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

Lyon Lab & Schatz Lab Journal Club
Han Fang
1/21/2013

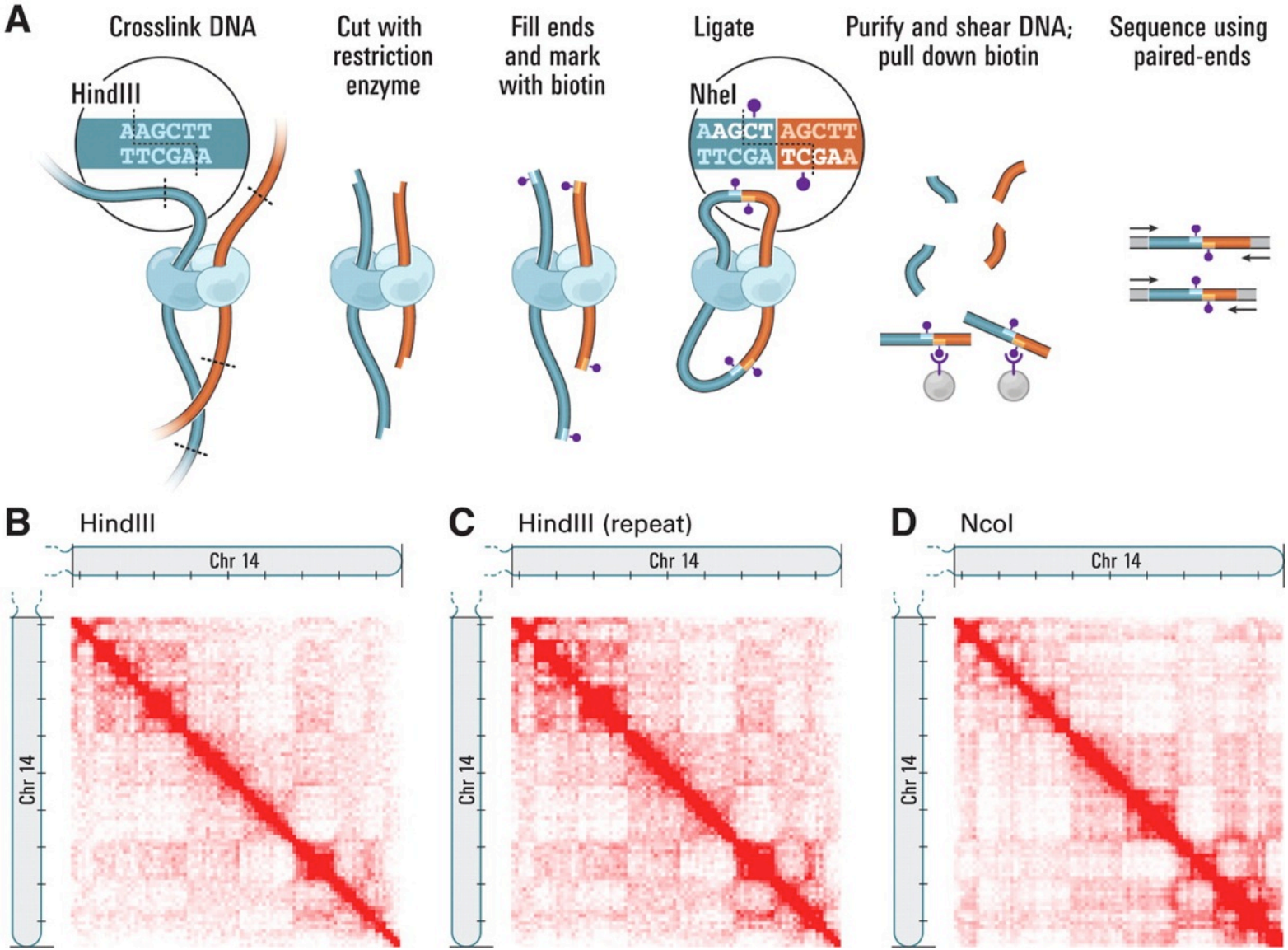
Science 9 October 2009:
Vol. 326 no. 5950 pp. 289-293
DOI: 10.1126/science.1181369

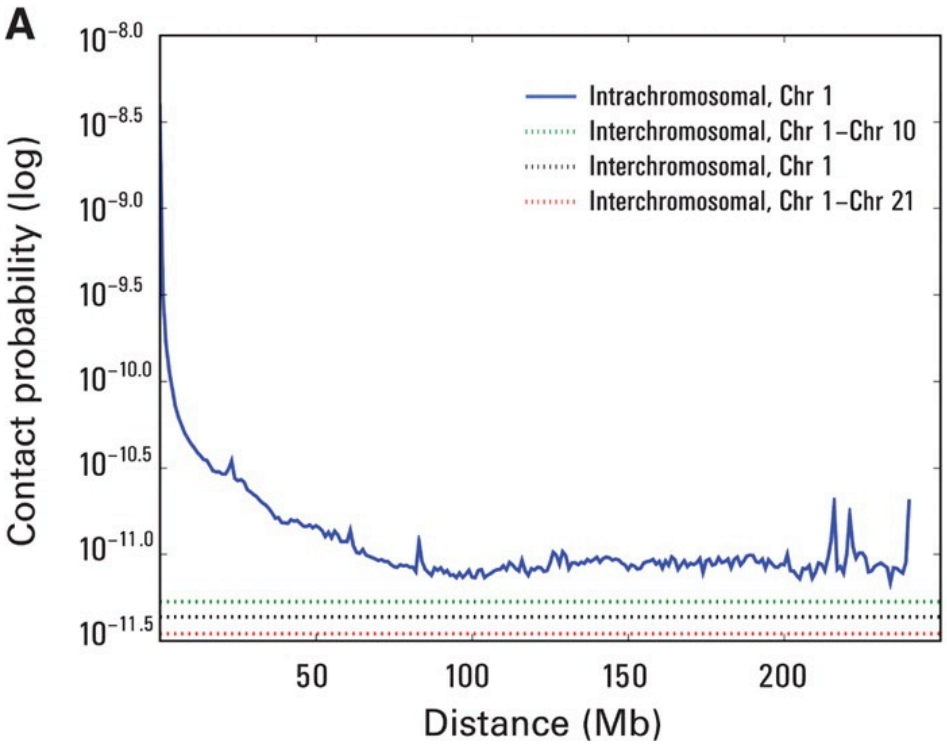
REPORT

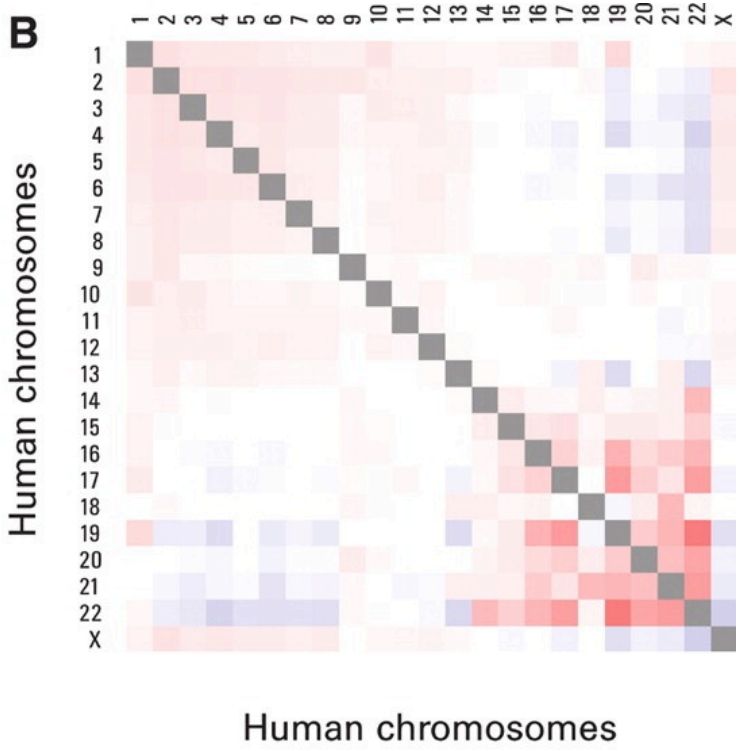
Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

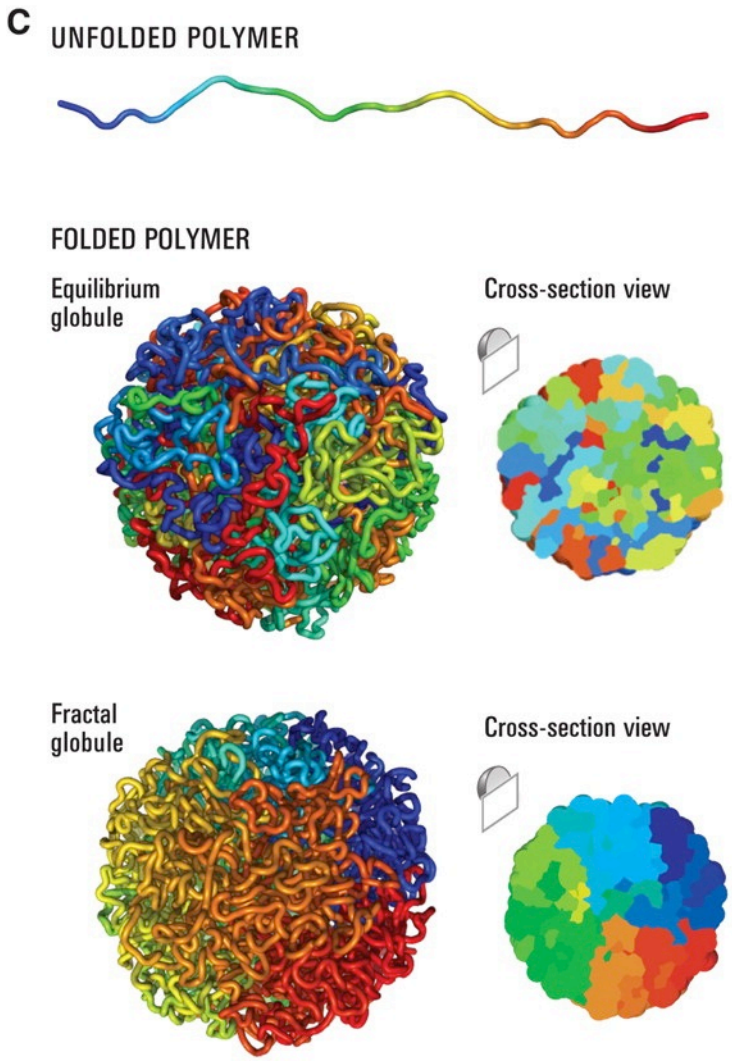
Erez Lieberman-Aiden[1,2,3,4,*], Nynke L. van Berkum[5,*], Louise Williams[1], Maxim Imakaev[2], Tobias Ragoczy[6,7], Agnes Telling[6,7], Ido Amit[1], Bryan R. Lajoie[5], Peter J. Sabo[8], Michael O. Dorschner[8], Richard Sandstrom[8], Bradley Bernstein[1,9], M. A. Bender[10], Mark Groudine[6,7], Andreas Gnirke[1], John Stamatoyannopoulos[8], Leonid A. Mirny[2,11], Eric S. Lander[1,12,13,†], Job Dekker[5,†]

