3% (7.5%)	0.2% (0.4%)	3.0% (4.0%)
	% sequence in trunks (% contigs) w/ orientation errors	1.1% (2.3%)	2.8% (7.5%)	1.7% (4.2%)	1.9% (3.5%)
	% sequence in trunks (% contigs) w/ high quality	93.0% (79.4%)	92.4% (56.8%)	93.6% (83.5%)	94.7% (89.6%)
	% high-quality sequence in trunks (% contigs) w/ ordering errors	0.1% (0.2%)	3.0% (2.0%)	0.1% (0.2%)	2.1% (2.5%)
	% high-quality sequence in trunks (% contigs) w/ orientation errors	0.3% (0.3%)	1.0% (0.8%)	0.4% (0.8%)	1.1% (1.6%)

Supplementary Table 1 | Metrics for the *LACHESIS* scaffolding results. This is a more detailed version of **Table 1**. Results for the human *de novo* assembly exclude the chimeric group not shown in **Figure 3**.

Figure	Dominant chrom(s)	Sequence length in grouped scaffolds				Sequence length in ordered scaffolds			
		Total (Mb)	Percent aligning to…			Total (Mb)	Percent aligning to…		
			Dominant chrom(s)	Other chroms	None		Dominant chrom(s)	Other chroms	None
3a	chr1	210.9	99.9%	0.01%	0.07%	202.6	100%	-	-
3b	chr2	224.4	99.9%	0.02%	0.05%	216.8	100%	-	-
3c	chr3	190.6	99.3%	0.6%	0.02%	182.8	99.3%	0.7%	-
3d	chr4	181.0	99.98%	0.01%	0.01%	173.6	100%	-	-
3e	chr5	170.5	99.9%	0.01%	0.09%	162.1	100%	-	-
3f	chr6	164.9	99.2%	0.8%	0.02%	156.8	99.2%	0.8%	-
3g	chr7	143.9	99.8%	0.03%	0.18%	134.8	100%	-	-
3h	chr8	136.7	99.8%	0.15%	0.01%	131.3	99.9%	0.1%	-
3i	chr9	106.7	99.9%	0.03%	0.09%	101.0	100%	-	-
3j	chr10	125.9	99.6%	0.3%	0.09%	119.8	99.8%	0.2%	-
3k	chr11	125.7	99.9%	0.01%	0.10%	118.3	99.99%	0.01%	-
3l	chr12	126.0	99.9%	0.1%	0.04%	119.8	99.9%	0.1%	-
3m	chr13	93.9	99.96%	0.007%	0.03%	92.2	100%	-	-
3n	chr14	84.8	99.7%	0.2%	0.05%	81.4	99.8%	0.2%	-
3o	chr15	75.5	99.8%	0.01%	0.2%	71.0	100%	-	-
3p	chr16	68.3	99.6%	0.06%	0.3%	64.3	100%	-	-
3q	chr17	73.4	99.7%	0.1%	0.2%	65.9	100%	-	-
3r	chr18	72.4	99.95%	0.02%	0.04%	70.8	100%	-	-
3s	chr19, chr22	82.8	99.9%	0.1%	0.03%	67.9	57.6%, 42.4%	-	-
3t	chr20, chr21	91.2	99.8%	0.2%	0.01%	88.0	63.2%, 36.6%	0.2%	-
3u	chrX	36.7	99.9%	0.03%	0.05%	34.8	100%	-	-
3v	chrX	104.5	99.5%	0.01%	0.4%	90.9	100%	-	-
Supp. Figure 4w	chr16	6.5	23.5%	47.6%	29.0%	2.3	42.8%	54.2%	3.0%

Supplementary Table 2 | Contents of *LACHESIS*' orderings in the human *de novo* assembly (Figure 3). For each of the 23 groups, there is a "dominant chromosome" in the reference genome to which the plurality of alignable sequence aligns. This chart shows what fraction of the scaffold length in each ordering aligns to the dominant chromosome, to other chromosomes, or to no chromosomes. The last row corresponds to the small, chimeric chromosome group described in the main text.

Repeat type	UCSC Genome Browser track name	Enrichment near the edges of mis-ordered scaffolds
Segmental duplications (>1 Kb length, >90% similarity)	Segmental Dups	**6.38**
Microsatellite repeats (dinucleotide, trinucleotide)	Microsatellite	**1.24**
Simple tandem repeats (4 or more nucleotides)	Simple Repeats	**2.87**
RepeatMasked regions	RepeatMasker	**0.93**
Interrupted repeats called by RepeatMasker	Interrupted Rpts	**0.94**

Supplementary Table 3 | Enrichment of repetitive sequences in error-prone regions. Human genomic regions corresponding to several different types of repetitive sequence elements were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu). For each scaffold in the human *de novo* assembly created by *LACHESIS*, a 5 Kb region was extracted around each of its ends. These edge regions were then overlapped with the repeat elements. The enrichment shown for each type of repeat element is the ratio of the frequency with which that element co-occurs with the ends of one of the 61 scaffolds marked with ordering errors, divided by the frequency with which it co-occurs with the ends of one of the 7,604 scaffolds not marked with ordering errors.

Supp. Figure	Dominant chrom(s)	Sequence length in grouped scaffolds				Sequence length in ordered scaffolds			
		Total (Mb)	Percent aligning to…			Total (Mb)	Percent aligning to…		
			Dominant chrom(s)	Other chroms	None		Dominant chrom(s)	Other chroms	None
7a	chr1	176.5	99.7%	0.3%	0.02%	150.8	99.7%	0.3%	-
7b	chr2	167.3	99.3%	0.6%	0.1%	149.4	99.4%	0.6%	-
7c	chr3	142.0	99.8%	0.2%	0.04%	119.8	99.9%	0.1%	-
7d	chr4	136.2	99.9%	0.008%	0.1%	118.0	100%	-	-
7e	chr5	136.2	99.93%	0.02%	0.05%	119.1	100%	-	-
7f	chr6	134.6	99.7%	0.2%	0.1%	114.1	99.8%	0.2%	-
7g	chr7	120.2	99.8%	0.01%	0.2%	102.4	100%	-	-
7h	chr8	119.6	97.8%	2.2%	0.04%	106.5	97.6%	2.4%	-
7i	chr9	113.9	99.93%	0.01%	0.06%	103.1	100%	-	-
7j	chr10	116.5	99.8%	0.1%	0.1%	101.0	99.9%	0.1%	-
7k	chr11	113.8	99.5%	0.4%	0.06%	106.8	99.5%	0.5%	-
7l	chr12	104.7	99.9%	0.02%	0.1%	89.9	100%	-	-
7m	chr13	106.1	99.8%	0.02%	0.2%	91.9	100%	-	-
7n	chr14	106.2	99.8%	0.002%	0.2%	92.1	100%	-	-
7o	chr15	95.2	99.96%	0.03%	0.02%	83.7	100%	-	-
7p	chr16	89.6	99.99%	0.003%	0.004%	79.0	100%	-	-
7q	chr17	84.3	99.7%	0.06%	0.2%	73.1	100%	-	-
7r	chr18	82.3	99.93%	0.06%	0.003%	73.1	100%	-	-
7s	chr19	55.6	99.94%	0.01%	0.04%	50.3	100%	-	-
7t	chrX	122.4	99.7%	0.1%	0.2%	90.9	99.9%	0.1%	-

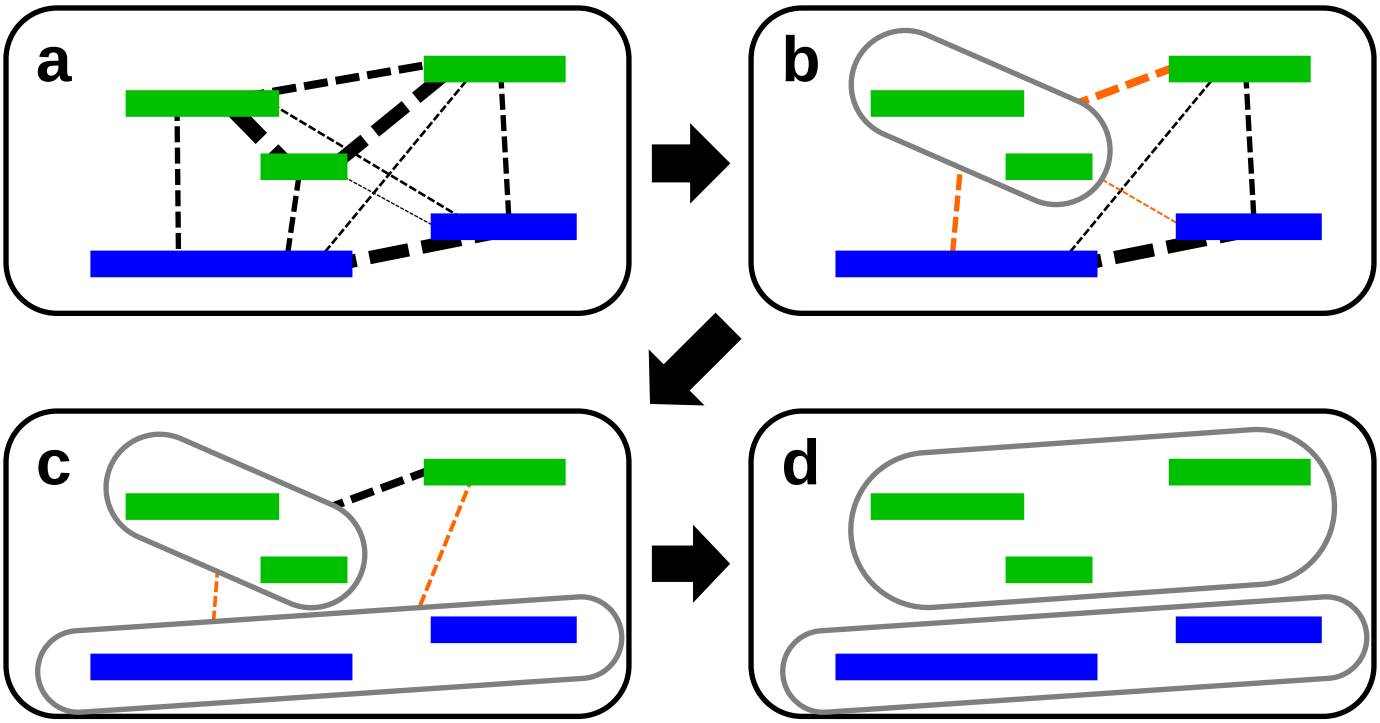
Supplementary Table 4 | Contents of *LACHESIS*' orderings in the mouse *de novo* assembly (Supplementary Figure 7). For each of the 20 groups, there is a "dominant chromosome" in the reference genome to which the plurality of alignable sequence aligns. This chart shows what fraction of the scaffold length in each ordering aligns to the dominant chromosome, to other chromosomes, or to no chromosomes.

Supp. Figure	Dominant chrom	Sequence length in grouped contigs				Sequence length in ordered contigs			
		Total (Mb)	Percent aligning to…			Total (Mb)	Percent aligning to…		
			Dominant chrom	Other chroms	No euchromatic sequence		Dominant chrom	Other chroms	No euchromatic sequence
9b	X	18.4	75.8%	2.3%	21.9%	12.9	85.0%	2.7%	12.3%
9c	4	2.5	46.5%	21.9%	31.7%	0.74	93.2%	4.0%	2.8%
9d	2	41.1	58.4%	1.9%	39.7%	32.6	71.1%	2.2%	26.7%
9e	3	40.4	82.6%	2.0%	15.4%	37.5	85.2%	1.9%	12.8%

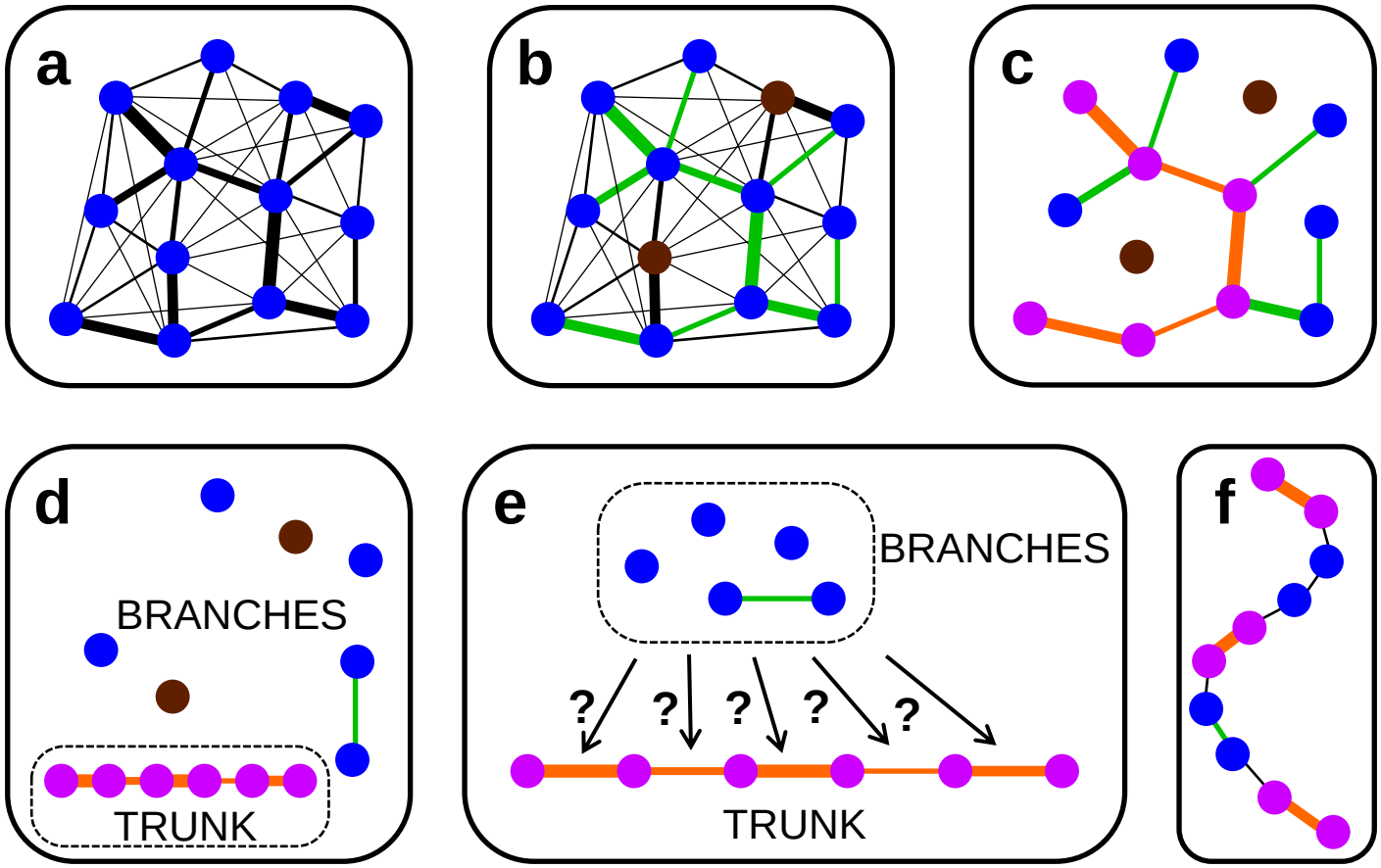
Supplementary Table 5 | Contents of *LACHESIS*' orderings in the *D. melanogaster de novo* assembly (Supplementary Figure 9). For each of the four groups, there is a "dominant chromosome" in the reference genome to which the majority of euchromatic sequence aligns. This chart shows what fraction of the contig length in each ordering aligns to the dominant chromosome, to other chromosomes, or to no euchromatic chromosome. Note that a substantial fraction of the *D. melanogaster* assembly may consist of heterochromatic sequences as it does not align to the four euchromatic chromosomes of the reference assembly.

