DNA Sequencing

Definition of Coverage

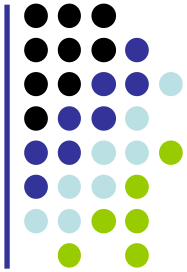
Length of genomic segment: **G**
Number of reads: **N**
Length of each read: **L**

Definition: Coverage **C = N L / G**

How much coverage is enough?

Lander-Waterman model: **Prob[not covered bp] = e^{-C}**
Assuming uniform distribution of reads, C=10 results in 1 gapped region /1,000,000 nucleotides

Repeats



Bacterial genomes: 5%
Mammals: 50%

Repeat types:

- **Low-Complexity DNA** (e.g. ATATATATACATA…)
- **Microsatellite repeats** $(a_1 \ldots a_k)^N$ where k ~ 3-6
  (e.g. CAGCAGTAGCAGCACCAG)
- **Transposons**
  - **SINE** (Short Interspersed Nuclear Elements)
    e.g., ALU: ~300-long, 10^6 copies
  - **LINE** (Long Interspersed Nuclear Elements)
    ~4000-long, 200,000 copies
  - **LTR retroposons** (Long Terminal Repeats (~700 bp) at each end)
    cousins of HIV
- **Gene Families** genes duplicate & then diverge (paralogs)
- **Recent duplications** ~100,000-long, very similar copies

Sequencing and Fragment Assembly

```
AGTAGCACAGAC
TACGACGAGACG
ATCGTGCGAGCG
ACGGCGTAGTGT
GCTGTACTGTCG
TGTGTGTGTACT
CTCCT
```

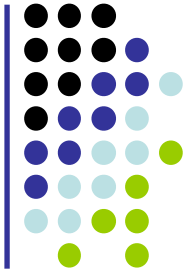
$3x10^9$ nucleotides

50% of human DNA is compose

Error!

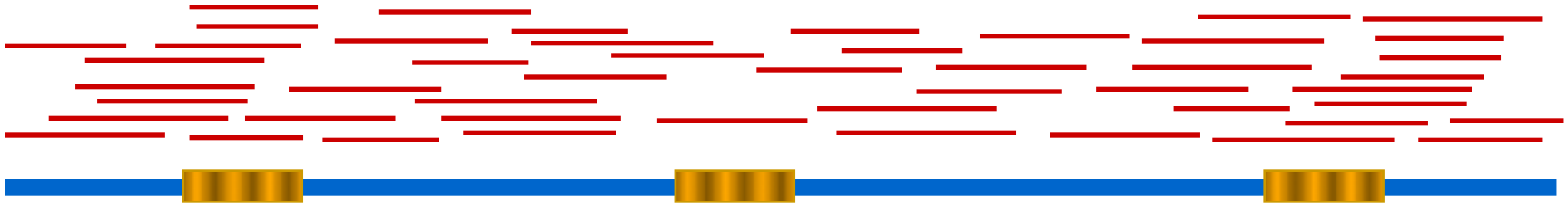
Glued together two distant regions

What can we do about repeats?



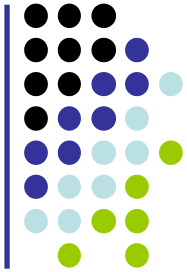
Two main approaches:

- Cluster the reads



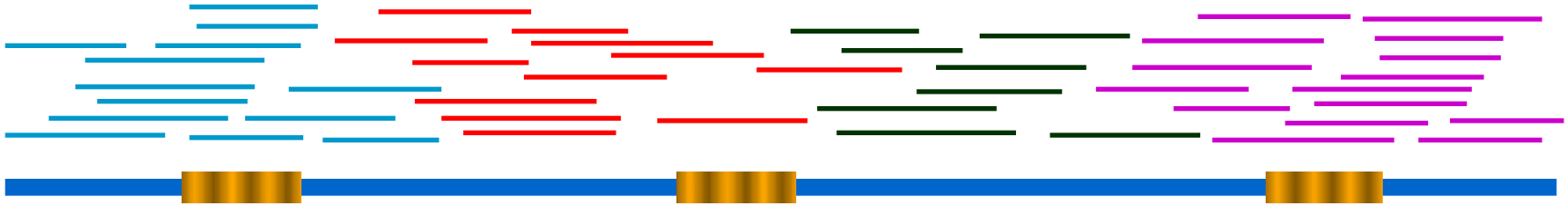
- Link the reads

What can we do about repeats?