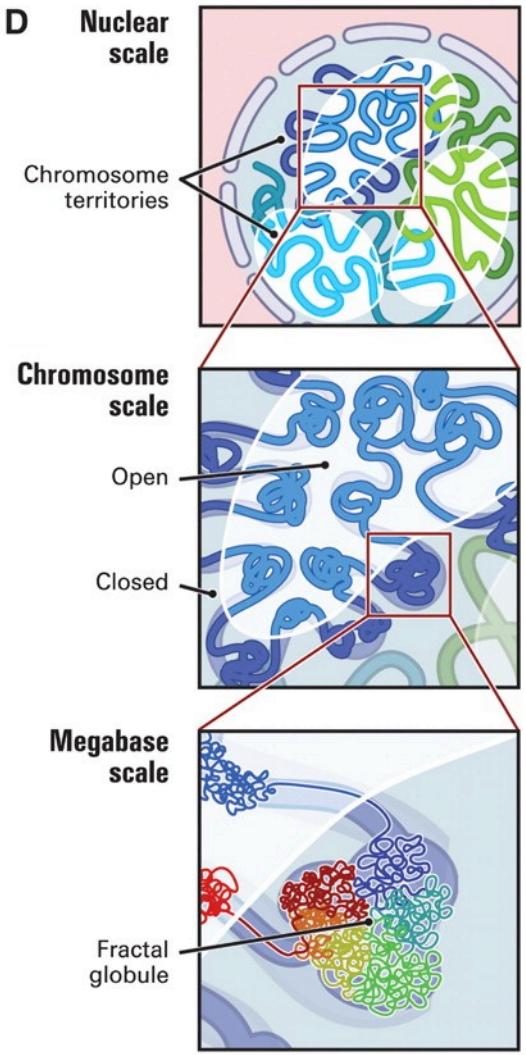
Lyon Lab


A
Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends
B
HindIII
Chr 14
Chr 14
C
HindIII (repeat)
Chr 14
Chr 14
D
NcoI
Chr 14
Chr 14


A
Contact probability (log)
$10^{-8.0}$
$10^{-8.5}$
$10^{-9.0}$
$10^{-9.5}$
$10^{-10.0}$
$10^{-10.5}$
$10^{-11.0}$
$10^{-11.5}$
Intrachromosomal, Chr 1
Interchromosomal, Chr 1–Chr 10
Interchromosomal, Chr 1
Interchromosomal, Chr 1–Chr 21
50
100
150
200
Distance (Mb)


B
Human chromosomes
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Human chromosomes


C
UNFOLDED POLYMER
FOLDED POLYMER
Equilibrium globule
Cross-section view
Fractal globule
Cross-section view


D
Nuclear scale
Chromosome territories
Chromosome scale
Open
Closed
Megabase scale
Fractal globule

NATURE BIOTECHNOLOGY | RESEARCH | ARTICLE

日本語要約

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

Joshua N Burton, Andrew Adey, Rupali P Patwardhan, Ruolan Qiu, Jacob O Kitzman & Jay Shendure

Affiliations | Contributions | Corresponding authors

Nature Biotechnology **31**, 1119–1125 (2013) | doi:10.1038/nbt.2727
Received 25 June 2013 | Accepted 02 October 2013 | Published online 03 November 2013

Speculation

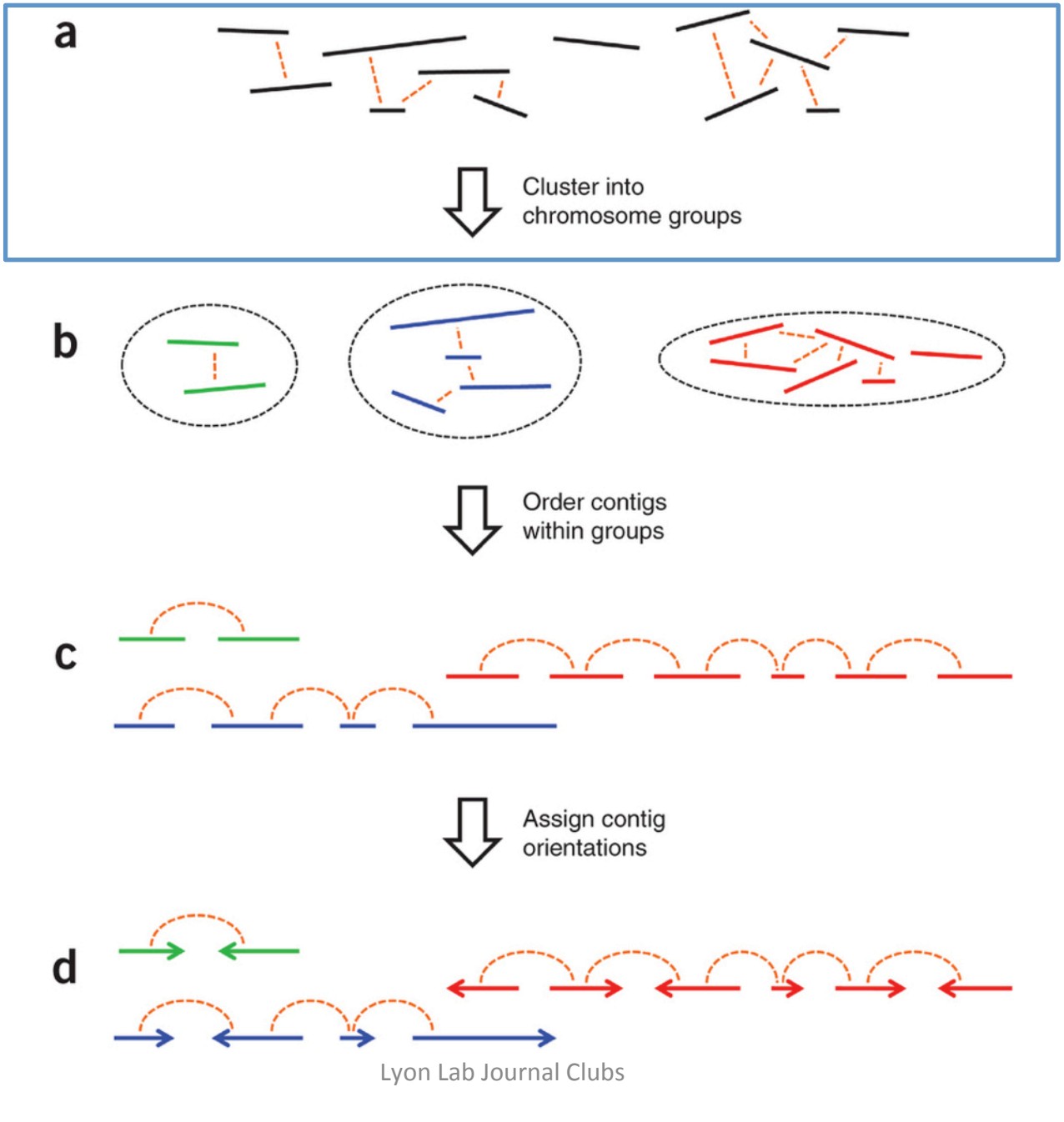
- Genome-wide chromatin interaction data sets might provide long-range information about the grouping and linear organization of sequences along entire chromosomes.

> LACHESIS (**l**igating **a**djacent **ch**romatin **e**nables **s**caffolding **i**n **s**itu)