Number of Hi-C pairs, before filtering	Percent of total Hi-C coverage	% (by length) of sequence clustered	Clustering error rate, excluding fusions	% (by length) of sequence ordered	Ordering error rate	Orienting error rate
51,493,359	7.0%	95.97%	0.36%	92.53%	14.8%	12.5%
113,961,921	15.5%	96.97%	0.25%	92.72%	6.3%	6.4%
175,873,230	24.0%	97.08%	0.16%	92.81%	4.6%	5.0%
237,662,270	32.4%	97.13%	0.16%	92.79%	4.0%	4.6%
404,341,129	55.1%	97.92%	0.14%	92.95%	1.0%	1.7%
568,435,079	77.4%	98.04%	0.13%	92.96%	0.8%	1.4%
734,185,216	100%	98.22%	0.15%	93.02%	0.5%	1.2%

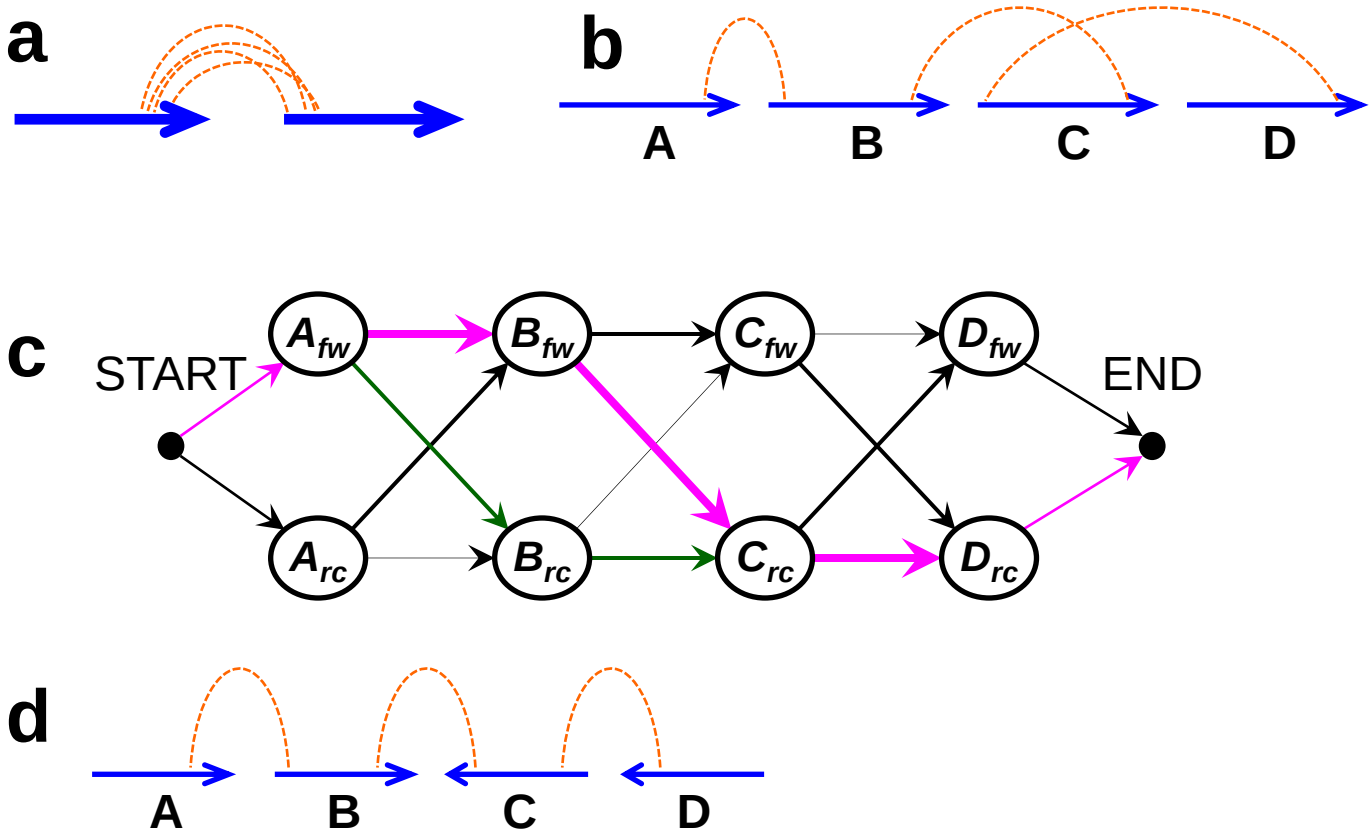
Supplementary Table 6 | The effect of Hi-C down-sampling on *LACHESIS* assembly quality. *LACHESIS* was provided with varying quantities of Hi-C read coverage with which to scaffold the shotgun human assembly. As read coverage increased, the total amount of sequence placed by *LACHESIS* increased slightly, while error rates decreased significantly. The bottom row describes the same assembly as in **Figure 2a, 2b** and **Supplementary Table 2**.



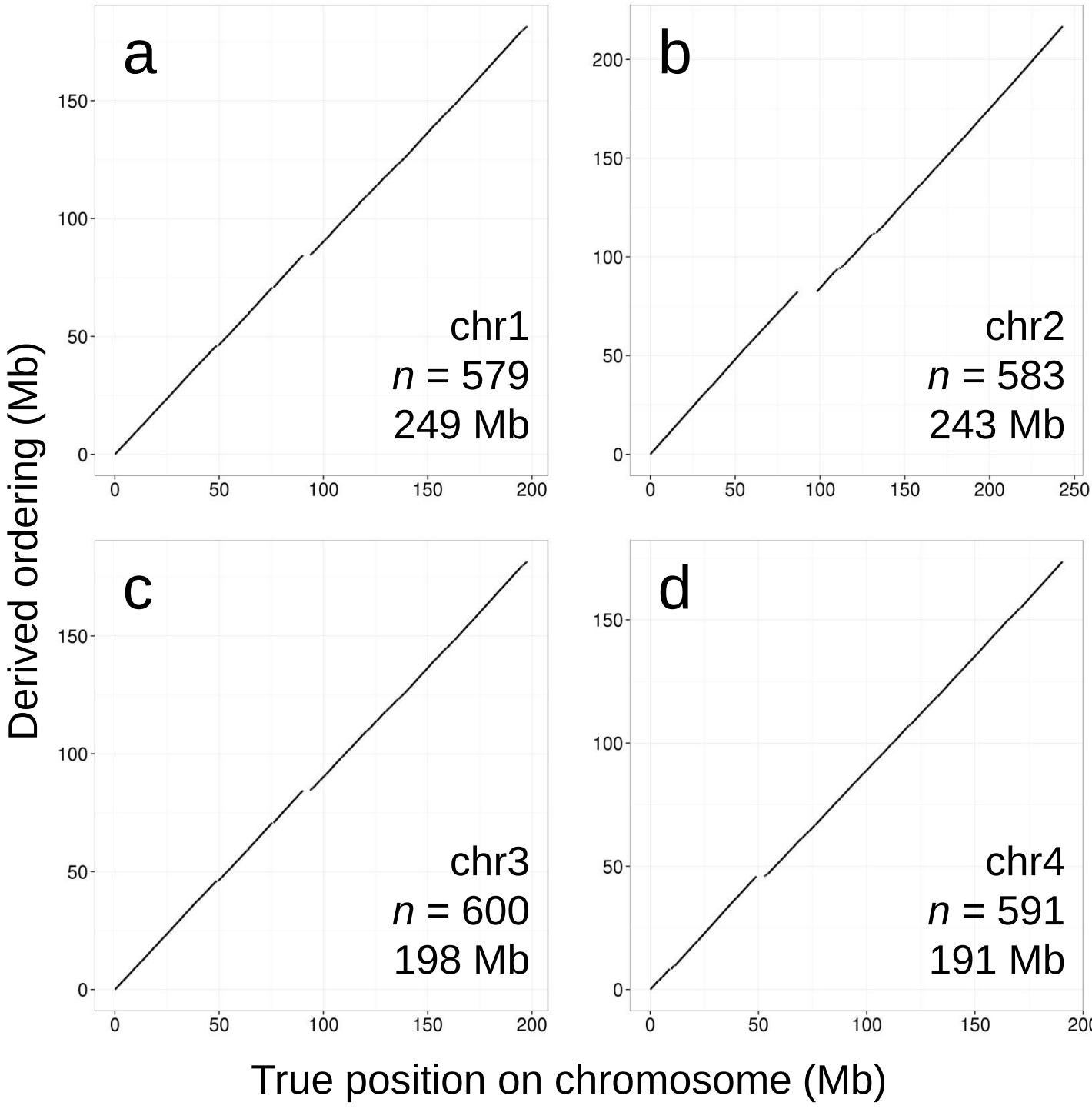
Supplementary Figure 1 | An illustrated overview of the *LACHESIS* clustering algorithm. a. An assembly consisting of five contigs, which in truth belong to two chromosomes (green and blue). Hi-C links between the contigs are shown as black dotted lines, with thicker lines indicating higher normalized link density. **b.** The agglomerative hierarchical clustering algorithm begins. The two contigs sharing the highest normalized link density are merged together to create a cluster (gray oval). The new link densities between this cluster and each other contig (orange dotted lines) are calculated as the average (normalized) linkage between the two contigs in this cluster and the other contig. **c.** Again, the two contigs sharing the highest normalized link density are merged to create a cluster. New average link densities are calculated (orange dotted lines); note that the link density between the two multi-contig clusters is the average of four original link densities. **d.** Another merge. The user-specified limit of two clusters has been reached, so the algorithm is complete. It has correctly found groups for each chromosome.



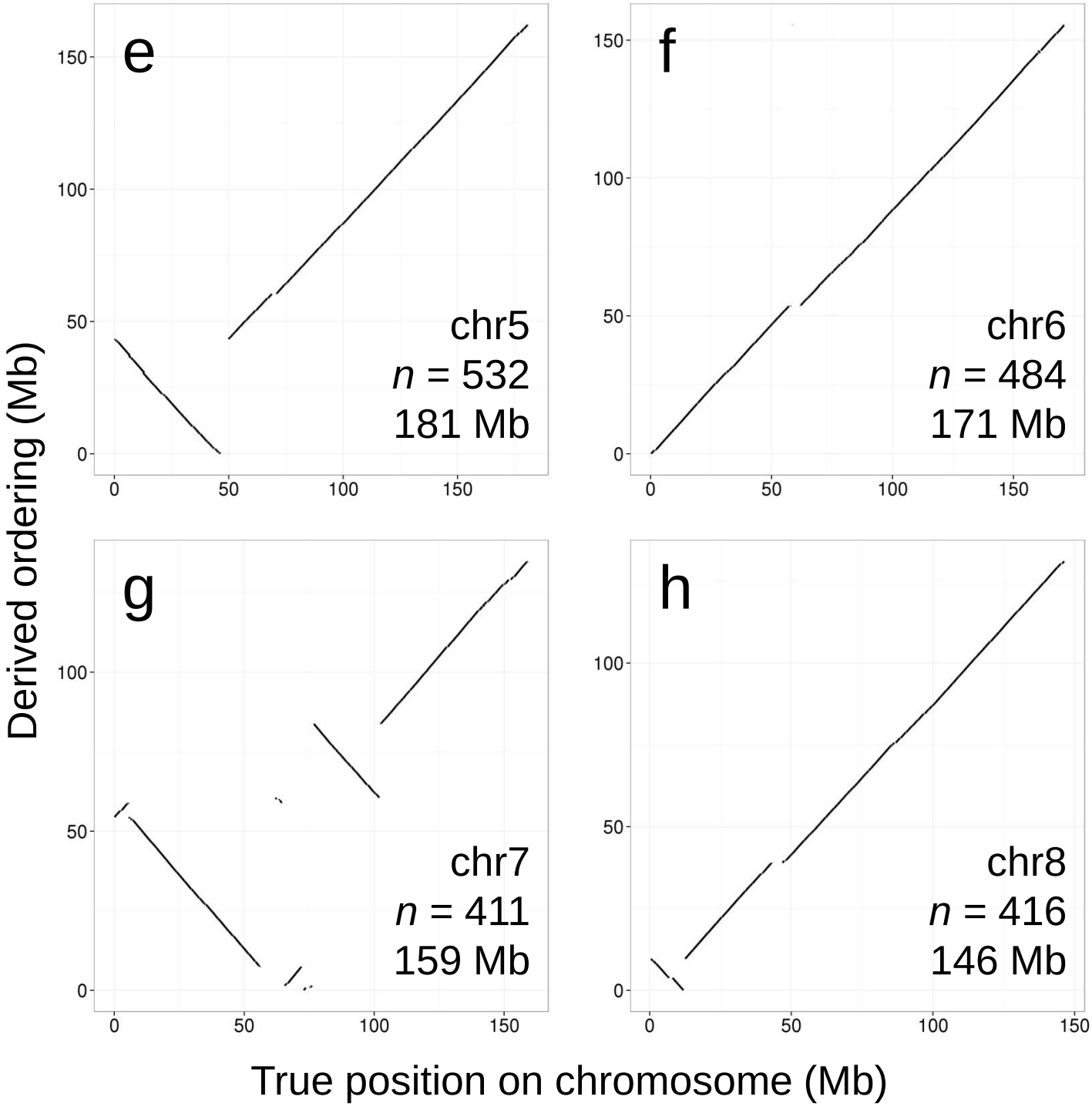
Supplementary Figure 2 | An illustrated overview of the *LACHESIS* ordering algorithm. a. A group of contigs depicted as a graph. Each blue vertex indicates a contig, and the edges between vertices indicate normalized Hi-C link densities (for clarity, edges are not shown between all pairs of contigs). **b.** A spanning tree (a set of edges that connects all vertices with no loops) is found (green edges). The edges of the spanning tree are chosen to have the maximum possible link densities. Short contigs (dark brown dots) are not included in the spanning tree. **c.** The longest path in the spanning tree (magenta dots, orange edges) is found. This path constitutes the "trunk", an initial contig ordering with high accuracy but low completeness. **d.** The trunk is removed from the spanning tree, leaving a set of vertices and edges called "branches", many of which consist of a single isolated vertex. **e.** Lastly, the branches are considered for reinsertion into the trunk at all possible positions and orientations. Each possible reinsertion site is given a "score" equal to the sum of the reciprocals of all link distances. Very short branches are not reinserted. **f.** The final contig ordering.



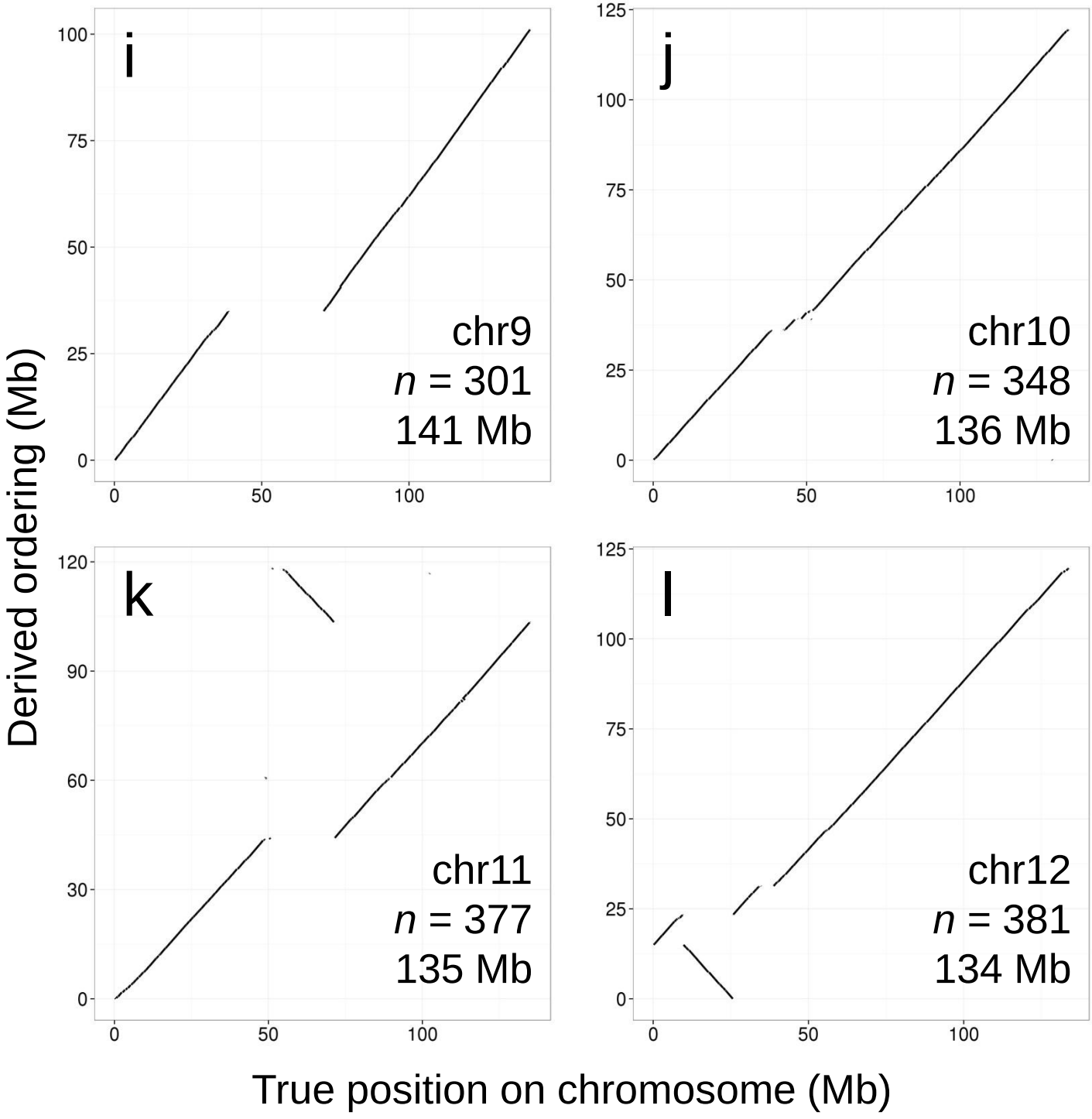
Supplementary Figure 3 | An illustrated overview of the *LACHESIS* orienting algorithm. a. A pair of contigs connected by several Hi-C links, with the exact location of the aligned Hi-C reads shown (orange dotted lines). All of the reads in these links are localized to one end of each contig, which suggests that the contigs should be placed in the orientation shown; any other orientation would increase the perceived length of the links. Note that this is the only time *LACHESIS* uses the exact location of the reads in a Hi-C link, as opposed to the mere fact of a link between two contigs. **b.** An ordering of four contigs A,B,C,D, with arbitrary initial orientations. The exact locations of the Hi-C read alignments between adjacent contigs are shown (for clarity, only one link per adjacency is shown). **c.** A weighted directed acyclic graph (WDAG) describing all possible ways in which these four contigs could be ordered. The edges exiting the start node and entering the end note all have the same weight. The edge weights between each pair of contigs (arrows) are set to the log-likelihoods of observing the Hi-C links between those two contigs in the two orientations, given that longer links are less likely; larger numbers (thicker arrows) indicate more likely orientations. The likeliest path through the WDAG (magenta arrows) is shown. The orientation quality score is calculated as the differential to the log-likelihood caused by choosing a particular orientation; for example, for contig B, the log-likelihood is the difference between the weights of the magenta arrows entering and leaving node $\boldsymbol{B_{fw}}$ and the weights of the alternative nodes entering and leaving $\boldsymbol{B_{rc}}$ (dark green arrows). **d.** The contig orientations corresponding to the likeliest path found in **c**.