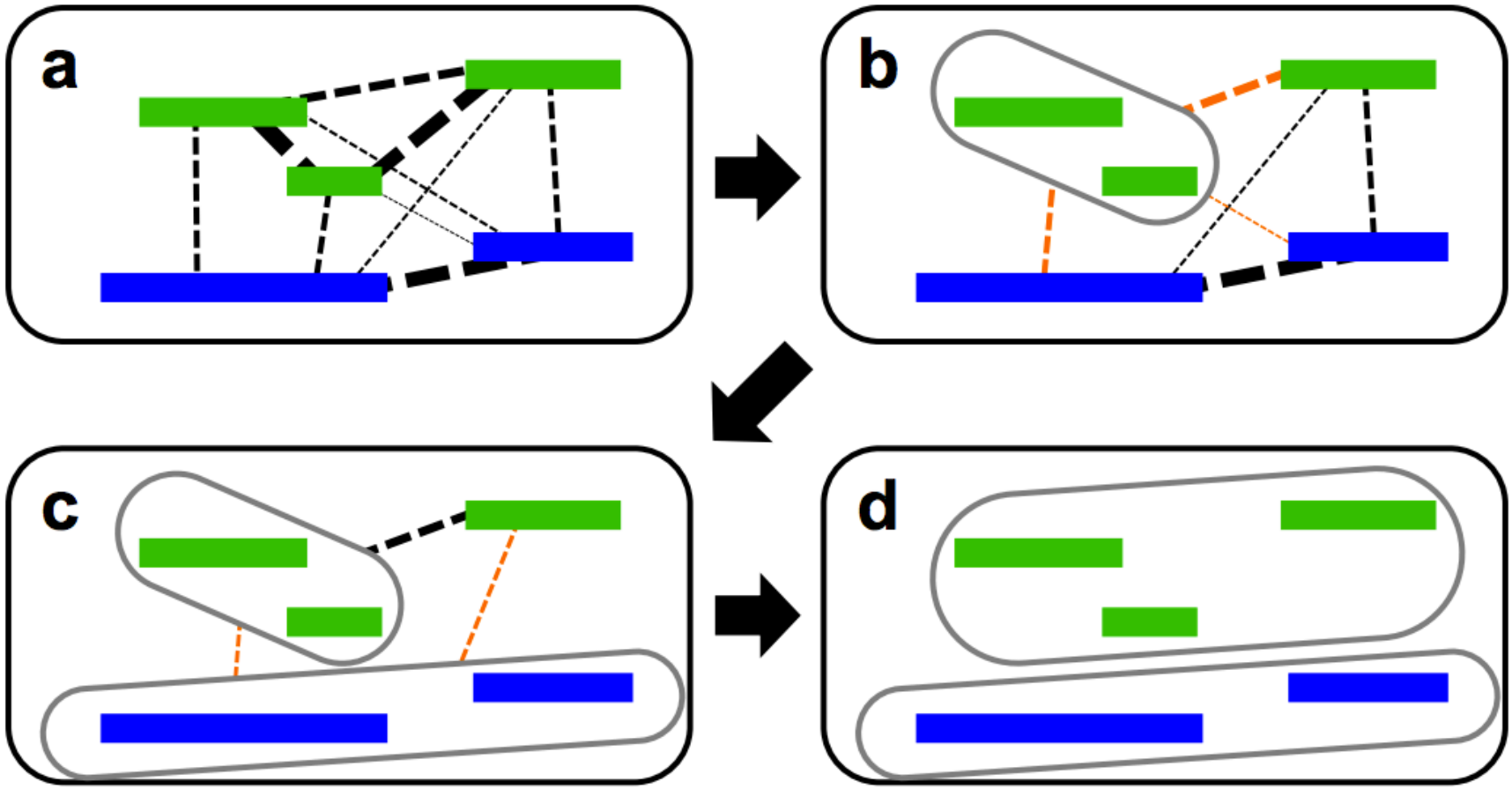
Overview

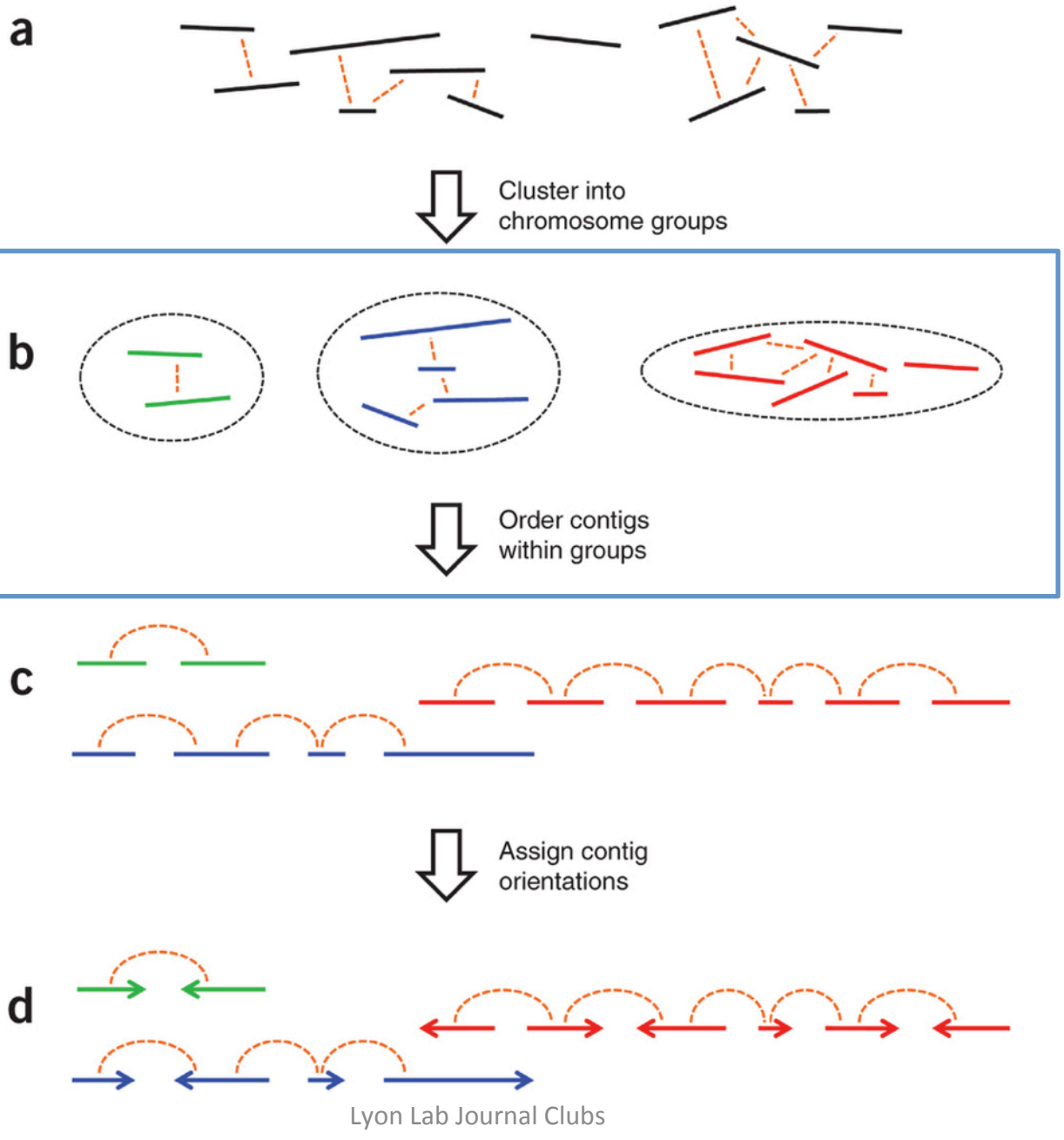
- Create shotgun assemblies (ALLPATH-LG)
- Align Hi-C reads to assemblies
- LACHESIS
  - Step1: **Cluster** contigs or scaffolds into chromosome groups
  - Step2: **Order** contigs or scaffolds within chromosome groups
  - Step3: **Orient** contigs or scaffolds

Lyon Lab

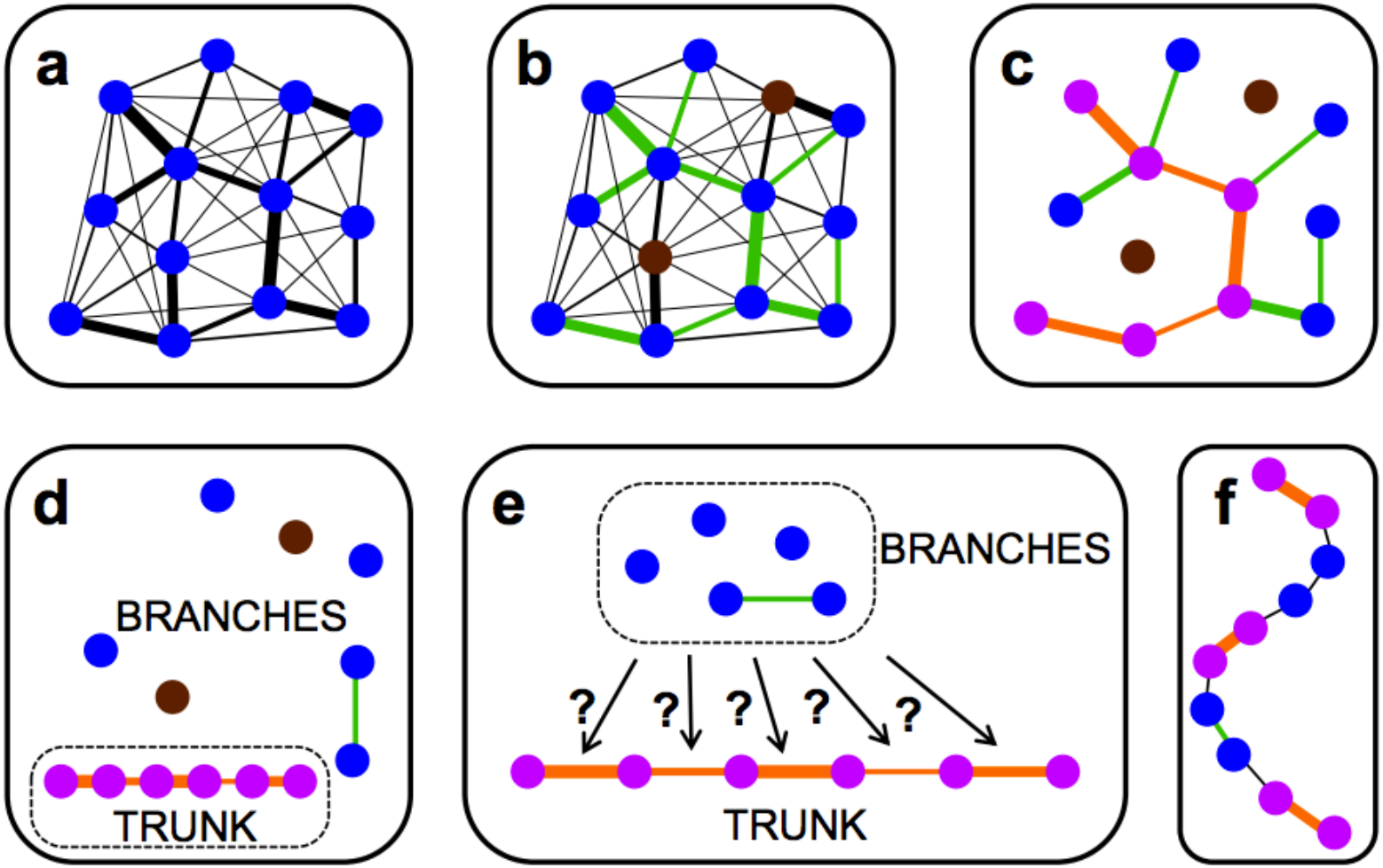

a
Cluster into
chromosome groups
b
Order contigs
within groups
c
Assign contig
orientations
d

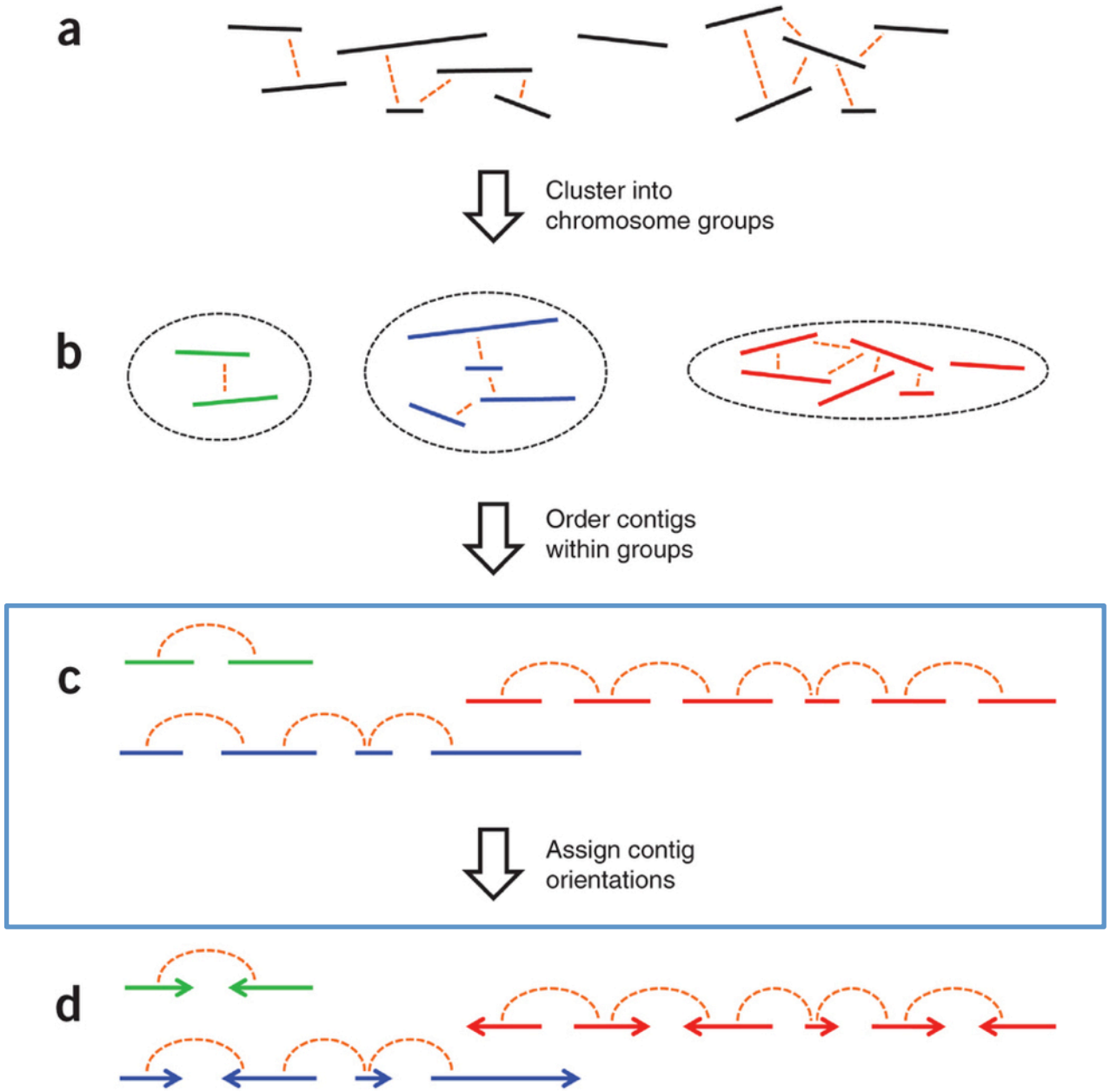
Clustering toy example



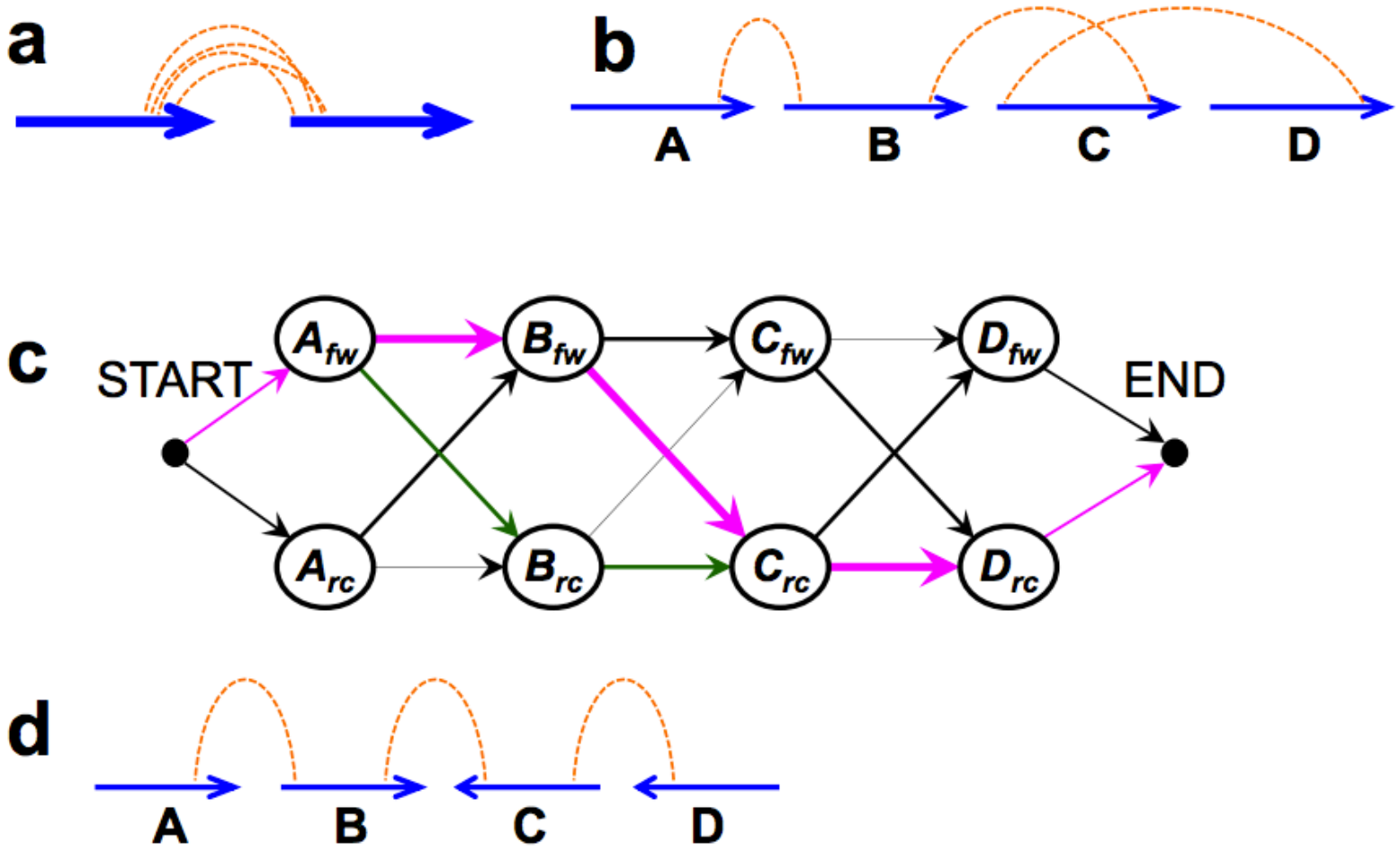

a
Cluster into chromosome groups
b
Order contigs within groups
c
Assign contig orientations
d

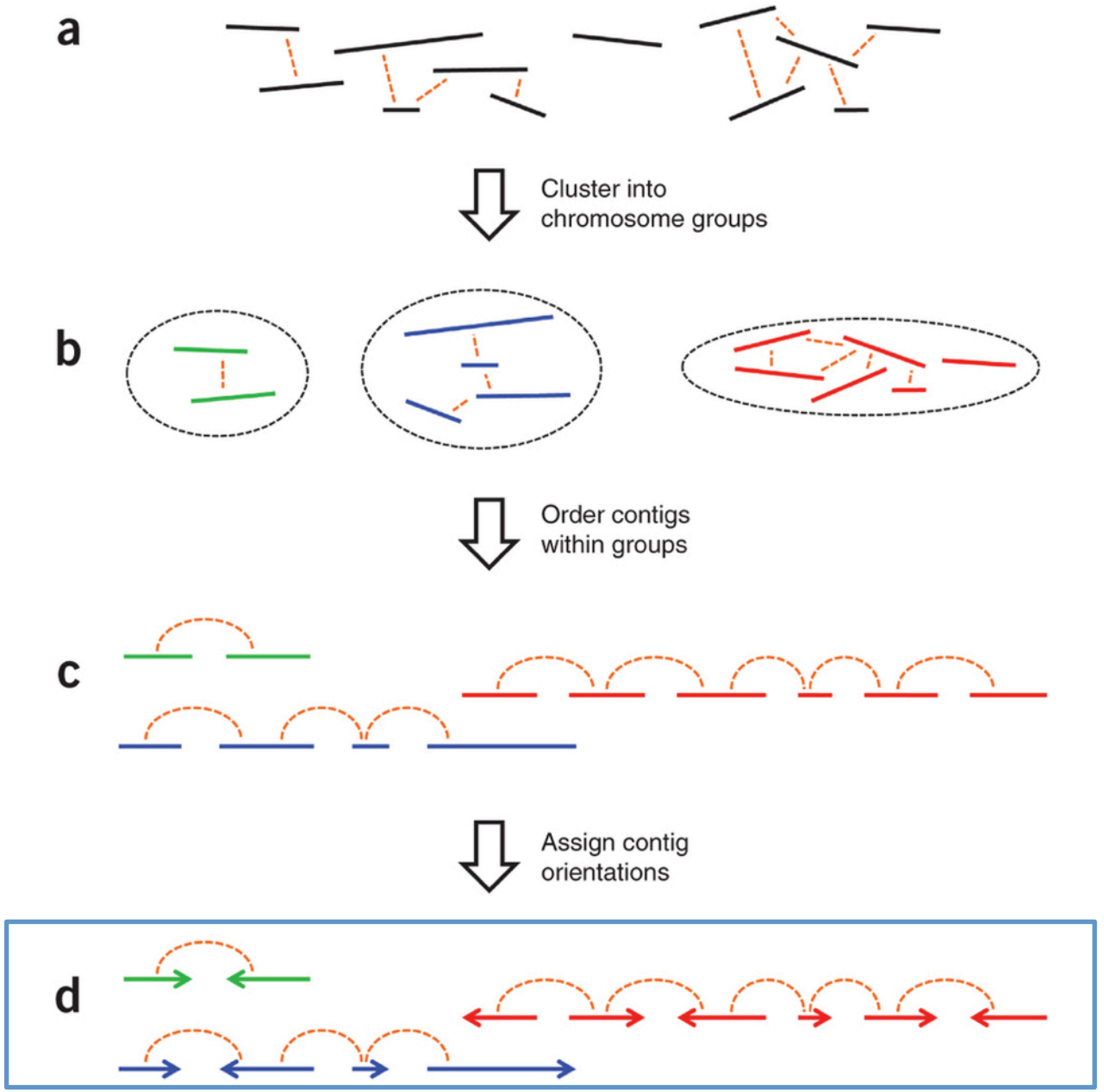
Ordering toy example




a
Cluster into chromosome groups
b
Order contigs within groups
c
Assign contig orientations
d

Orienting toy example




a
Cluster into chromosome groups
b
Order contigs within groups
c
Assign contig orientations
d