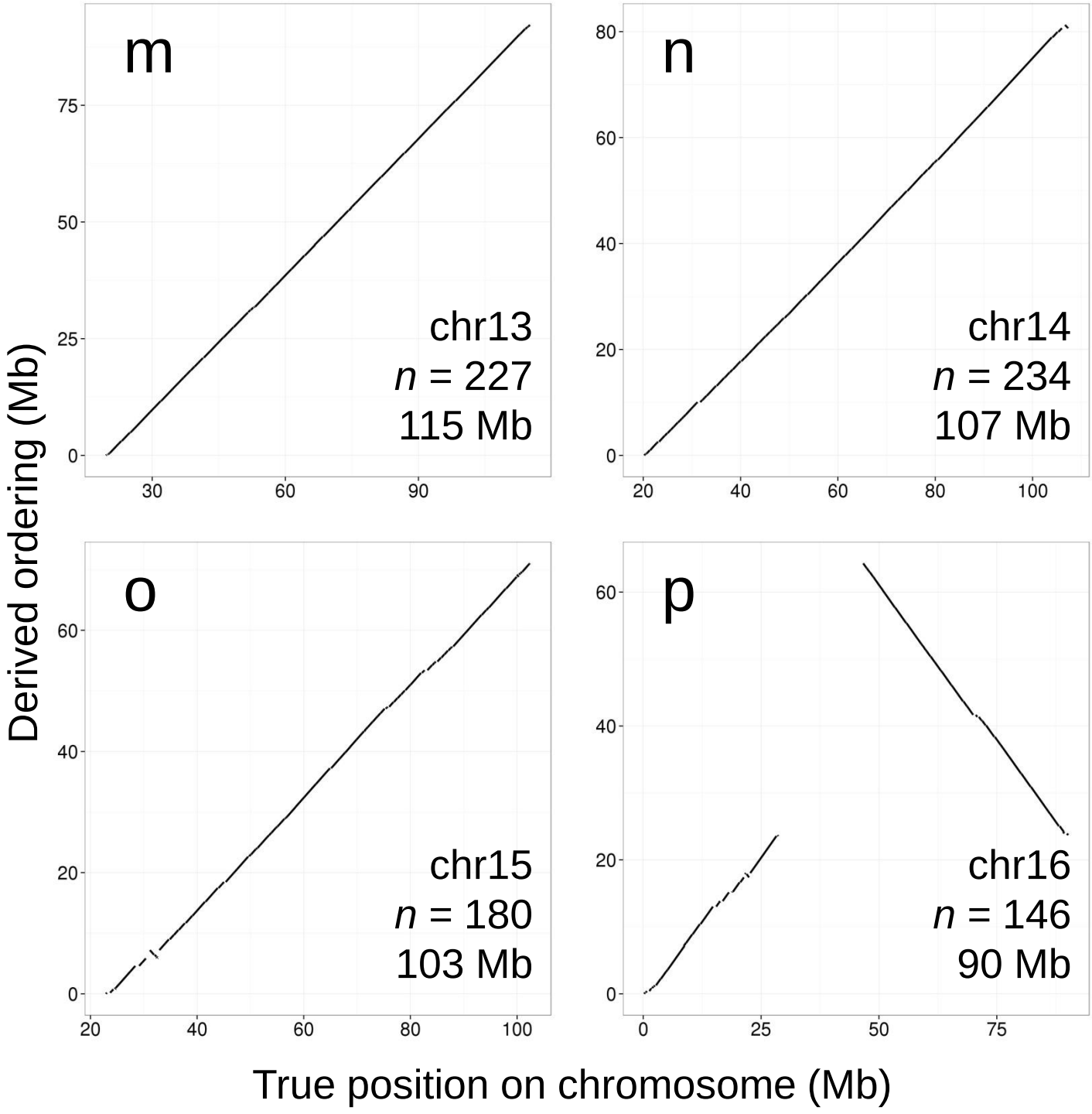
Supplementary Figure 4 (page 1 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3**. **w**, the chimeric group not shown in **Figure 3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.



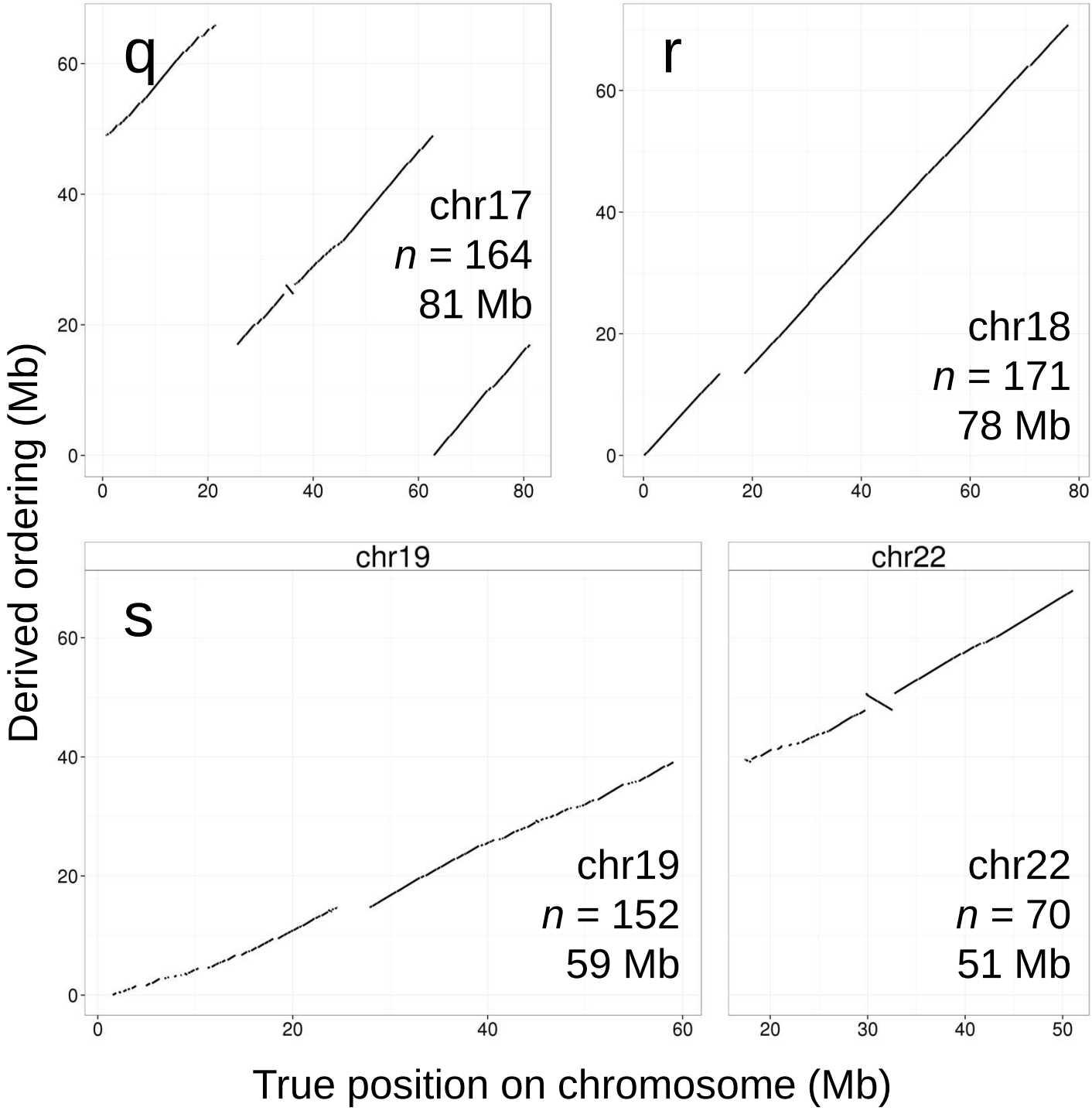
Supplementary Figure 4 (page 2 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3**. **w**, the chimeric group not shown in **Figure 3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.



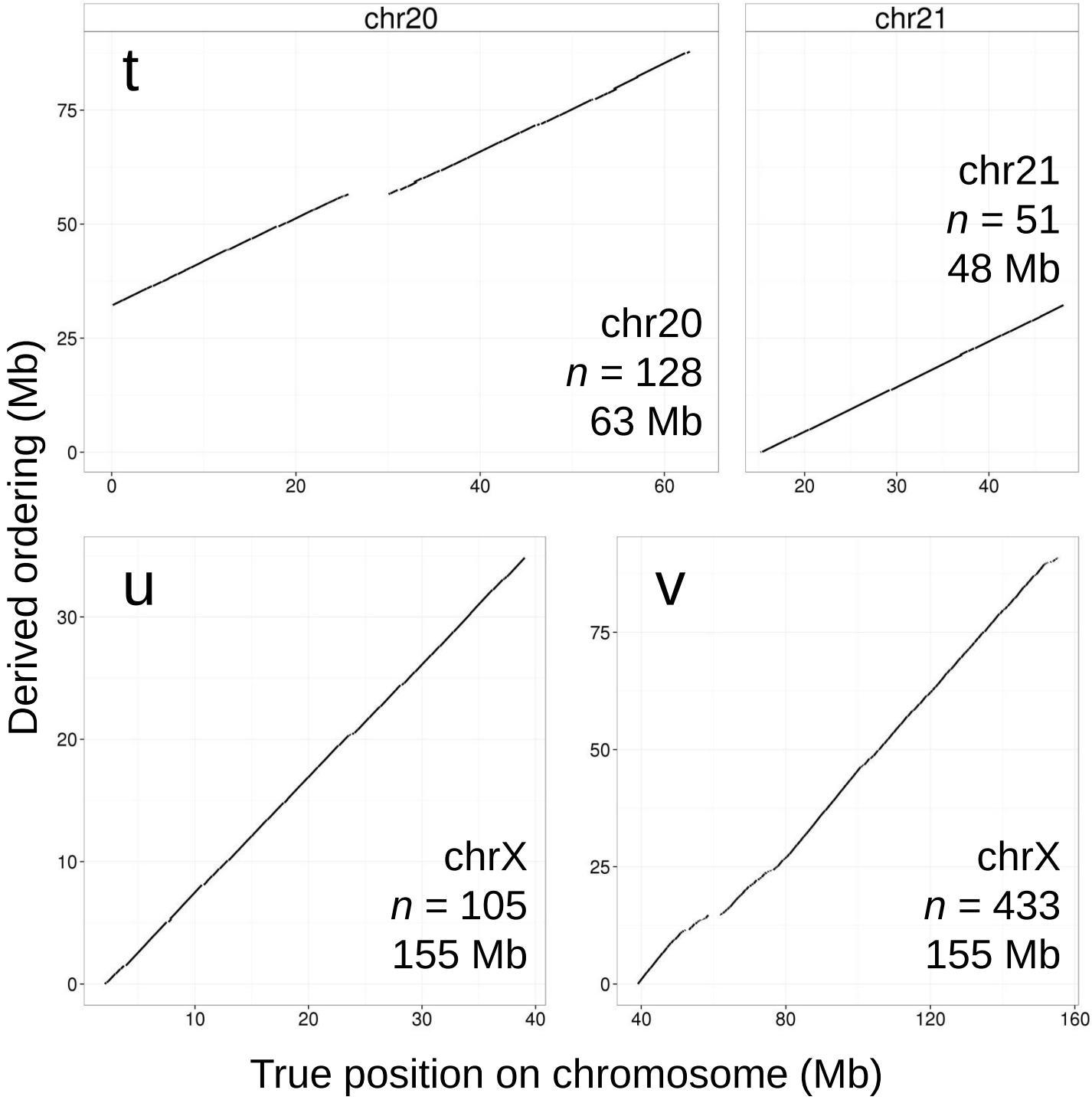
Supplementary Figure 4 (page 3 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3**. **w**, the chimeric group not shown in **Figure 3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.



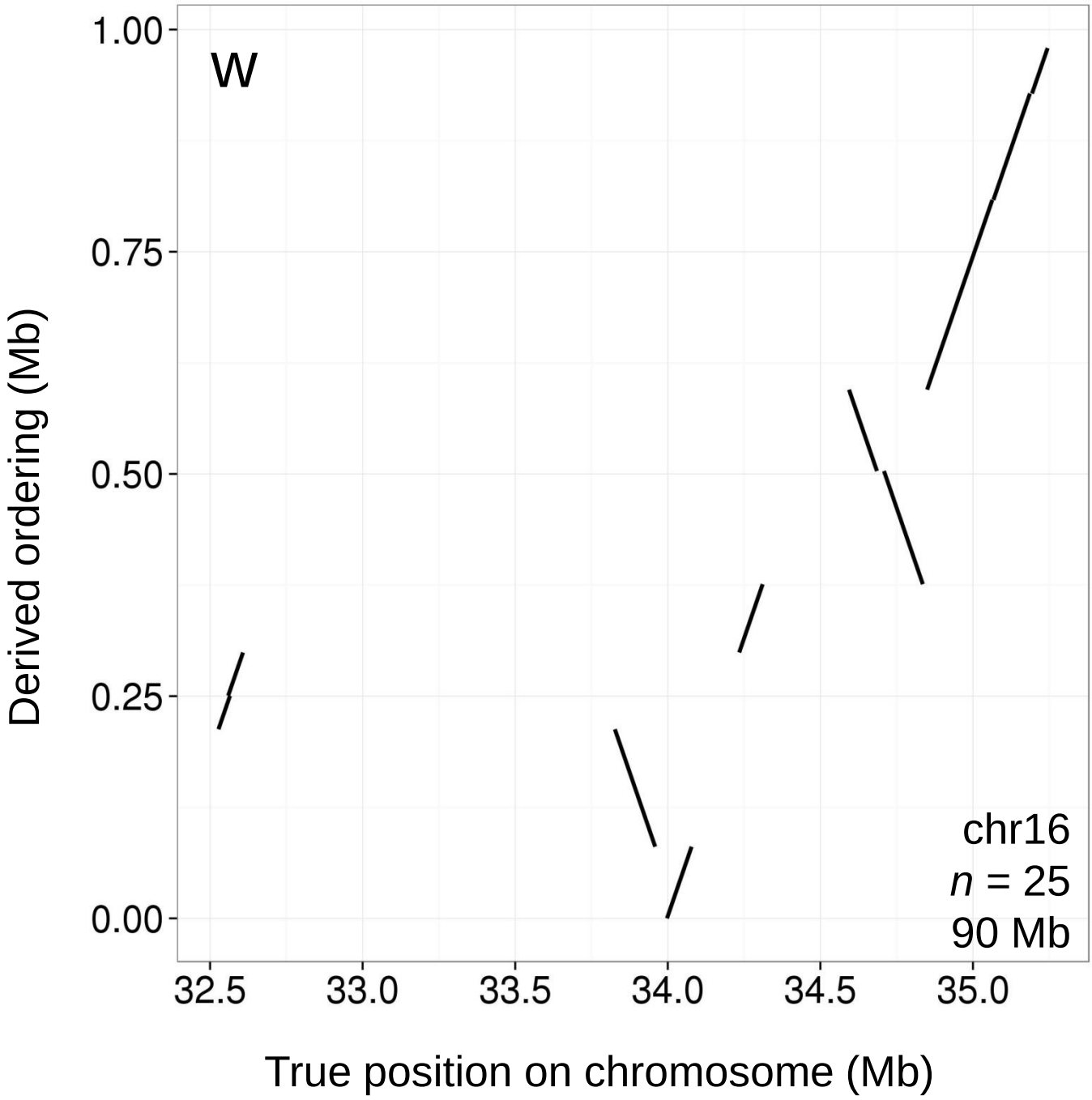
Supplementary Figure 4 (page 4 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3**. **w**, the chimeric group not shown in **Figure 3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.