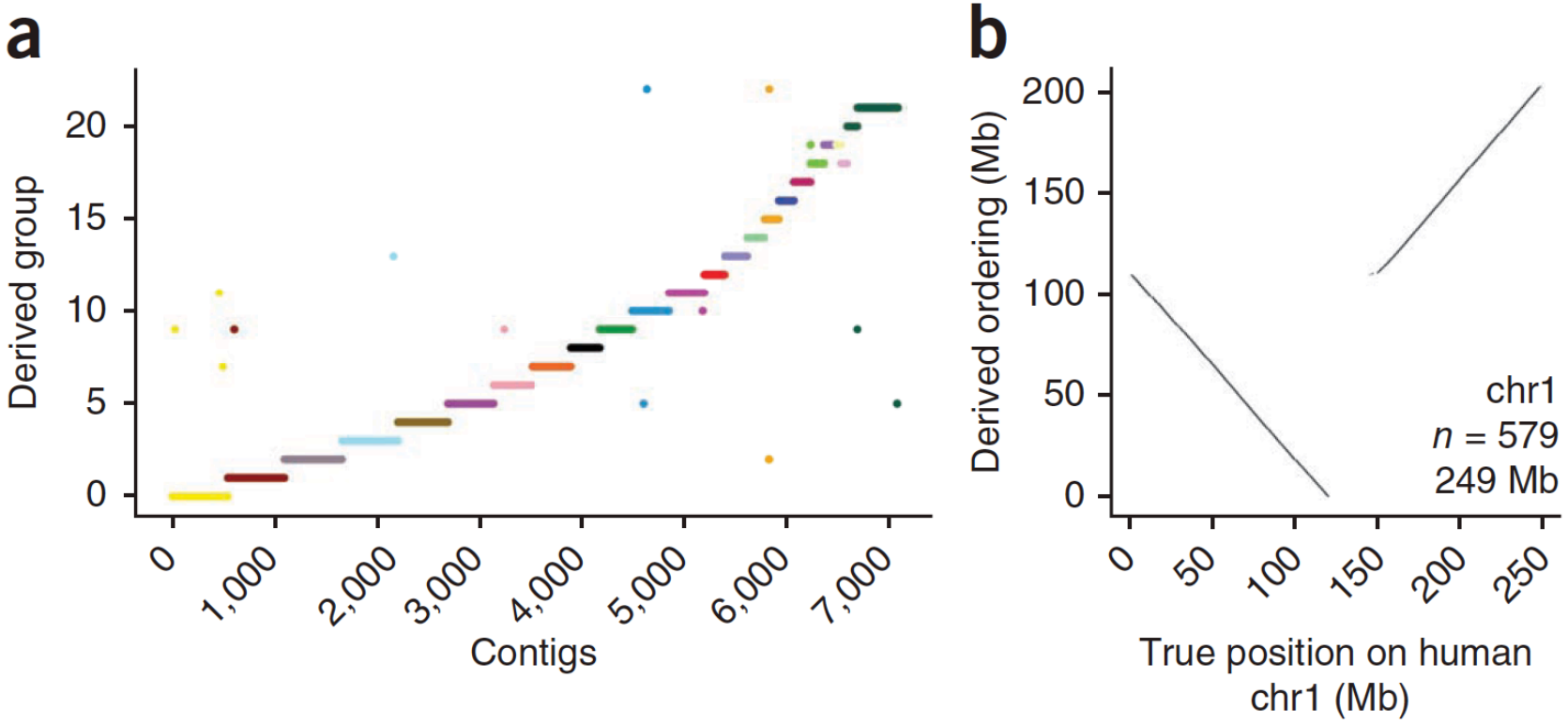
Chromosome-scale assembly of mammalian genomes

Table 1 Metrics for LACHESIS-based scaffolding of shotgun assemblies

Metric	*De novo* assemblies		
	Human	Mouse	*Drosophila*
Shotgun assembly metrics			
Total assembly length, including gaps (Mb)	2,739	2,370	127
Number of contigs or scaffolds	18,921	25,964	7,109
N50 contig or ungapped scaffold size (Kb)	437	224	68
Clustering			
% sequence (% contigs) clustered into groups	98.2 (71.5)	98.0 (87.8)	81.2 (64.3)
% clustered sequence (% contigs) mis-clustered	0.14 (1.4)	0.24 (0.5)	3.4 (10.5)
Ordering			
% clustered sequence (% contigs) ordered	94.4 (55.3)	86.7 (42.7)	82.0 (24.5)
% ordered sequence (% contigs) w/ordering errors	0.5 (0.8)	0.5 (1.1)	4.6 (5.2)
% ordered sequence (% contigs) w/orientation errors	1.2 (2.5)	1.9 (4.6)	4.1 (6.1)
High-quality predictions			
% ordered sequence (% contigs) w/high quality	92.8 (79.0)	93.3 (82.9)	94.1 (88.1)
% high-quality sequence (% contigs) w/ordering errors	0.3 (0.4)	0.3 (0.7)	3.3 (3.4)
% high-quality sequence (% contigs) w/orientation errors	0.4 (0.5)	0.5 (1.0)	2.5 (2.7)

The human and mouse shotgun assemblies are based on read-pairs from short-insert and ~2.5 Kb jumping libraries, whereas the *Drosophila* shotgun assembly is based solely on read-pairs from short-insert libraries[6].
The human and mouse shotgun assemblies consist of scaffolds, whereas the *Drosophila* shotgun assembly consists of contigs. LACHESIS places scaffolds or contigs into groups and then orders and orients them within each group.
An ordering error means that a contig or scaffold's position is out of the expected order with respect to its neighbors.
An orientation error means that its orientation is not the orientation implied by its position with respect to its immediate predecessor. 'High-quality predictions' refers to a subset of contigs or scaffolds whose position and orientation in their ordering is deemed more certain; the threshold for high quality is chosen for convenience for each assembly.

Clustering and ordering mammalian sequences with LACHESIS – human genome



LACHESIS ordering of scaffolds in a *de novo* human assembly

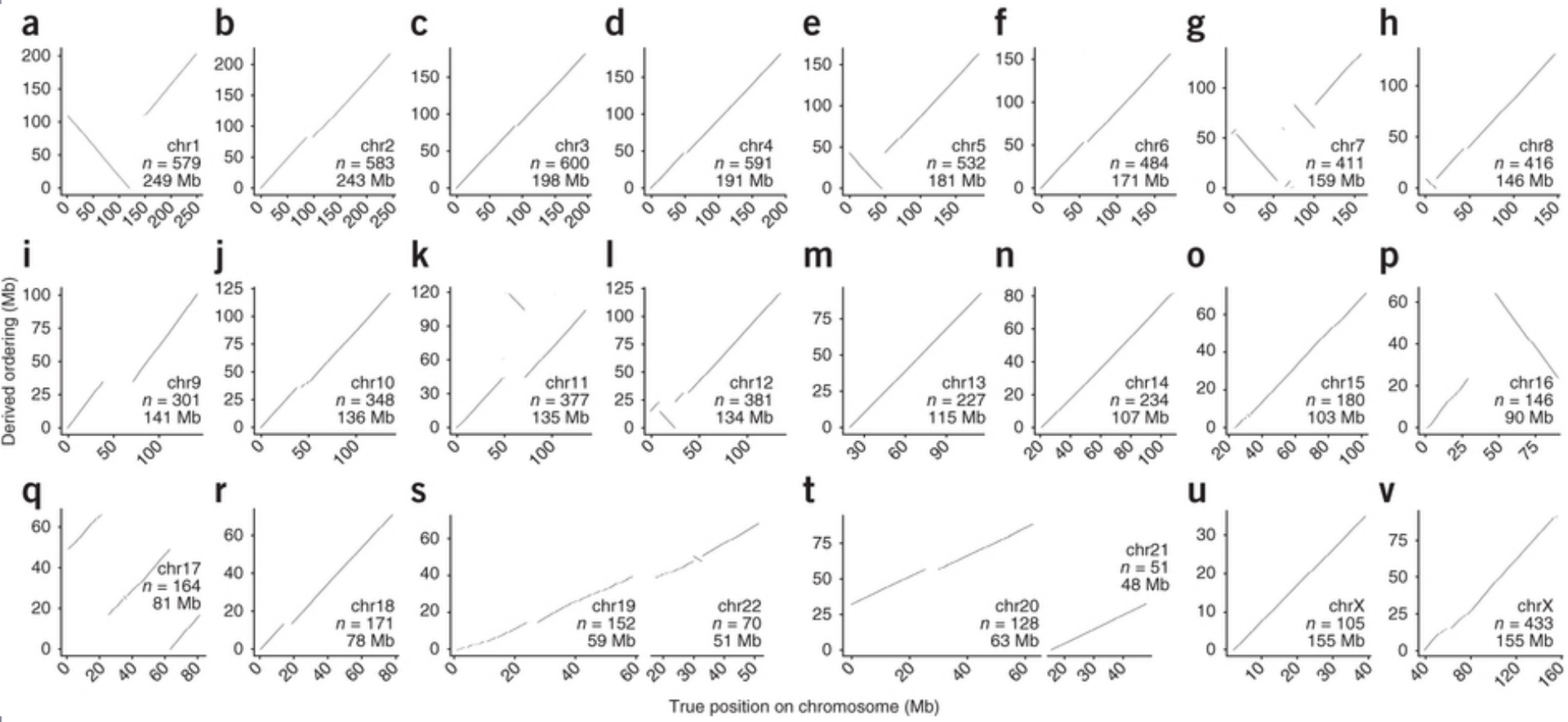
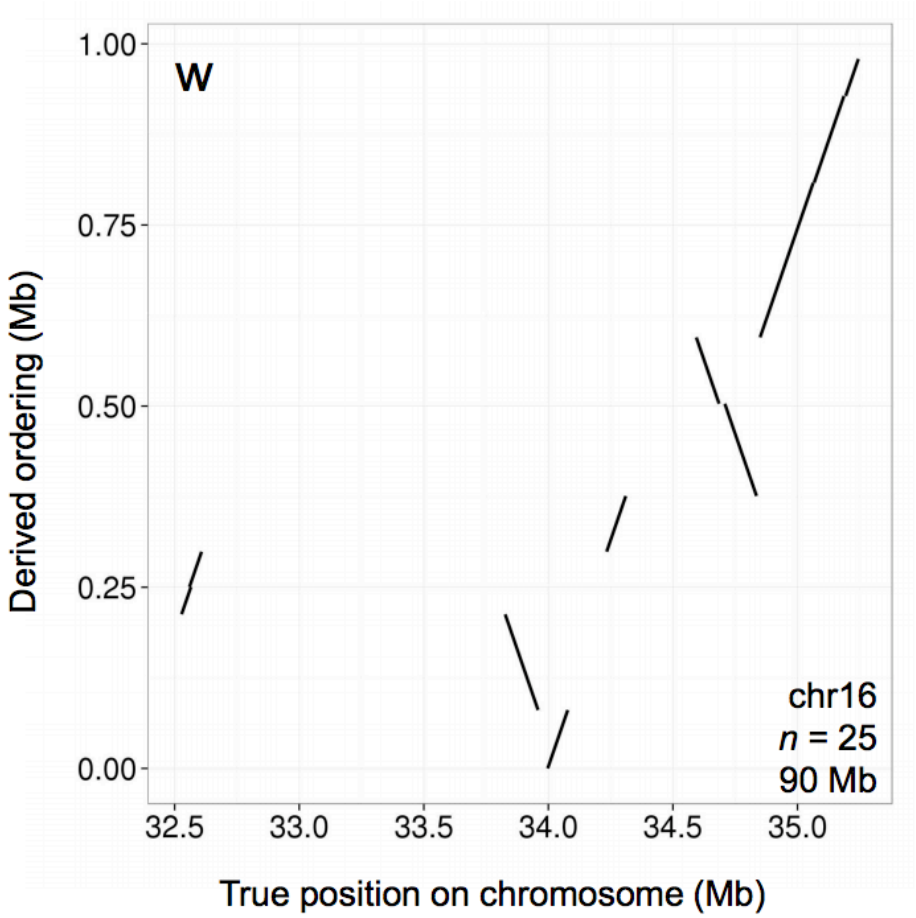


Figure	Dominant chrom(s)	Sequence length in grouped scaffolds				Sequence length in ordered scaffolds			
		Total (Mb)	Percent aligning to…			Total (Mb)	Percent aligning to…		
			Dominant chrom(s)	Other chroms	None		Dominant chrom(s)	Other chroms	None
3a	chr1	210.9	99.9%	0.01%	0.07%	202.6	100%	-	-
3b	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).

One chimeric group contains very little sequences



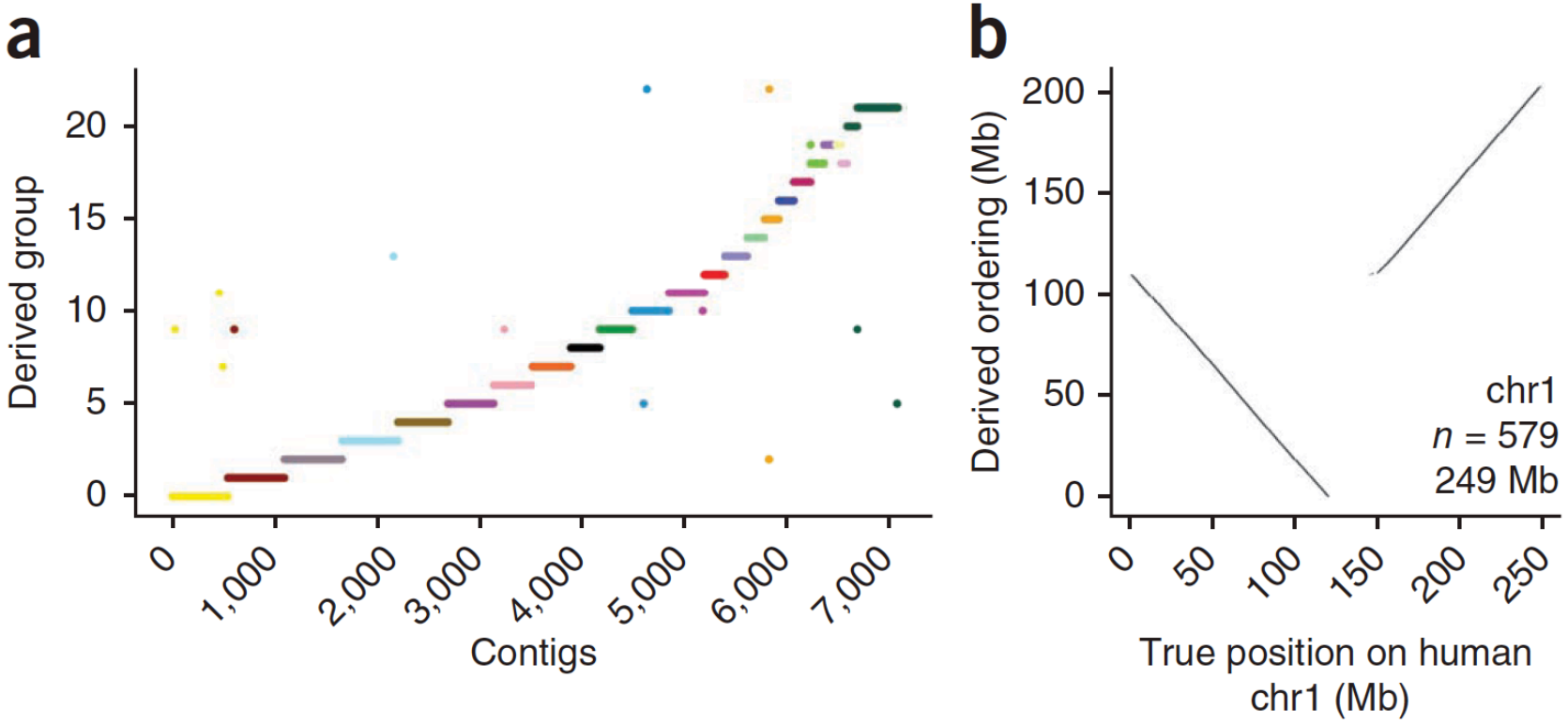
Chromosome-scale assembly of mammalian genomes

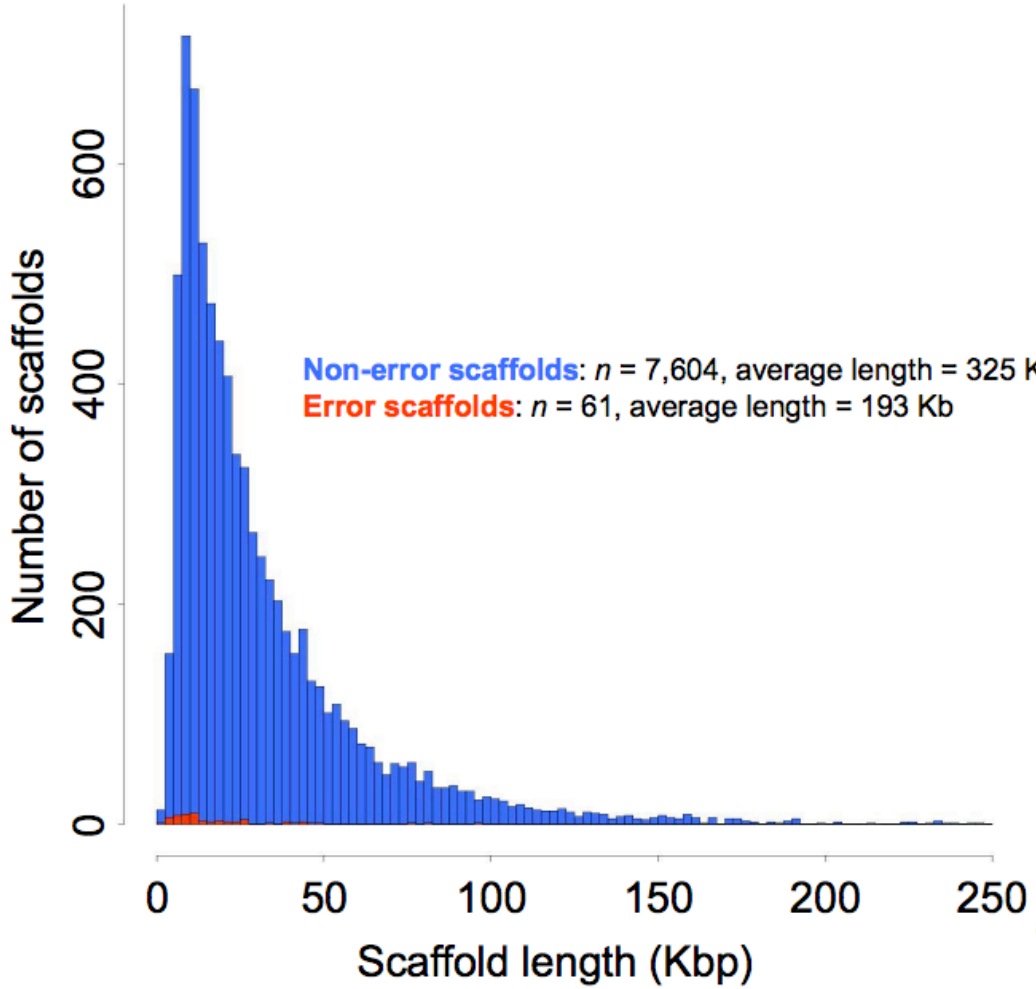
Table 1 Metrics for LACHESIS-based scaffolding of shotgun assemblies

Metric	*De novo* assemblies		
	Human	Mouse	*Drosophila*
Shotgun assembly metrics			
Total assembly length, including gaps (Mb)	2,739	2,370	127
Number of contigs or scaffolds	18,921	25,964	7,109
N50 contig or ungapped scaffold size (Kb)	437	224	68
Clustering			
% sequence (% contigs) clustered into groups	98.2 (71.5)	98.0 (87.8)	81.2 (64.3)
% clustered sequence (% contigs) mis-clustered	0.14 (1.4)	0.24 (0.5)	3.4 (10.5)
Ordering			
% clustered sequence (% contigs) ordered	94.4 (55.3)	86.7 (42.7)	82.0 (24.5)
% ordered sequence (% contigs) w/ordering errors	0.5 (0.8)	0.5 (1.1)	4.6 (5.2)
% ordered sequence (% contigs) w/orientation errors	1.2 (2.5)	1.9 (4.6)	4.1 (6.1)
High-quality predictions			
% ordered sequence (% contigs) w/high quality	92.8 (79.0)	93.3 (82.9)	94.1 (88.1)
% high-quality sequence (% contigs) w/ordering errors	0.3 (0.4)	0.3 (0.7)	3.3 (3.4)
% high-quality sequence (% contigs) w/orientation errors	0.4 (0.5)	0.5 (1.0)	2.5 (2.7)

The human and mouse shotgun assemblies are based on read-pairs from short-insert and ~2.5 Kb jumping libraries, whereas the *Drosophila* shotgun assembly is based solely on read-pairs from short-insert libraries[6].
The human and mouse shotgun assemblies consist of scaffolds, whereas the *Drosophila* shotgun assembly consists of contigs. LACHESIS places scaffolds or contigs into groups and then orders and orients them within each group.
An ordering error means that a contig or scaffold's position is out of the expected order with respect to its neighbors.
An orientation error means that its orientation is not the orientation implied by its position with respect to its immediate predecessor. 'High-quality predictions' refers to a subset of contigs or scaffolds whose position and orientation in their ordering is deemed more certain; the threshold for high quality is chosen for convenience for each assembly.