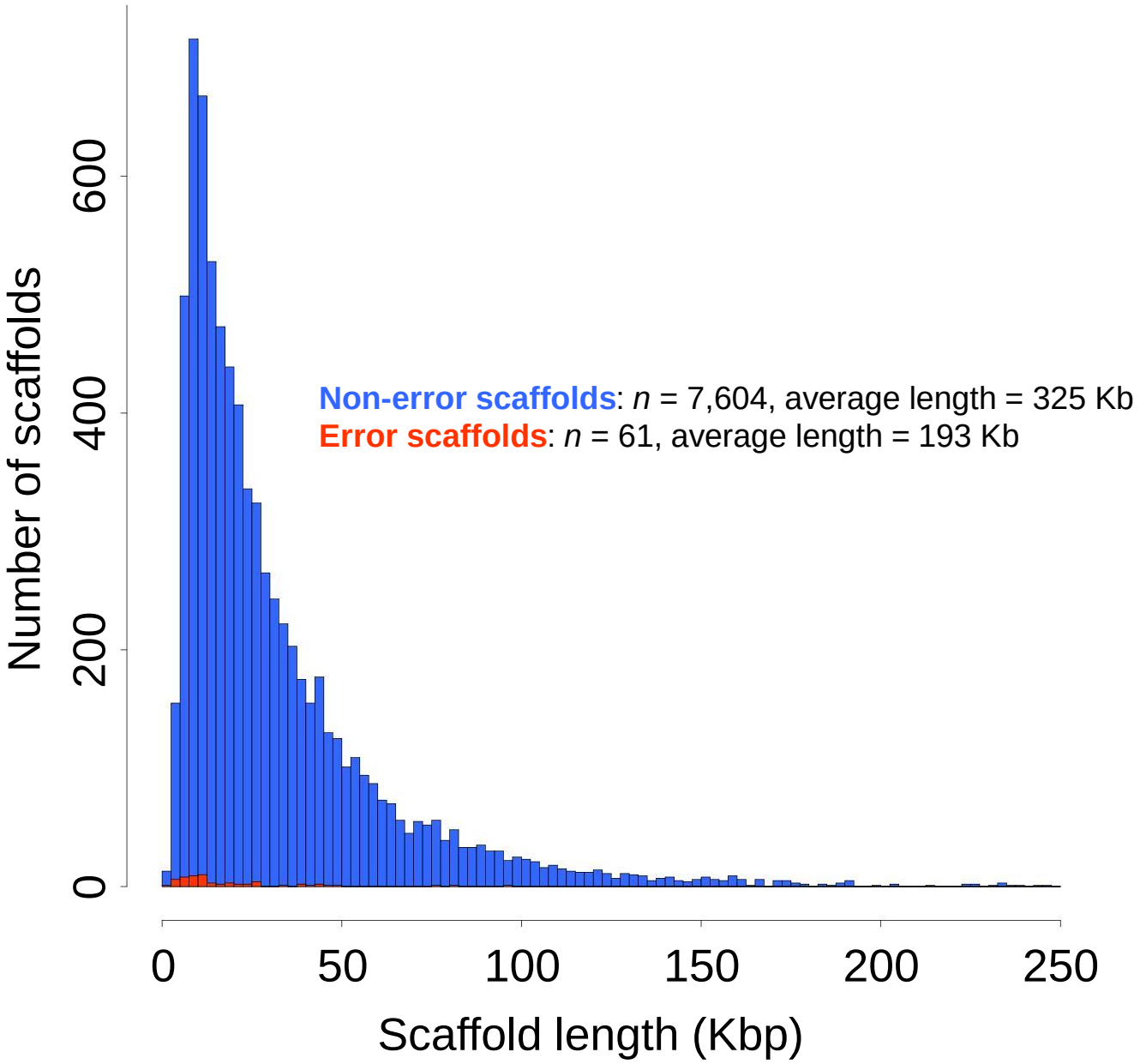
Supplementary Figure 4 (page 5 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3**. **w**, the chimeric group not shown in **Figure 3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.



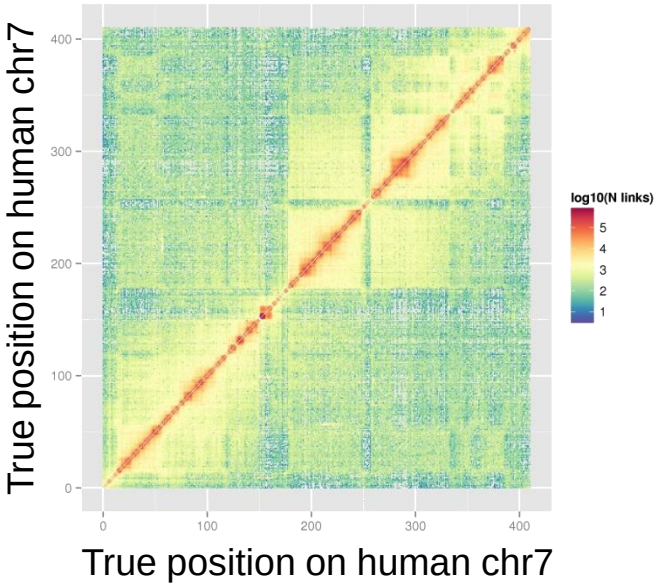
Supplementary Figure 4 (page 6 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3**. **w**, the chimeric group not shown in **Figure 3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.

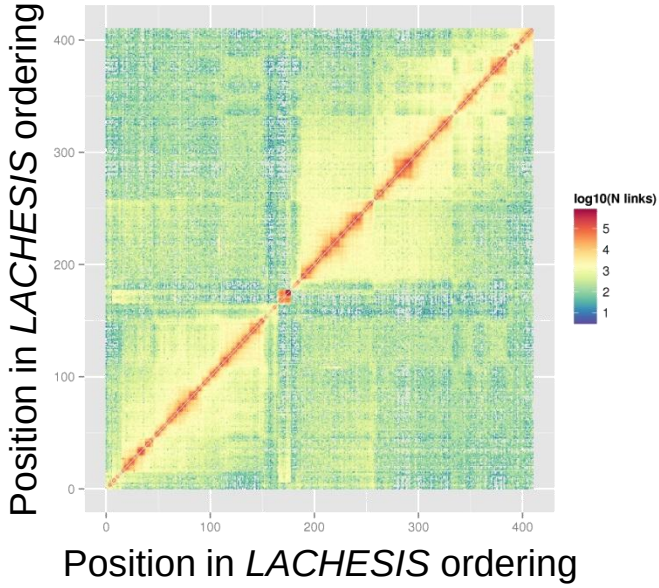


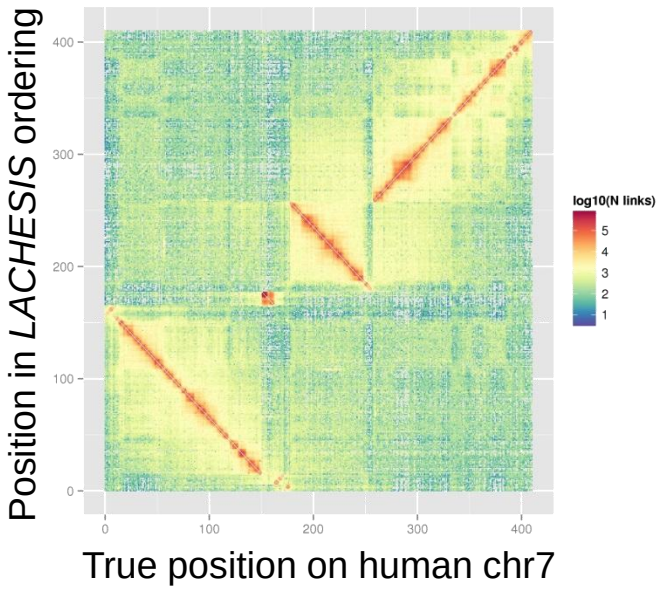
Supplementary Figure 4 (page 7 of 7) | *LACHESIS* ordering and orienting results on the 23 groups of scaffolds in the human *de novo* assembly. Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome. These plots are larger versions of the plots in **Figure 3**. **w**, the chimeric group not shown in **Figure 3**, showing the small region on chromosome 16 that constitutes the dominant chromosome for this group.



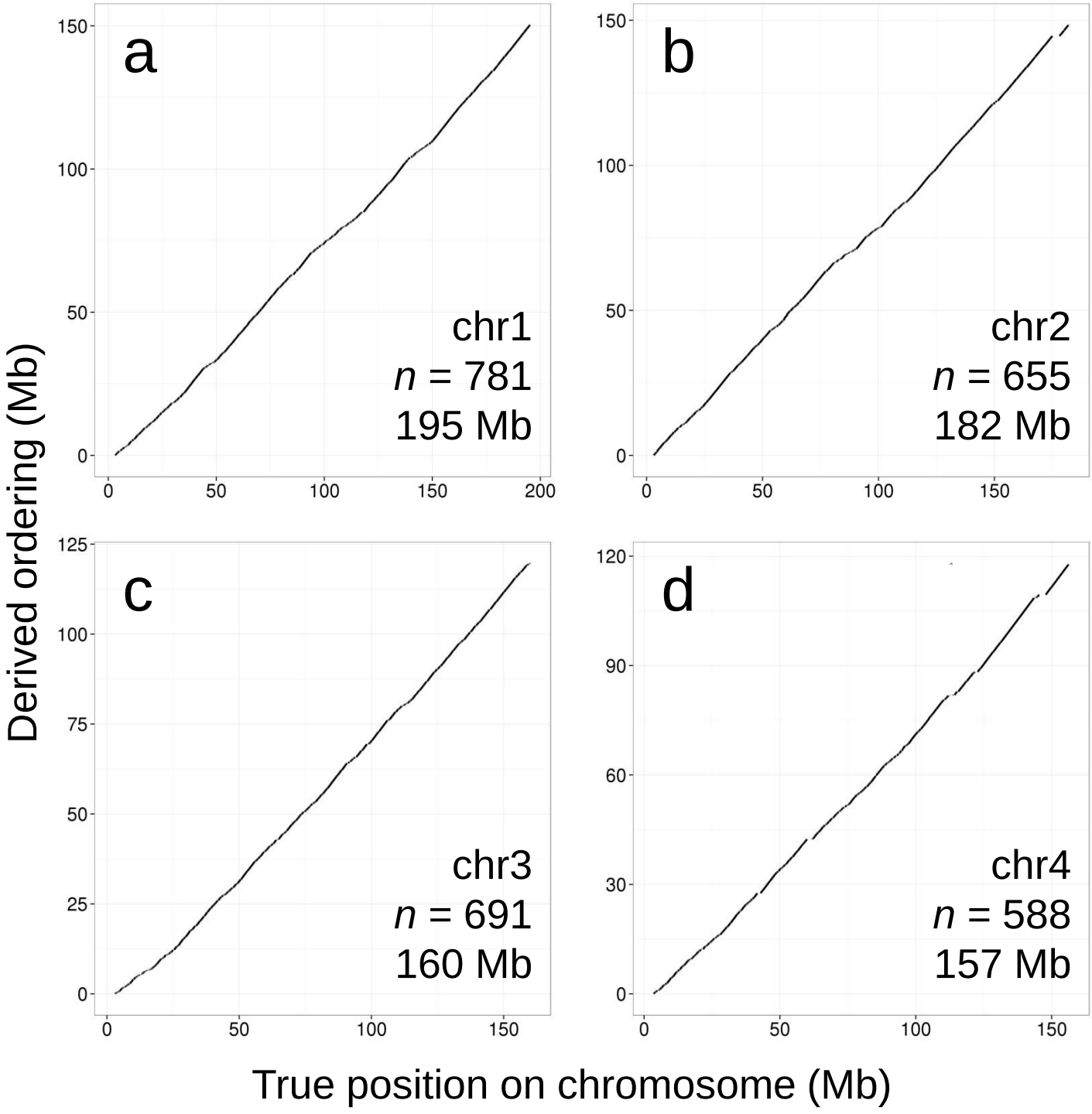
Supplementary Figure 5 | Scaffolds associated with ordering errors tend to be shorter than correctly ordered scaffolds. A histogram of the lengths of all scaffolds in the *de novo* human assembly which *LACHESIS* places in orderings and which map to the human reference. Scaffolds marked with ordering errors are shown in red; all other scaffolds are shown in blue. For clarity, six scaffolds of length >250 Kbp (none of which have ordering errors) are not shown.