Clustering and ordering mammalian sequences with LACHESIS – human genome chromosome 1

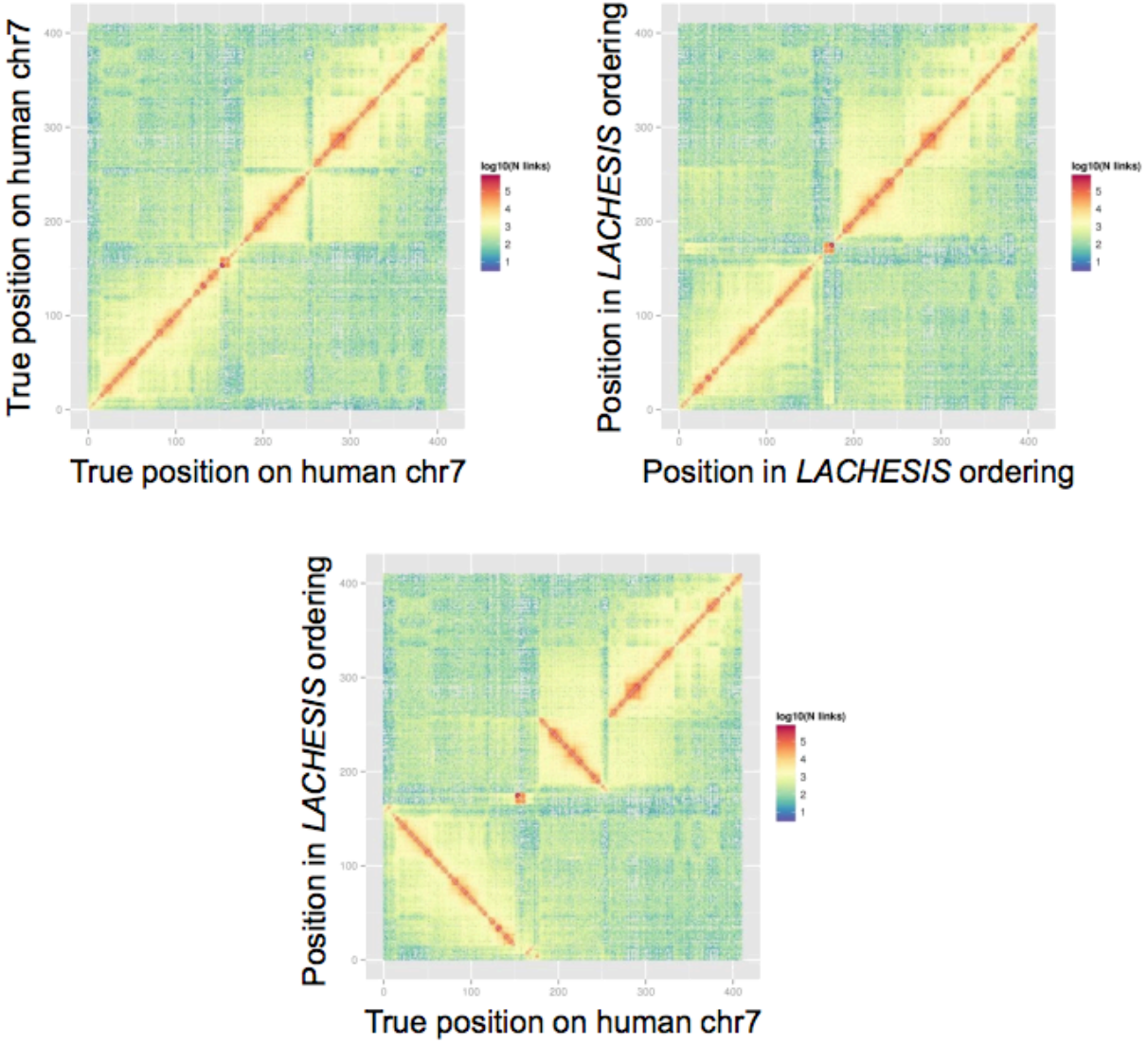




Repeat type	UCSC Genome Browser track name	Enrichment near the edges of mis-ordered scaffolds
Segmental duplications (>1 Kb length, >90% similarity)	Segmental Dups	**6.38**
Microsatellite repeats (dinucleotide, trinucleotide)	Microsatellite	**1.24**
Simple tandem repeats (4 or more nucleotides)	Simple 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

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.



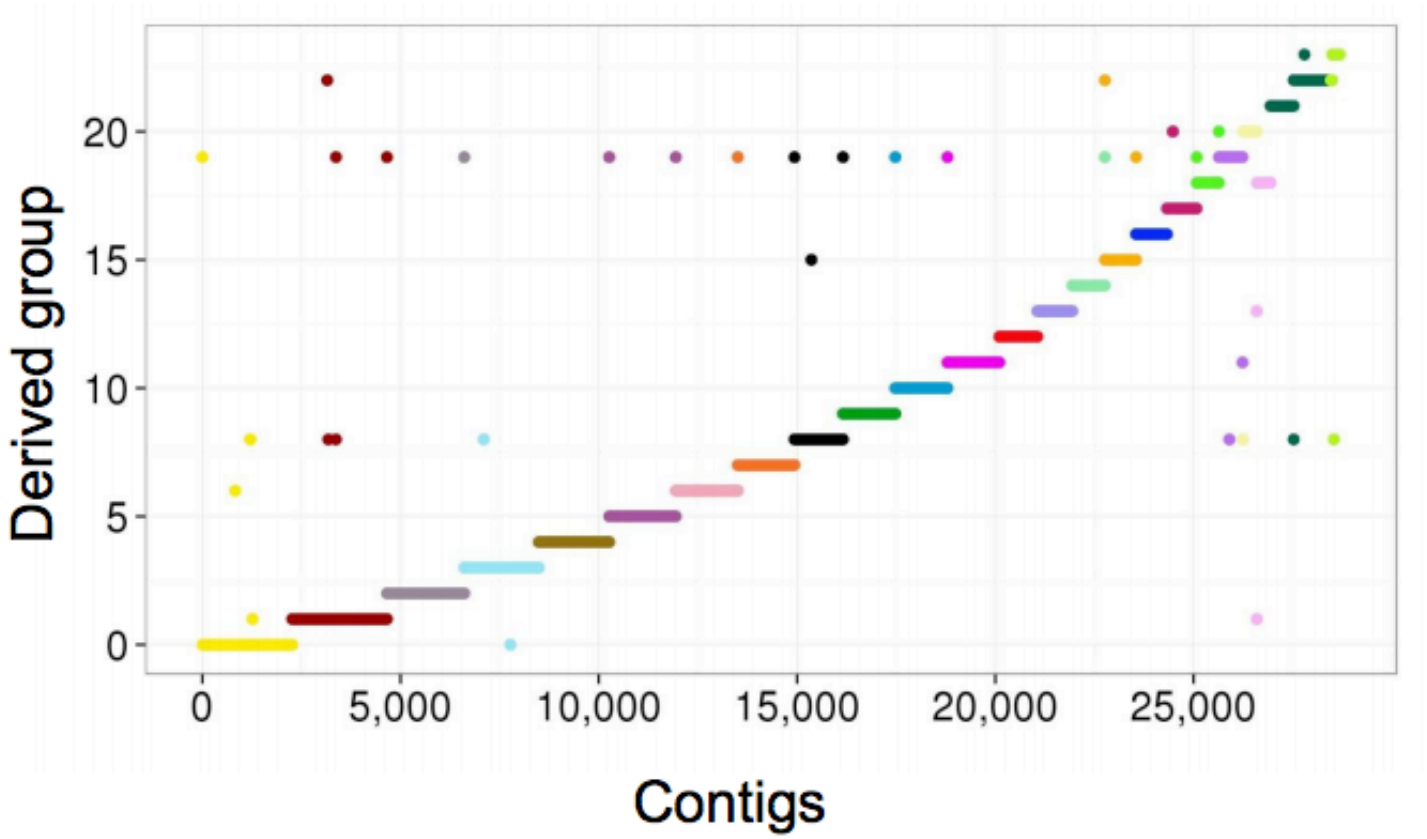
Supplementary Figure 6 | Example of *LACHESIS* assembly errors due to long-range chromatin interactions. Shown are three heatmaps of the density of Hi-C links between scaffolds of the *de novo* human

Robustness to contig size and Hi-C data quality

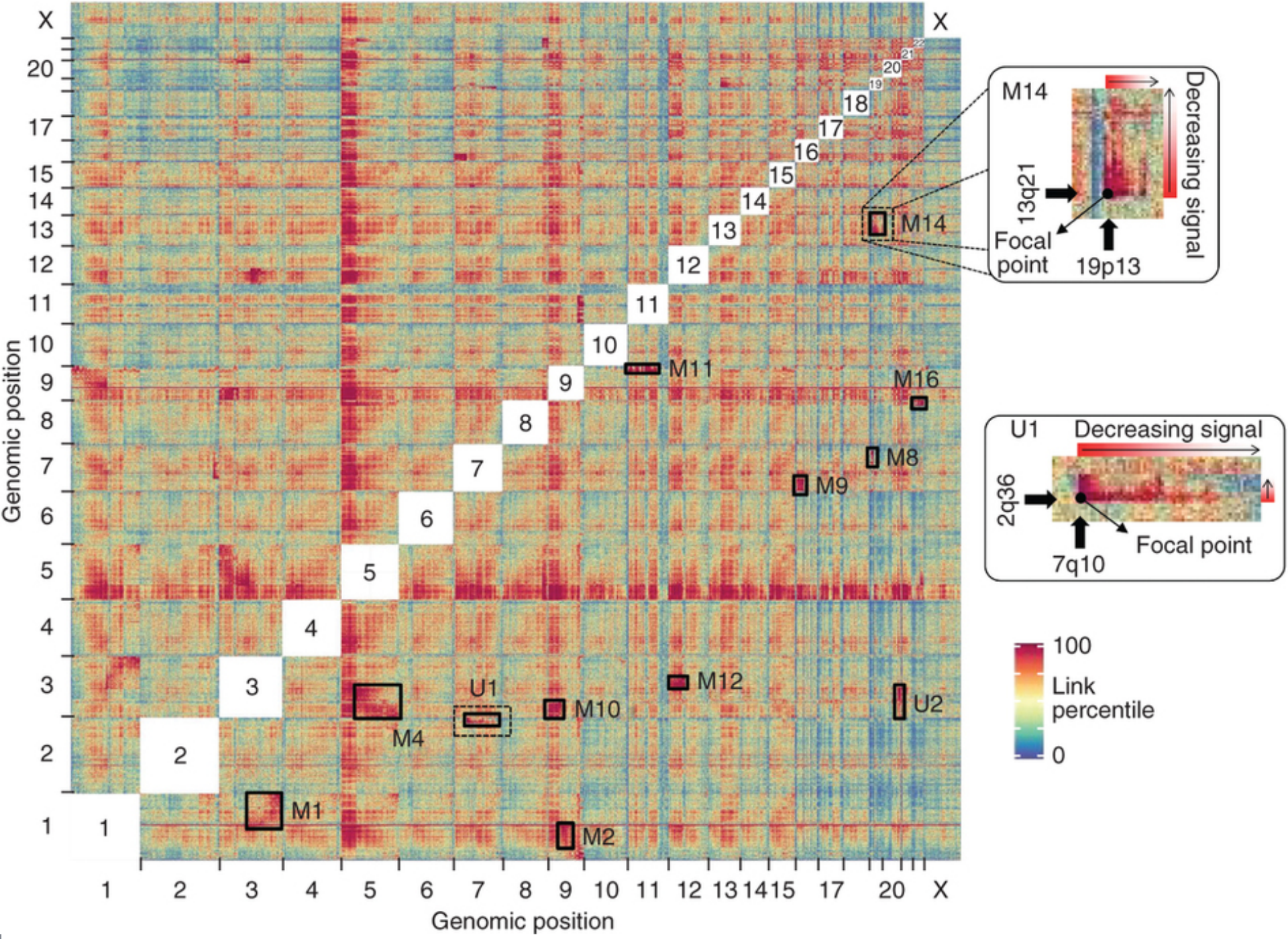
Table 2: Metrics for LACHESIS-based scaffolding of simulated assemblies