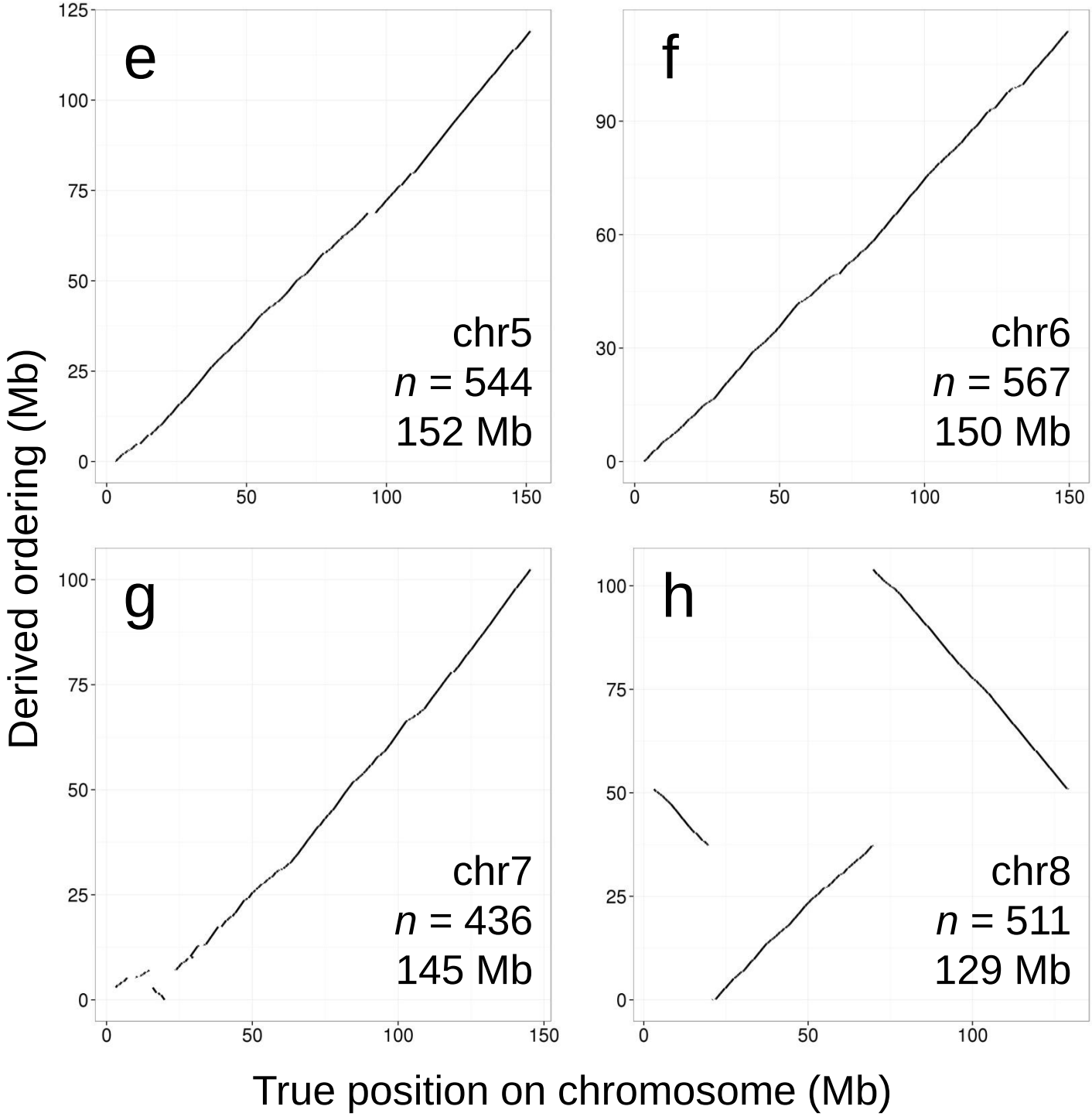




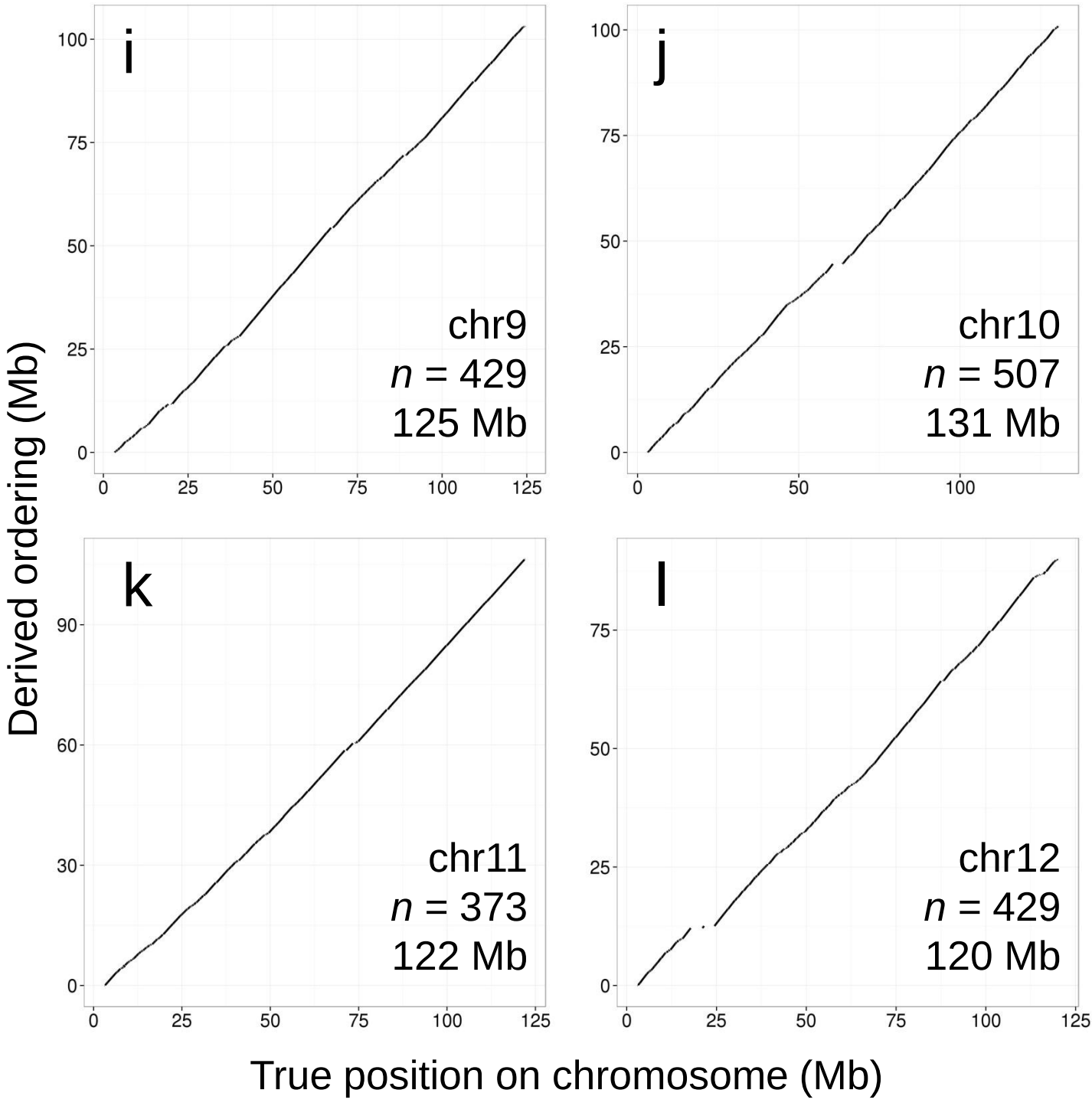
Supplementary Figure 6 | Example of *LACHESIS* assembly errors due to long-range chromatin interactions. Shown are three heatmaps of the density of Hi-C links between scaffolds of the *de novo* human assembly for chromosome 7. Only mapping contigs of length ≥10 Kb are shown. **a.** The scaffolds are ordered on both axes by their true position on chromosome 7. Note the presence of large domains with long-range internal interactions (squares along diagonal). **b.** The scaffolds are ordered on both axes by their position in the *LACHESIS* ordering in the group corresponding to chromosome 7. **c.** The scaffolds are ordered on the *x*-axis by their true position, and on the *y*-axis by their position in the *LACHESIS* ordering, revealing incorrect fusions of domains. Compare to **Supplementary Figure 4g**.



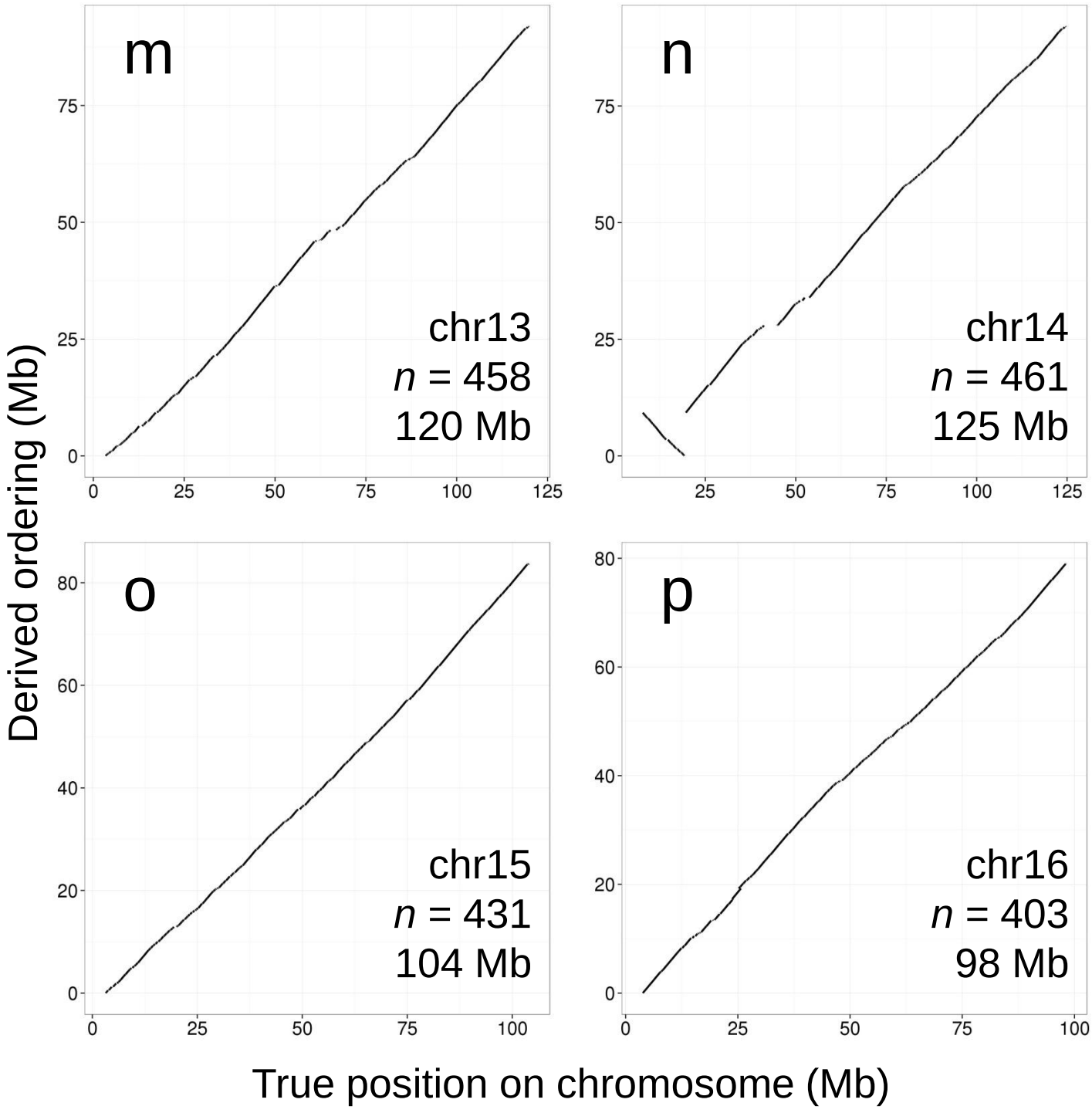
Supplementary Figure 7 (page 1 of 5) | *LACHESIS* ordering and orienting results on the 20 groups of scaffolds in the mouse *de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Supplementary Table 4**). Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome.



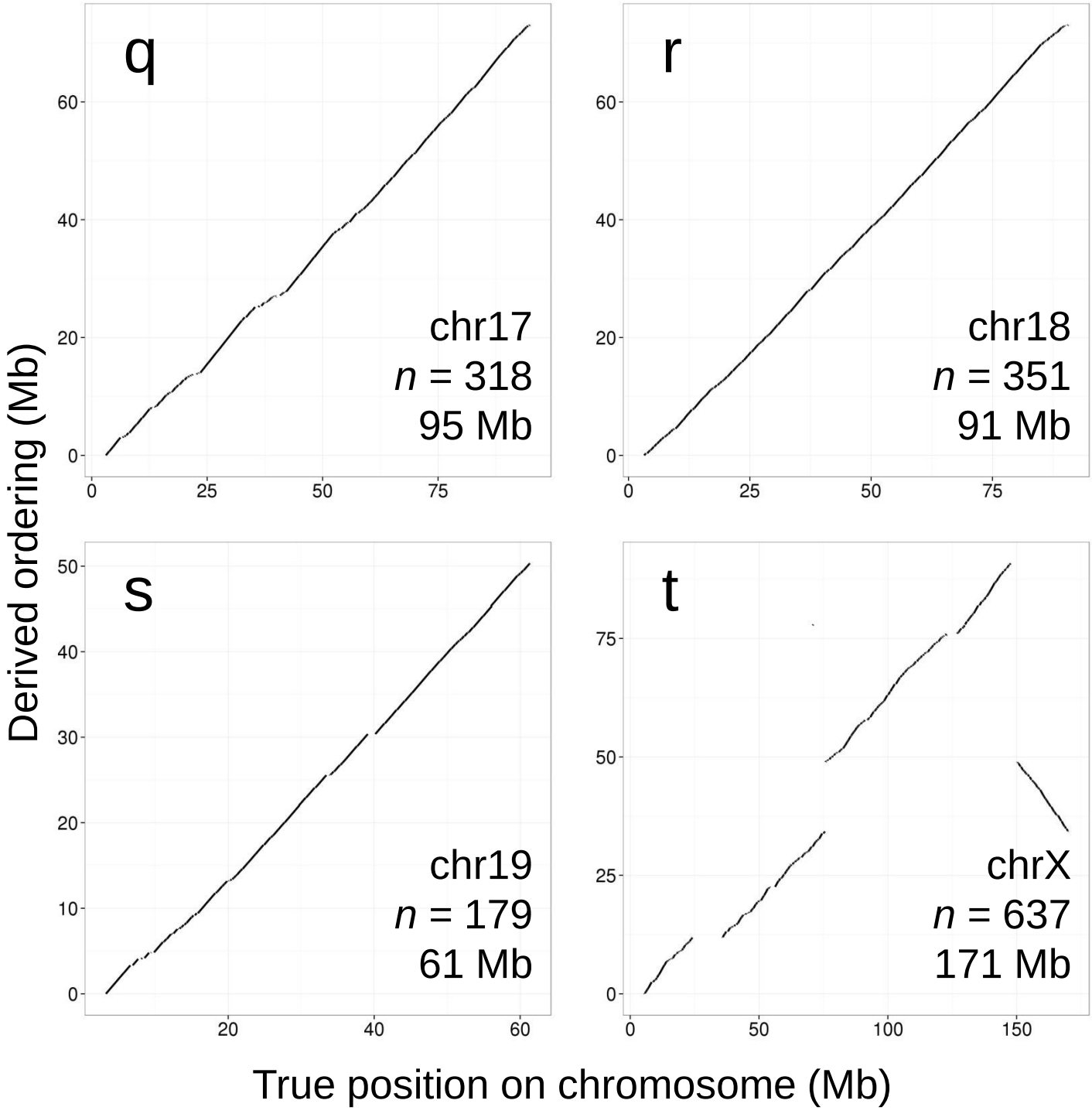
Supplementary Figure 7 (page 2 of 5) | *LACHESIS* ordering and orienting results on the 20 groups of scaffolds in the mouse *de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Supplementary Table 4**). Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome.



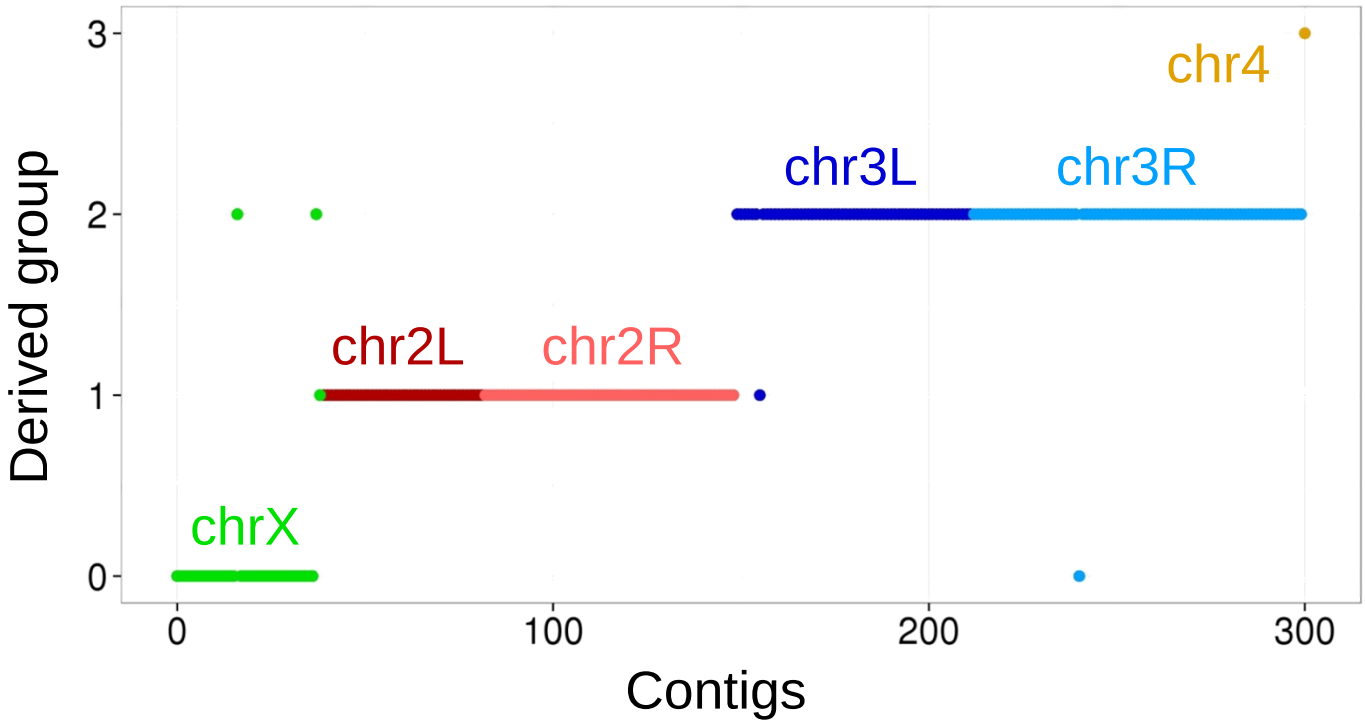
Supplementary Figure 7 (page 3 of 5) | *LACHESIS* ordering and orienting results on the 20 groups of scaffolds in the mouse *de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Supplementary Table 4**). Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome.



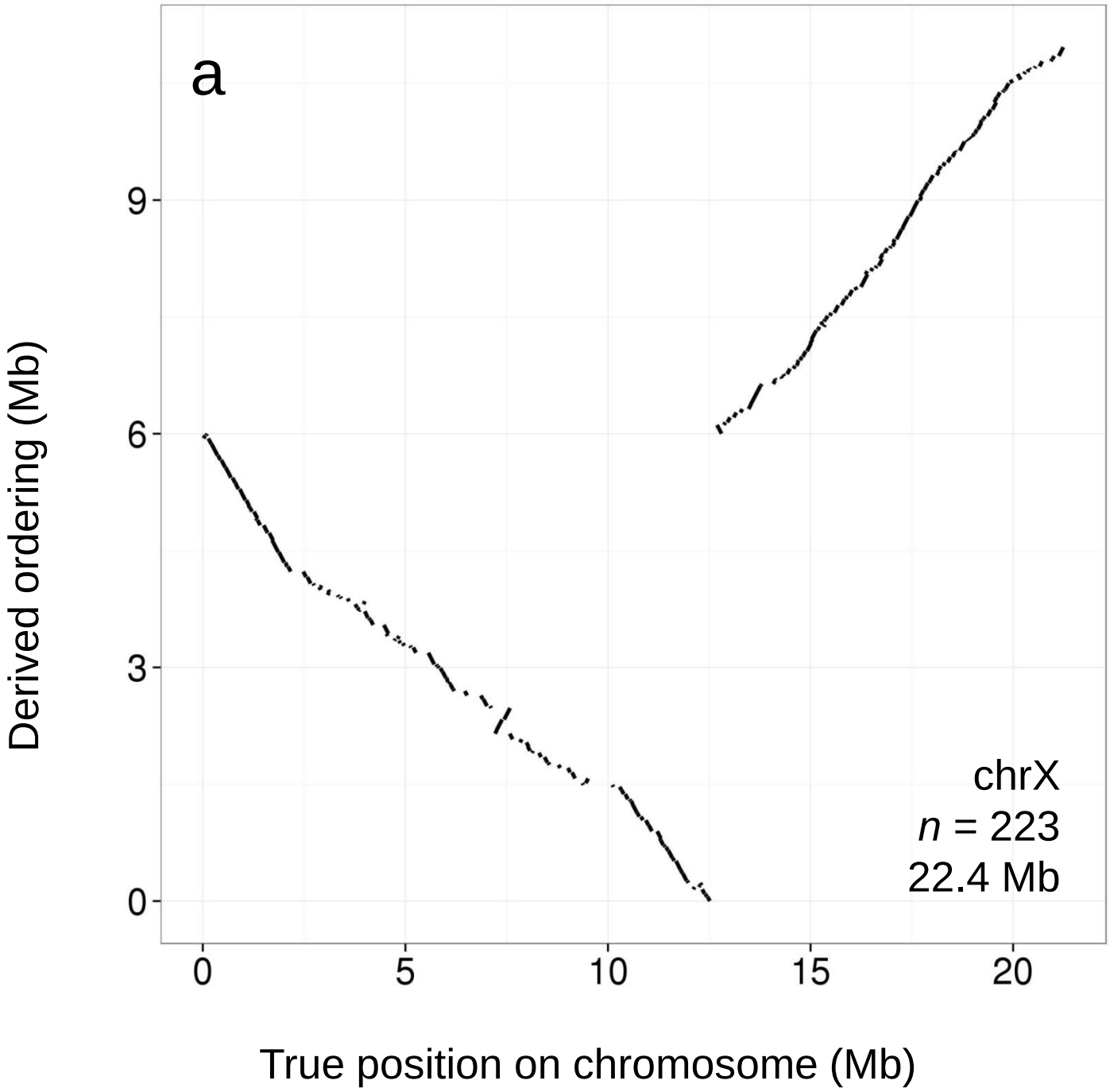
Supplementary Figure 7 (page 4 of 5) | *LACHESIS* ordering and orienting results on the 20 groups of scaffolds in the mouse *de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Supplementary Table 4**). Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome.