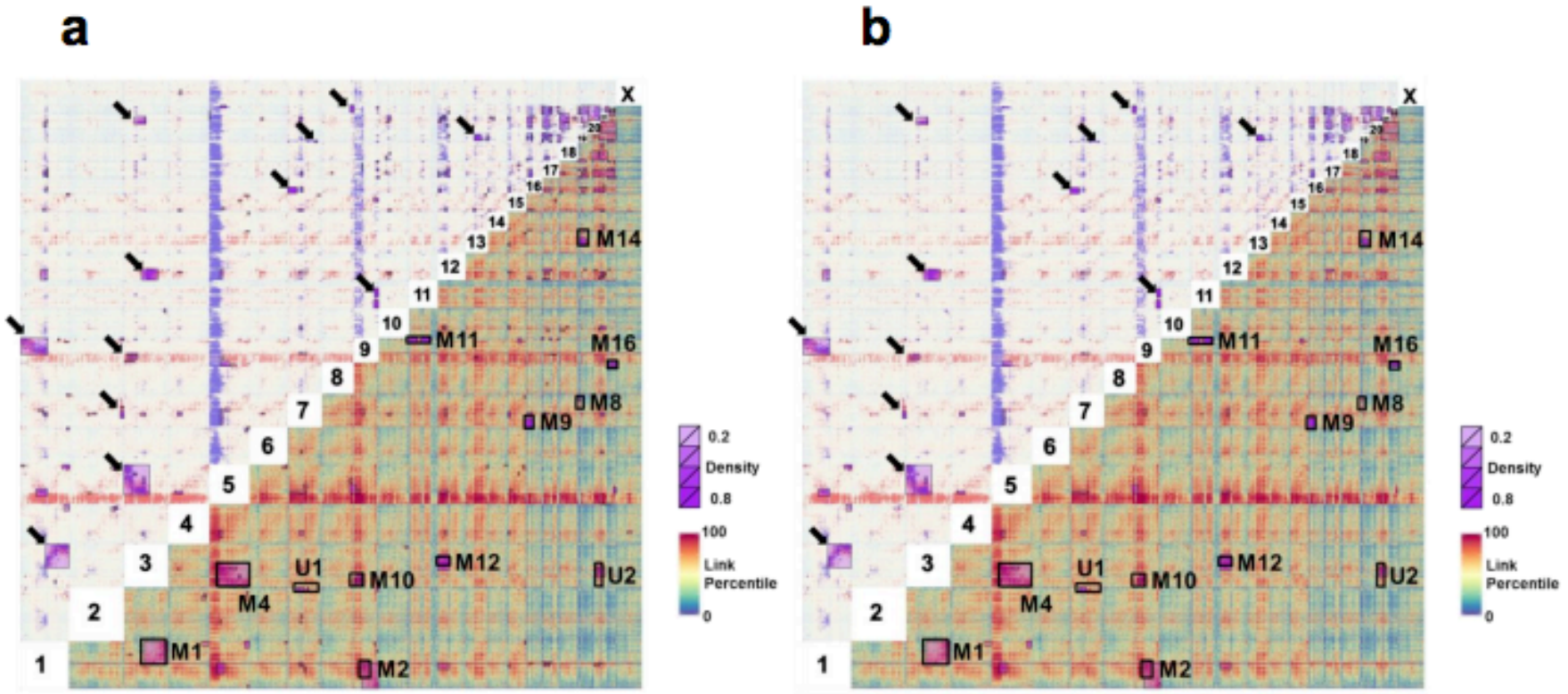
Metric	Simulated contig size						
	10 Kb	20 Kb	50 Kb	100 Kb	200 Kb	500 Kb	1 Mb
Number of contigs	309,579	154,794	61,927	30,970	15,489	6,206	3,113
% sequence clustered into groups	30.1	74.2	91.9	92.7	92.9	93.1	93.4
% clustered sequence mis-clustered	1.6	0.47	0.41	0.46	0.66	0.66	0.26
% clustered sequence ordered	48.5	79.9	98.9	99.8	99.97	99.93	99.98
% ordered sequence w/ordering errors	37.2	18.0	4.4	2.2	1.4	0.8	0.8
% ordered sequence w/orientation errors	44.8	28.7	7.7	2.6	1.2	0.8	0.7

Simulated assemblies were created by breaking up the human reference genome into simulated contigs of varying sizes, and then using LACHESIS to cluster, order and orient the simulated contigs. The simulated contigs' expected order and orientation are derived from their true position in the reference genome. Ordering and orientation errors are defined as in Table 1.

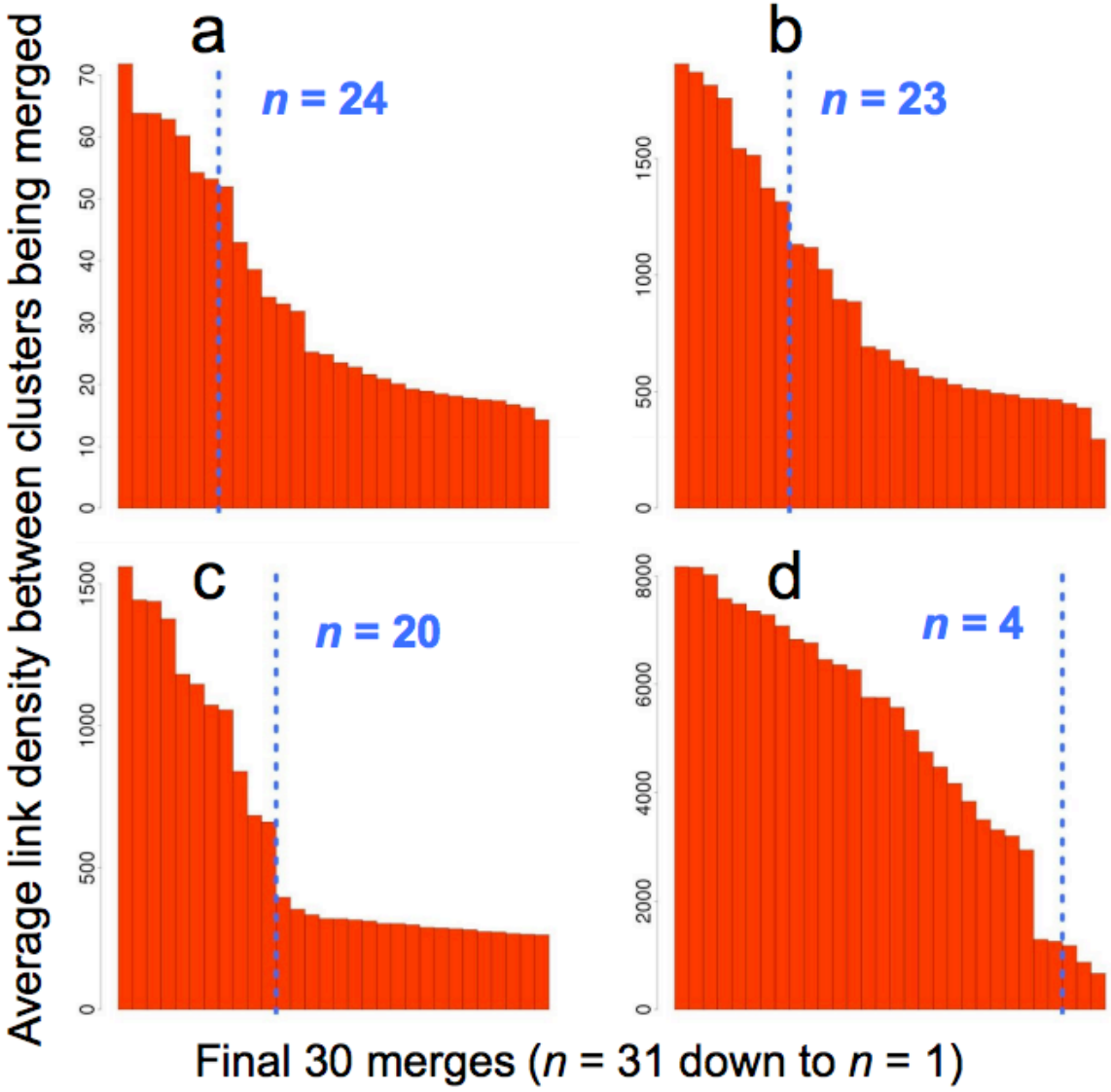


Supplementary Figure 10 | *LACHESIS* clustering results on simulated 100 Kb contigs of the human reference genome. The human genome was split into simulated 100 Kb contigs and *LACHESIS* was used to


Genomic position
Genomic position
M14
Decreasing signal
13q21
Focal point
19p13
U1
Decreasing signal
2q36
Focal point
7q10
100
Link percentile
0
M1
M2
M4
U1
M10
M11
M12
M16
M8
M9
U2



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



Supplementary Figure 13 | Difficulty of calling the number of chromosomes from Hi-C link data alone.

Possible solutions:

- Ordering before deciding chromosome groups
- Statistical methods for optical numbers of clusters prediction

Summary

Pros:

- Assign, order and orient genomic sequences to chromosomes, including across megabase-scale centromere gaps
- Validating chromosomal translocations in cancer genomes

Cons:

- Required a large amount of material (10^6-10^8 cells)
- The contiguity required for the initial *de novo* assembly (>50kb)
- Clustering step requires # of chromosomal groups

Future work:

- Reducing input requirements
- Further validate LACHSIS in diverse species
- Multiple restriction enzymes (small contigs)
- Integrate LACHESIS and the initial assembly (ALLPATHS-LG)