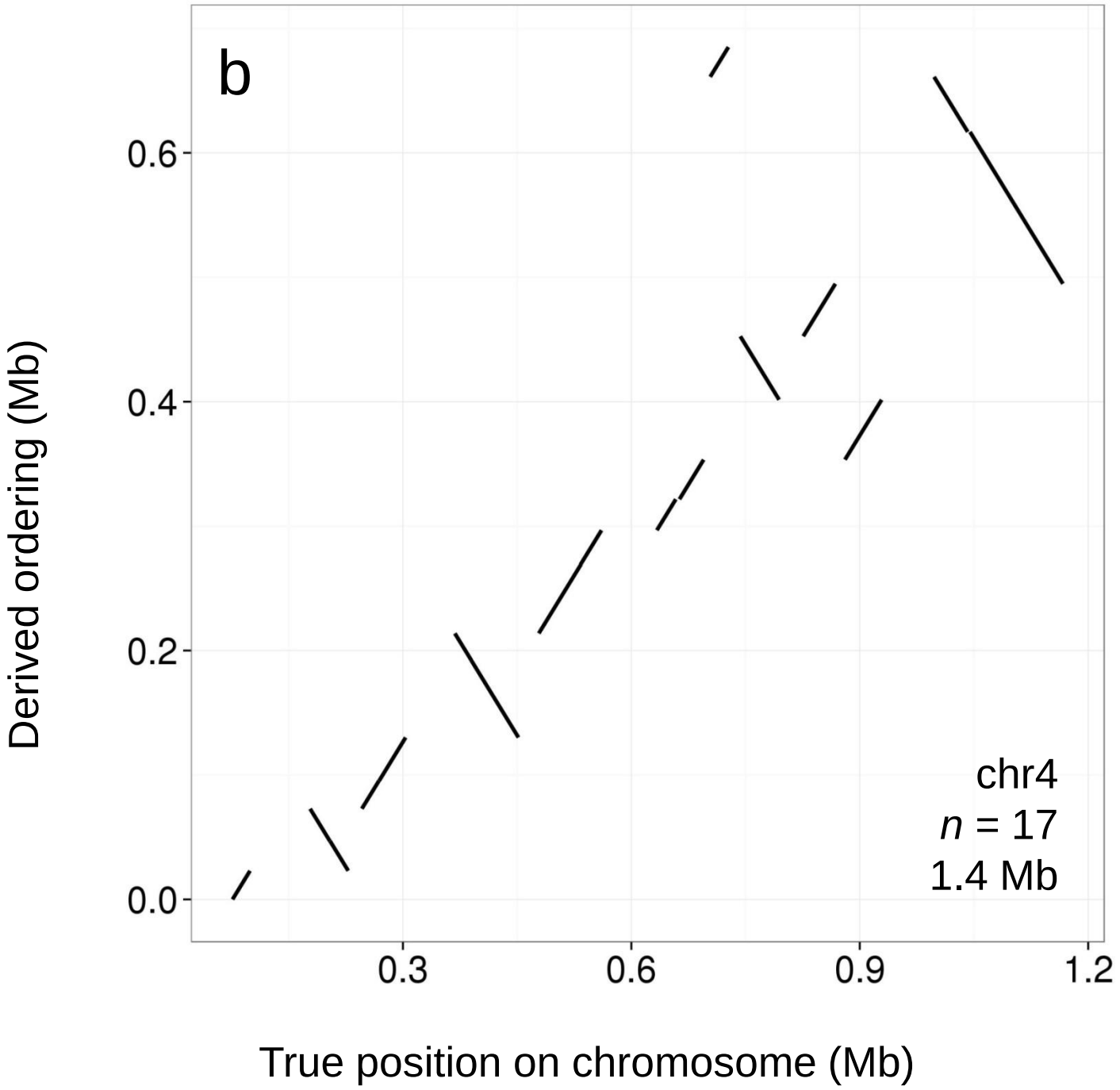
Supplementary Figure 7 (page 5 of 5) | *LACHESIS* ordering and orienting results on the 20 groups of scaffolds in the mouse *de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Supplementary Table 4**). Listed in each panel are the identity of the dominant chromosome, the number of scaffolds in the derived ordering, and the reference length of the dominant chromosome.



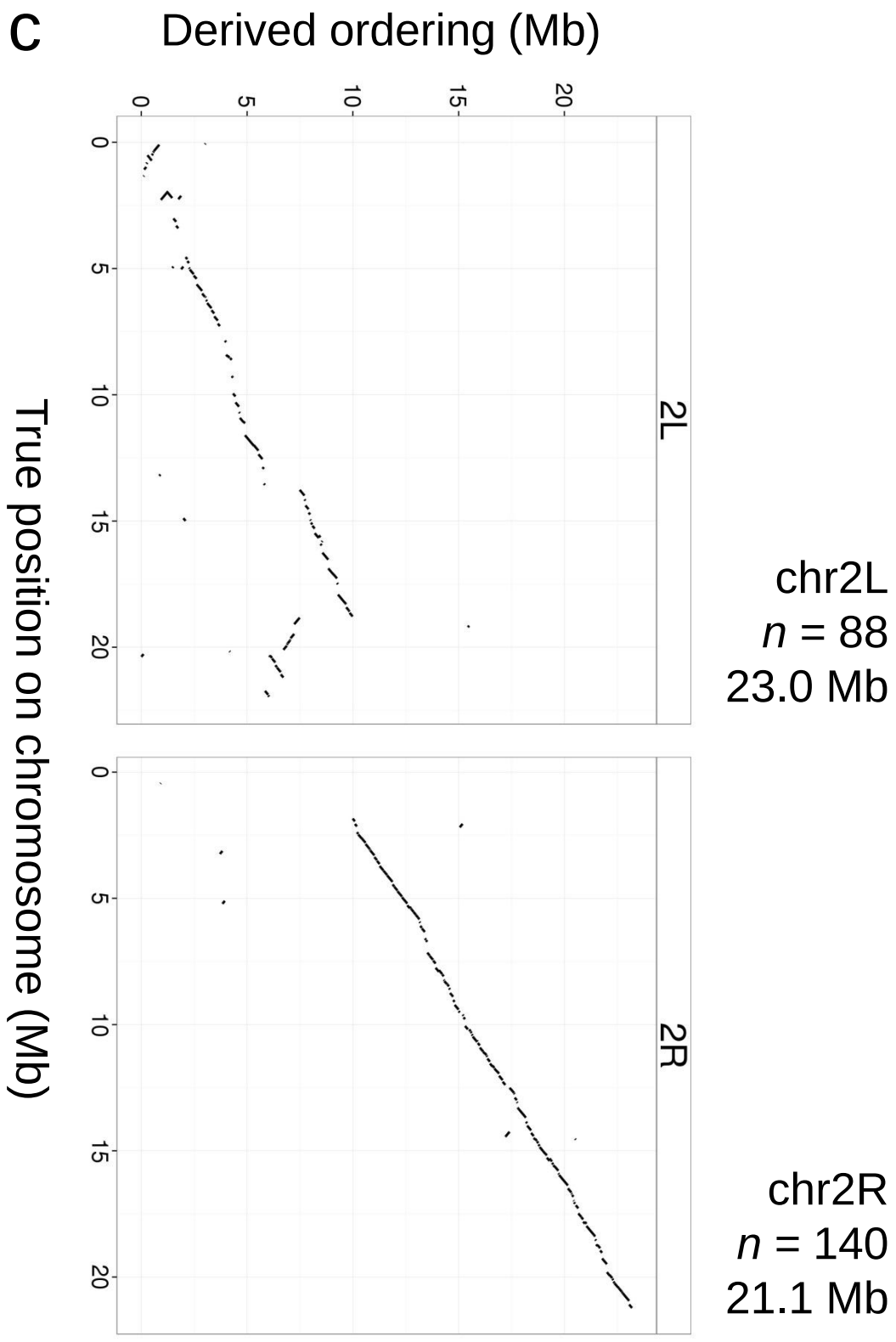
Supplementary Figure 8 | *LACHESIS* clustering results on the *Drosophila de novo* assembly. Shown on the *x*-axis are the 301 contigs (of 4,568 total contigs; total length: 43.8 Mb) that are long (≥250 GATC restriction sites) and not repetitive (Hi-C link density less than 2 times average), which *LACHESIS* used as informative for clustering. The *y*-axis shows the four groups created by *LACHESIS*, with the order chosen for the purposes of clarity. Each contig is shown as a dot, with a color indicating the chromosome to which the contig truly aligns, including the chromosome arm in the case of chromosomes 2 and 3.



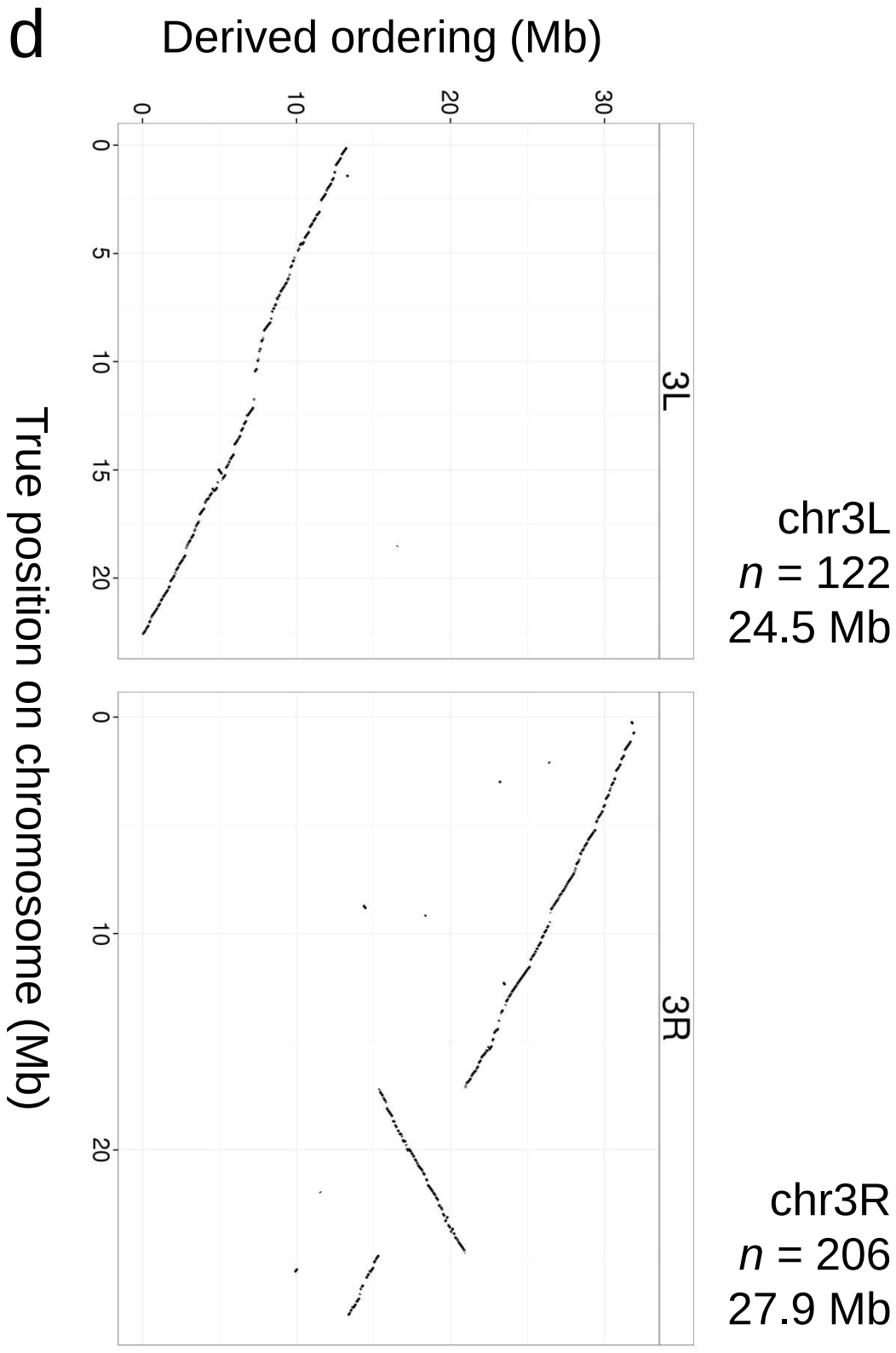
Supplementary Figure 9 (page 1 of 4) | *LACHESIS* ordering and orienting results on the 4 groups of contigs in the *Drosophila de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Supplementary Table 5**). Also listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.



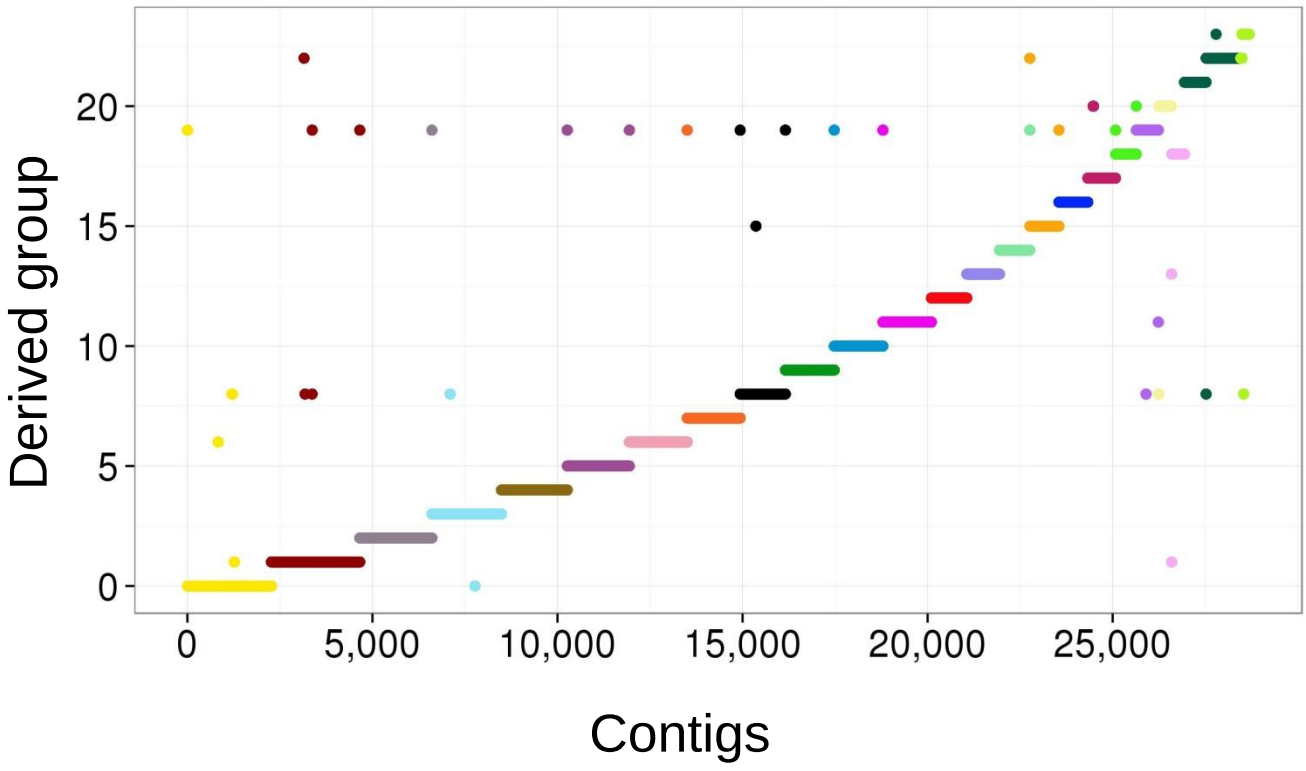
Supplementary Figure 9 (page 2 of 4) | *LACHESIS* ordering and orienting results on the 4 groups of contigs in the *Drosophila de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Supplementary Table 5**). Also listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.



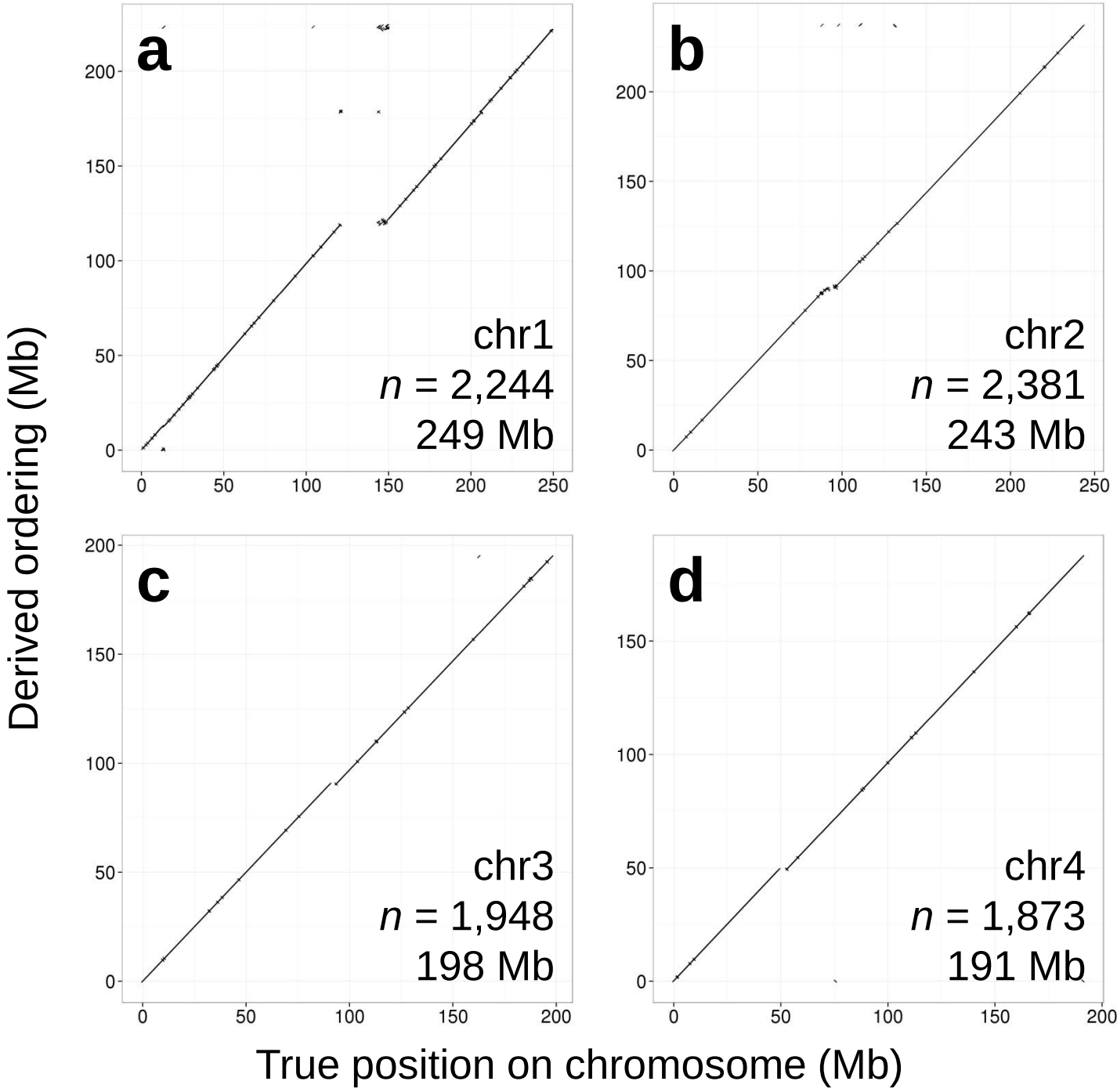
Supplementary Figure 9 (page 3 of 4) | *LACHESIS* ordering and orienting results on the 4 groups of contigs in the *Drosophila de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Supplementary Table 5**). Also listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.



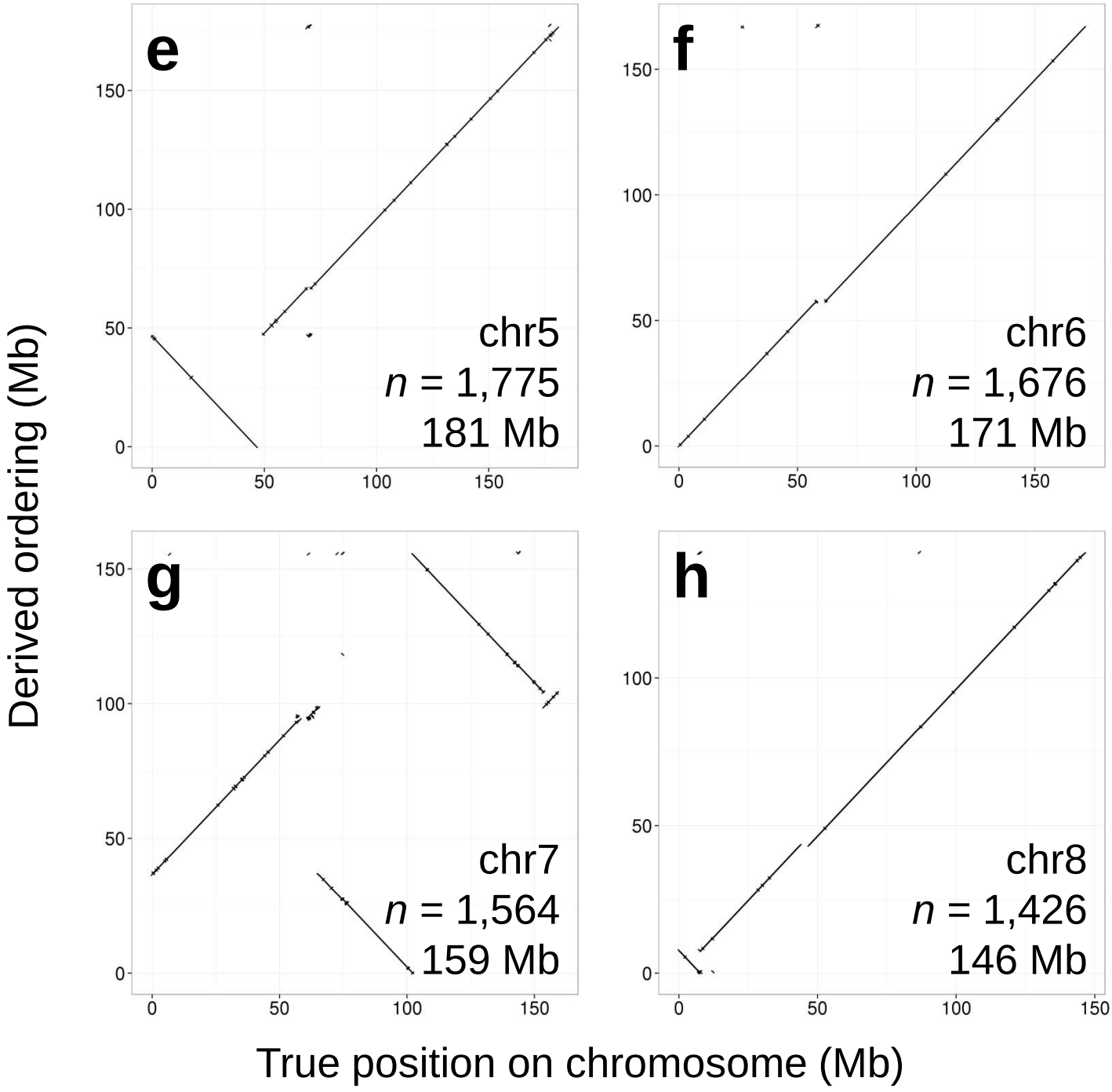
Supplementary Figure 9 (page 4 of 4) | *LACHESIS* ordering and orienting results on the 4 groups of contigs in the *Drosophila de novo* assembly. For each ordering, only the contigs on the "dominant chromosome" – that is, the chromosome containing the plurality of aligned sequence – are shown (*see* **Supplementary Table 5**). Also listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.



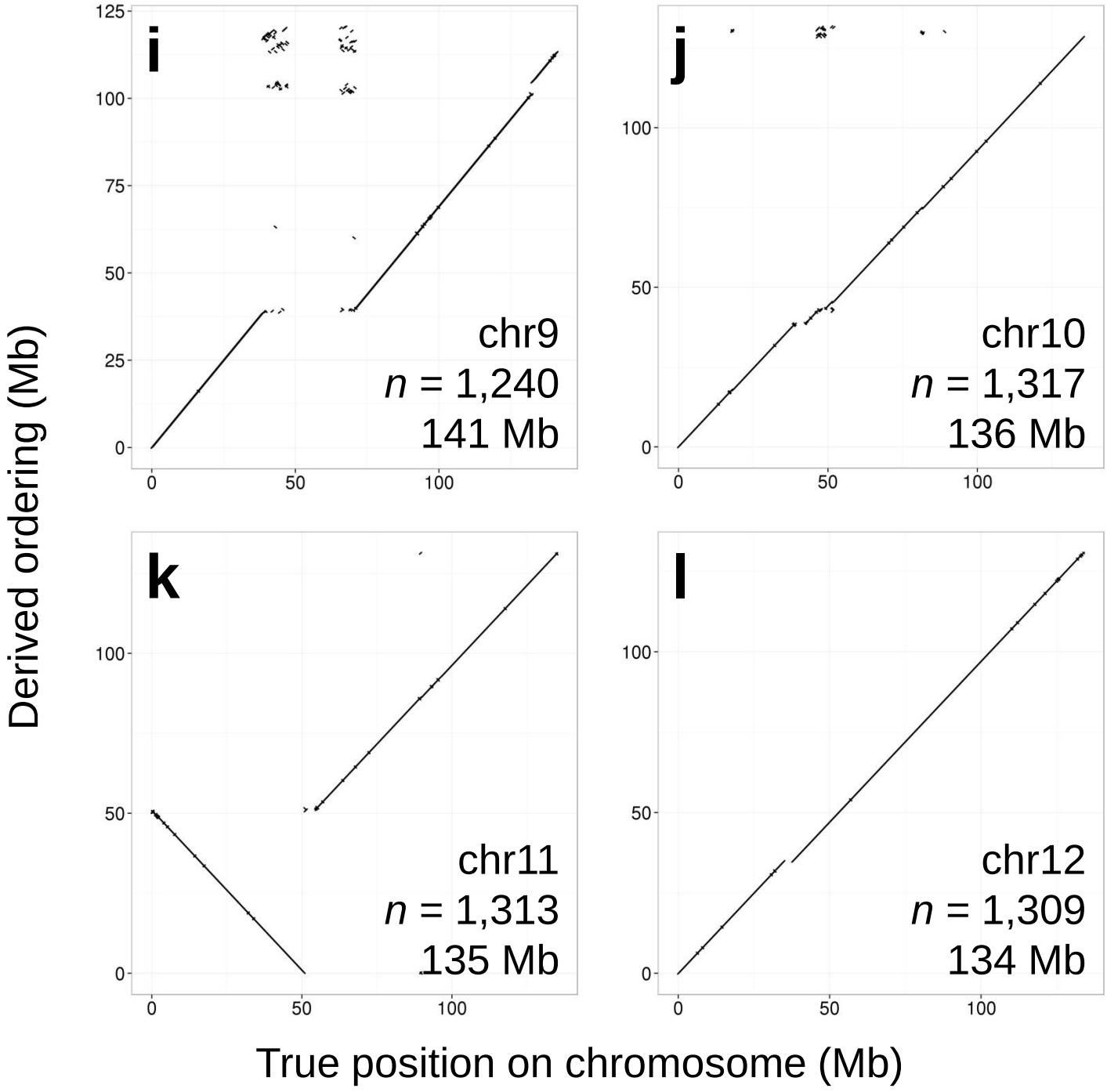
Supplementary Figure 10 | *LACHESIS* clustering results on simulated 100 Kb contigs of the human reference genome. The human genome was split into simulated 100 Kb contigs and *LACHESIS* was used to cluster these contigs into groups. The 28,689 clustered contigs (total length: 2.87 Gb) are ordered on the *x*-axis in order of ascending chromosome number and then position on the chromosome. The *y*-axis represents the 24 groups created by LACHESIS, with the order chosen for the purposes of clarity. Each 100 Kb contig is shown as a dot, with a color indicating the chromosome on which it belongs. The color scheme is the standard SKY (spectral karyotyping) color scheme for human. Not shown are the 2,281 contigs (7.4%) not placed into groups due to lack of unique sequence content, mostly corresponding to centromeres.



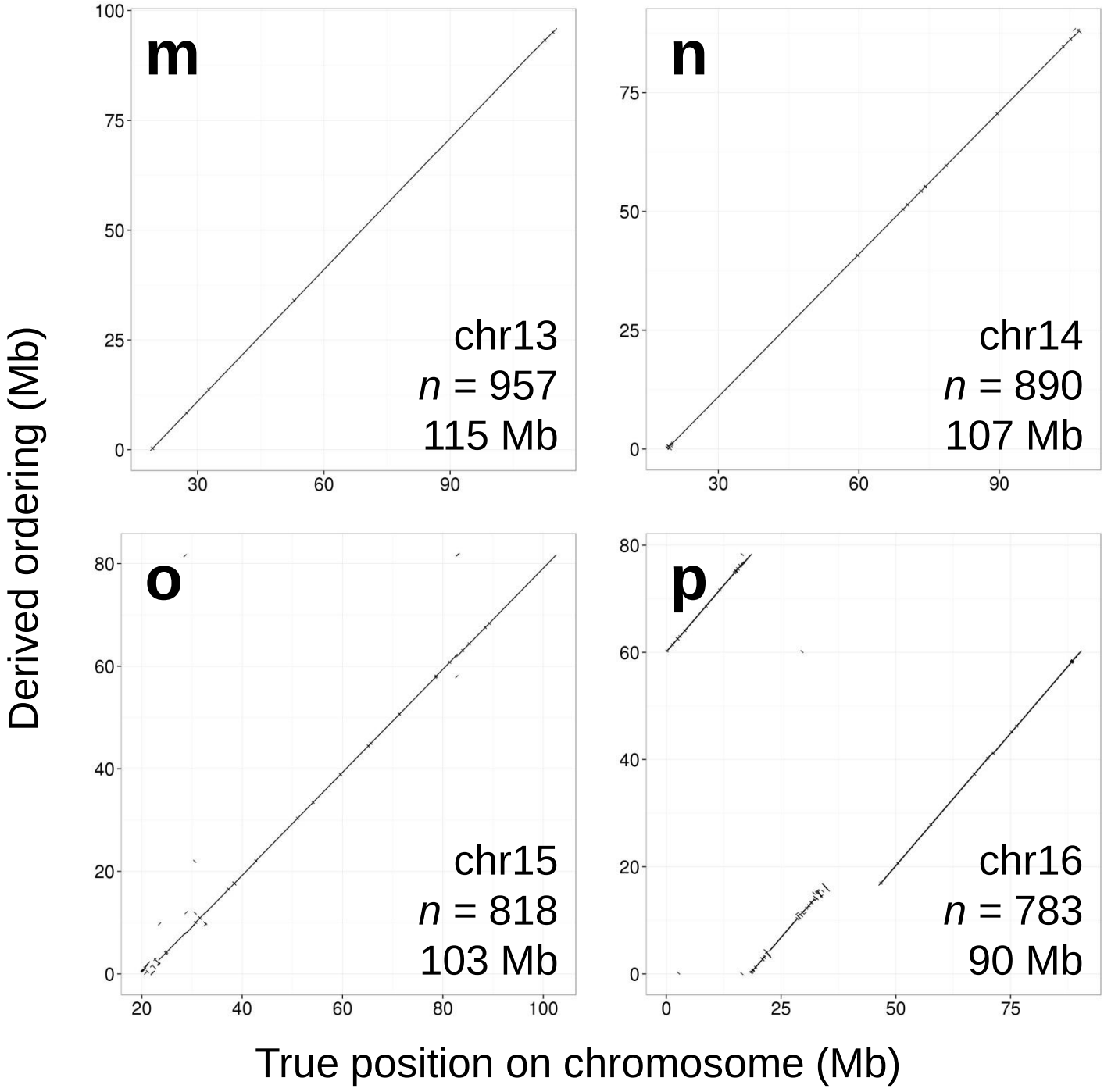
Supplementary Figure 11 (page 1 of 6) | *LACHESIS* ordering and orienting results on all 23 groups of simulated 100 Kb contigs in the human reference genome. Listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.



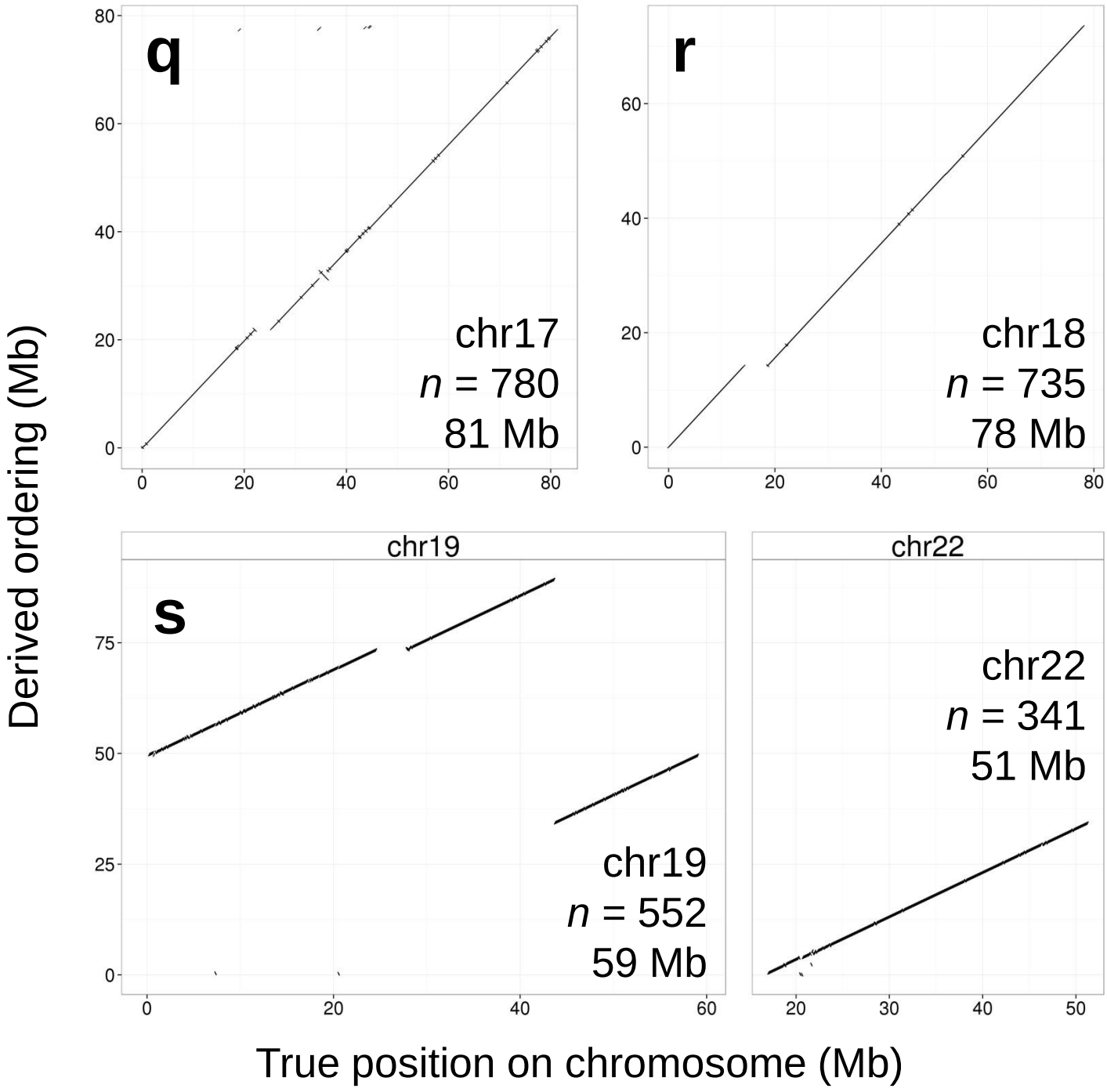
Supplementary Figure 11 (page 2 of 6) | *LACHESIS* ordering and orienting results on all 23 groups of simulated 100 Kb contigs in the human reference genome. Listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.



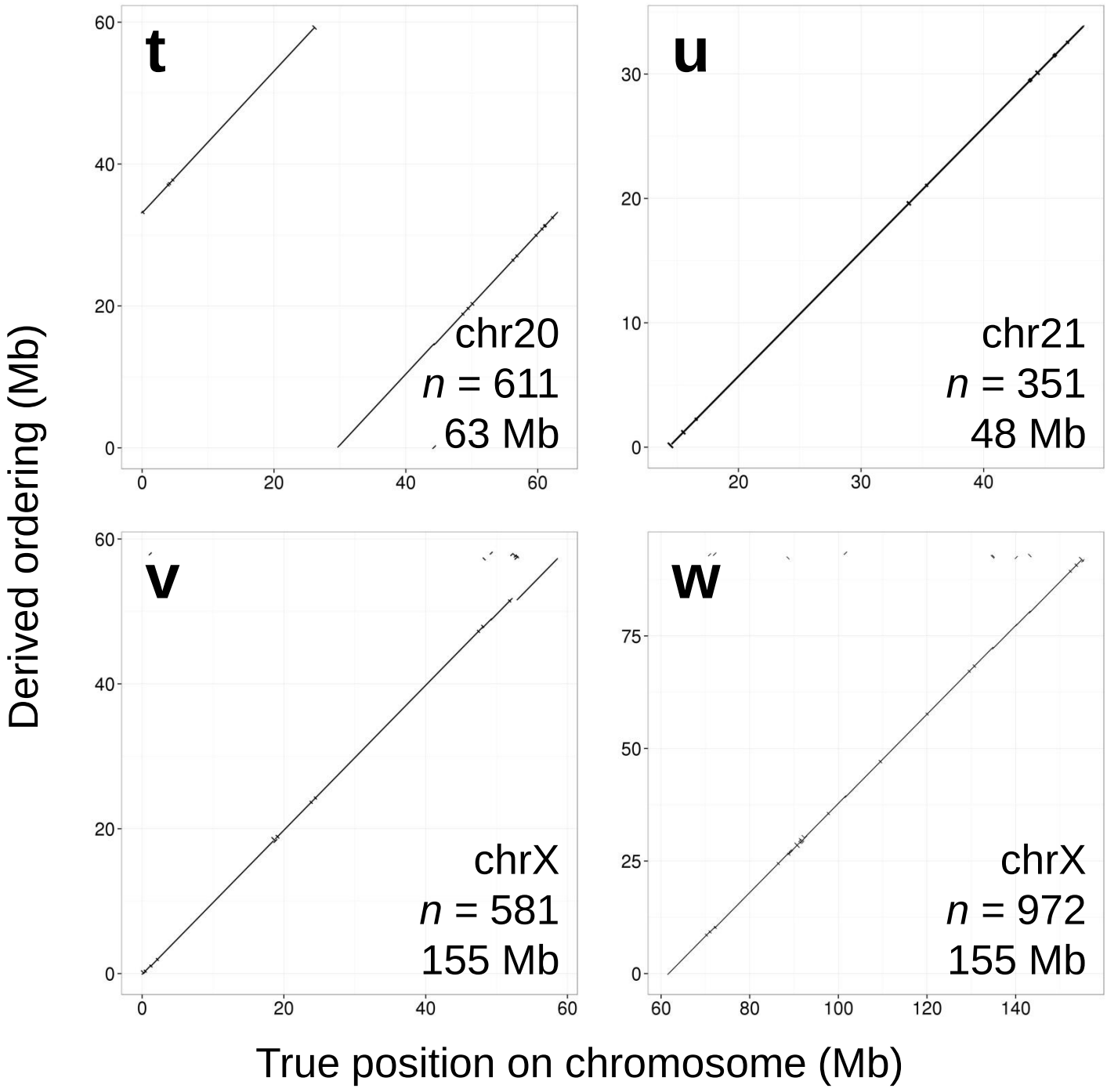
Supplementary Figure 11 (page 3 of 6) | *LACHESIS* ordering and orienting results on all 23 groups of simulated 100 Kb contigs in the human reference genome. Listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.



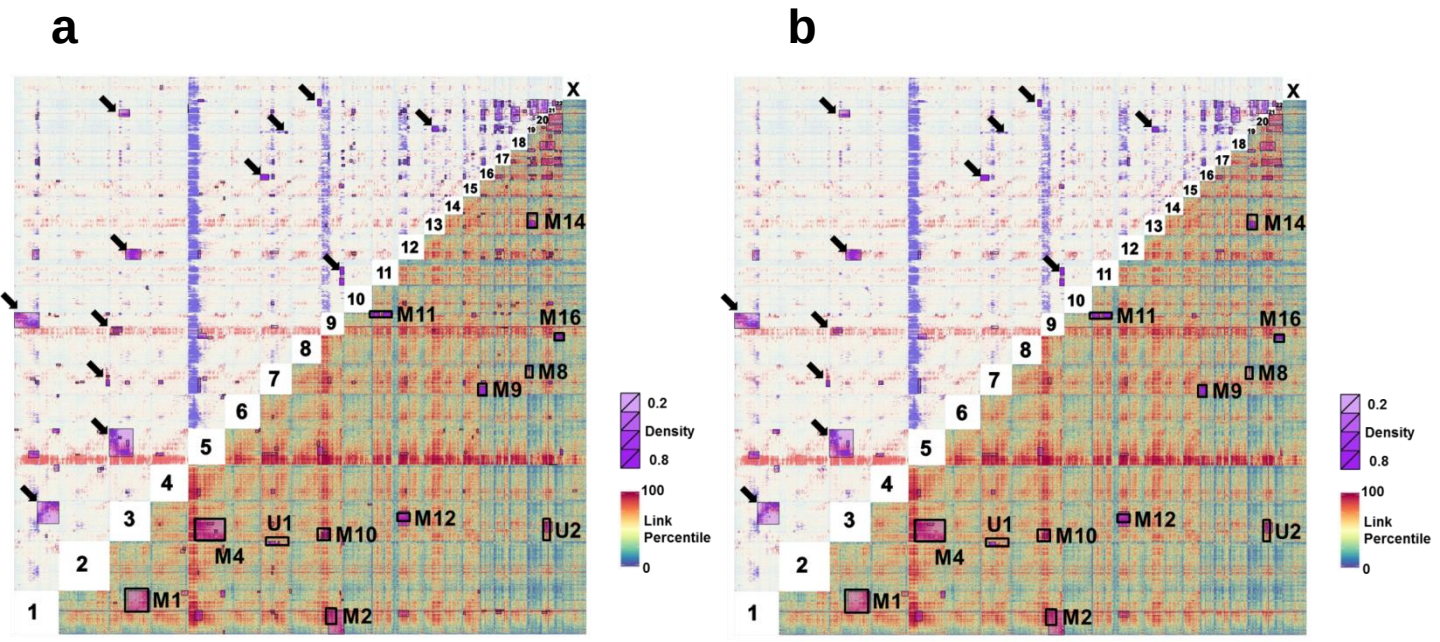
Supplementary Figure 11 (page 4 of 6) | *LACHESIS* ordering and orienting results on all 23 groups of simulated 100 Kb contigs in the human reference genome. Listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.



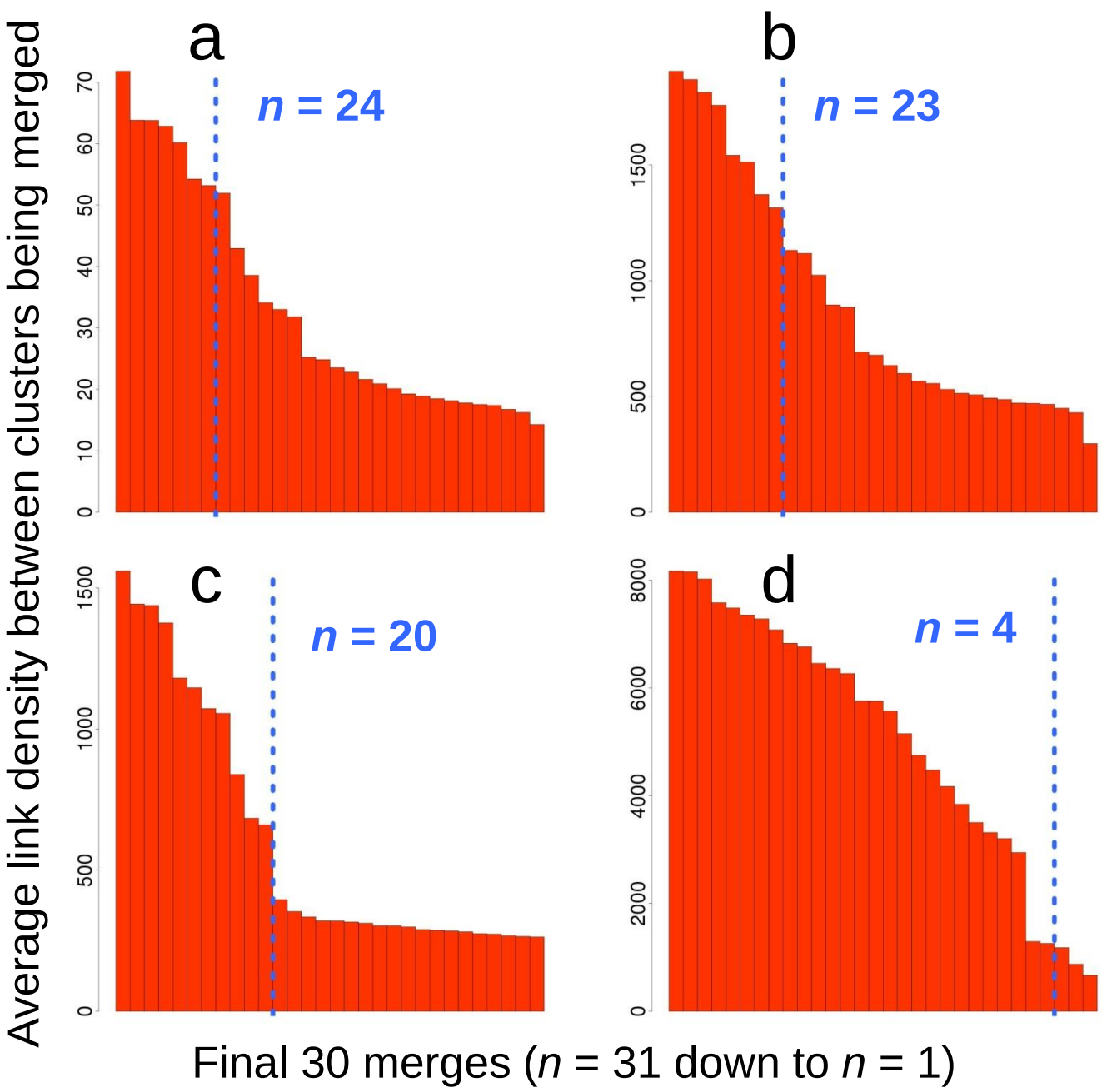
Supplementary Figure 11 (page 5 of 6) | *LACHESIS* ordering and orienting results on all 23 groups of simulated 100 Kb contigs in the human reference genome. Listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.



Supplementary Figure 11 (page 6 of 6) | *LACHESIS* ordering and orienting results on all 23 groups of simulated 100 Kb contigs in the human reference genome. Listed in each panel are the identity of the dominant chromosome, the number of contigs in the derived ordering, and the reference length of the dominant chromosome.



Supplementary Figure 12 | Using Hi-C to detect interchromosomal rearrangements in HeLa with high sensitivity. In the top left half of each image, blue horizontal lines represent outlying stretches of link scores with ≥10 windows, of which ≥80% of windows are ≥1 standard deviation above the mean of the row. Likewise, vertical red lines represent similar outlying stretches with respect to columns. Windows called as both row and column outliers are designated "outlier windows". Regions with excessive outlier stretch calls (*e.g.*, chromosome 5p) are only called as rows or columns and not likely both, thus reducing noise from globally high-scoring regions of the genome. Outlier window points are then clustered and called as potential fusions (purple boxes) and scored according to the density of outlier points within the window. **a.** An inclusive approach yields 100% sensitivity for detecting previously identified marker chromosomes, but only 8% specificity (assuming no additional marker chromosomes beyond those previously identified). False positive calls are largely due to increased interchromosomal contact among the smaller, gene-rich chromosomes, known to occur in healthy cells. **b.** Specificity can be increased by filtering based on cluster area. Specificity increases to 31% but sensitivity drops to 92%, with a bias towards rearrangements involving larger chromosomes or large regions of chromosomes.



Supplementary Figure 13 | Difficulty of calling the number of chromosomes from Hi-C link data alone. At each step of the *LACHESIS* clustering algorithm, the two clusters with the highest average link density are found and merged (**Supplementary Figure 1**). *x*-axis, the final 30 merges; *y*-axis, average link densities of each merge. The average link density decreases monotonically as merges are made. In practice, merging stops when a predetermined number of clusters is reached (blue line); importantly, this number is determined from *a priori* knowledge of the chromosome number rather than from the link densities shown here. **a.** The human simulated assembly with 100 Kb bins. **b.** The human *de novo* assembly. Note that the first several merges beyond $n = 23$, corresponding to fusions of the small chromosomes, have fairly high link densities. **c.** The mouse *de novo* assembly. **d.** The *Drosophila de novo* assembly. Note that the link densities imply $n = 6$, corresponding to a split of the arms of fly chromosomes 2 and 3, is a better solution than $n = 4$.