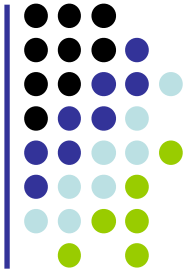
Two main approaches:

- Cluster the reads



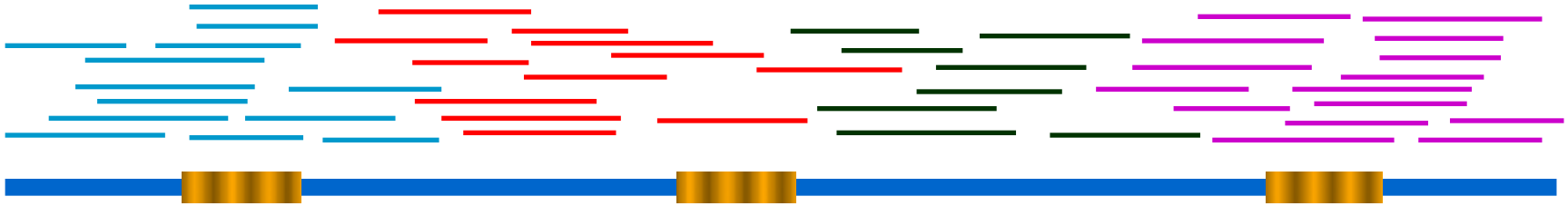
- Link the reads

What can we do about repeats?

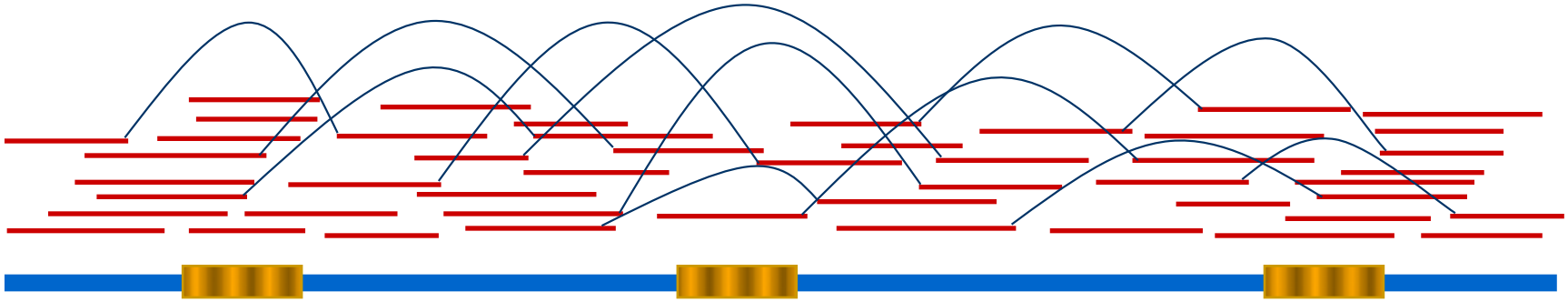


Two main approaches:

- Cluster the reads

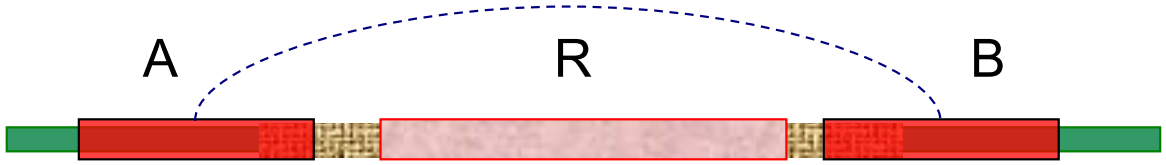


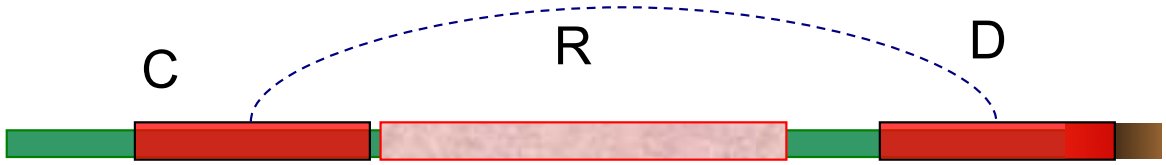
- Link the reads



Sequencing and Fragment Assembly

$3x10^9$ nucleotides



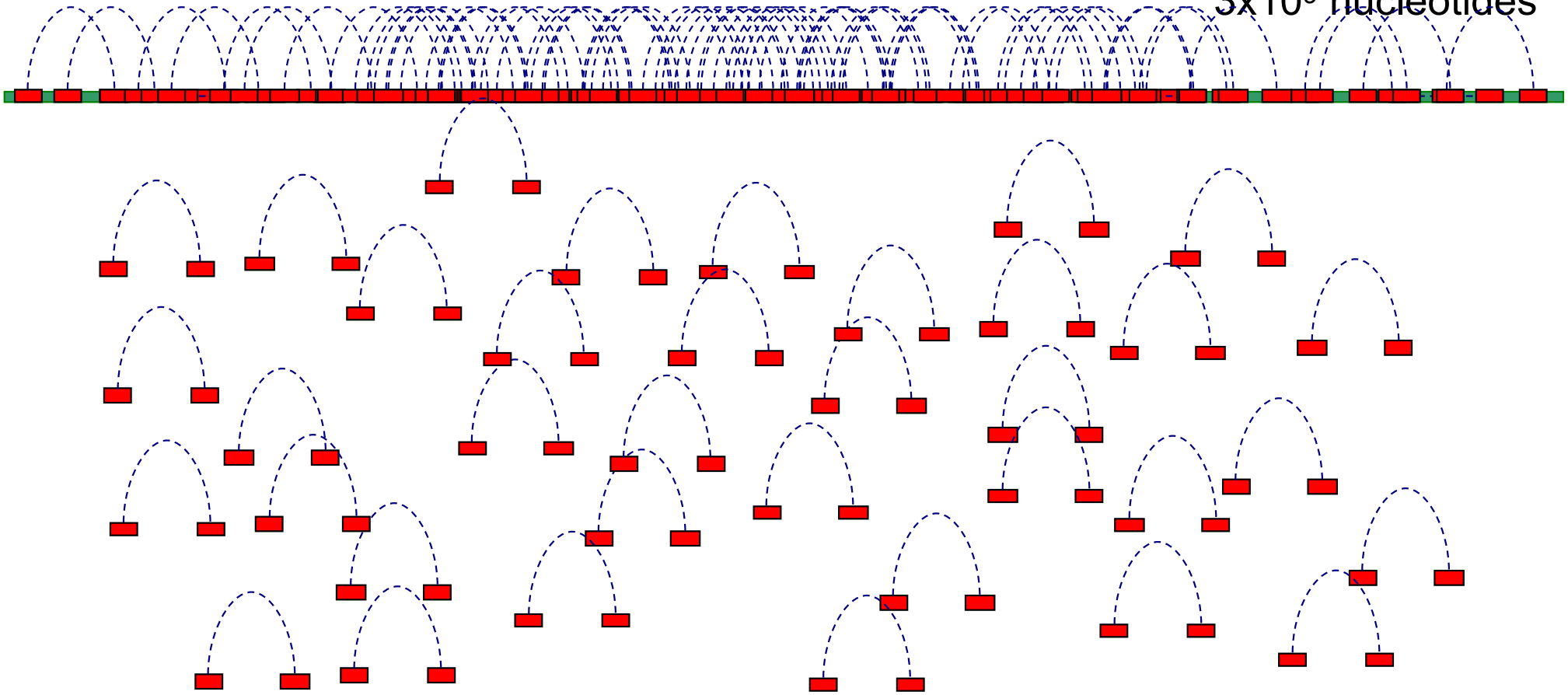


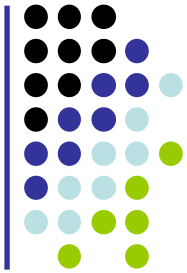
ARB, CRD

or

ARD, CRB ?

Sequencing and Fragment Assembly

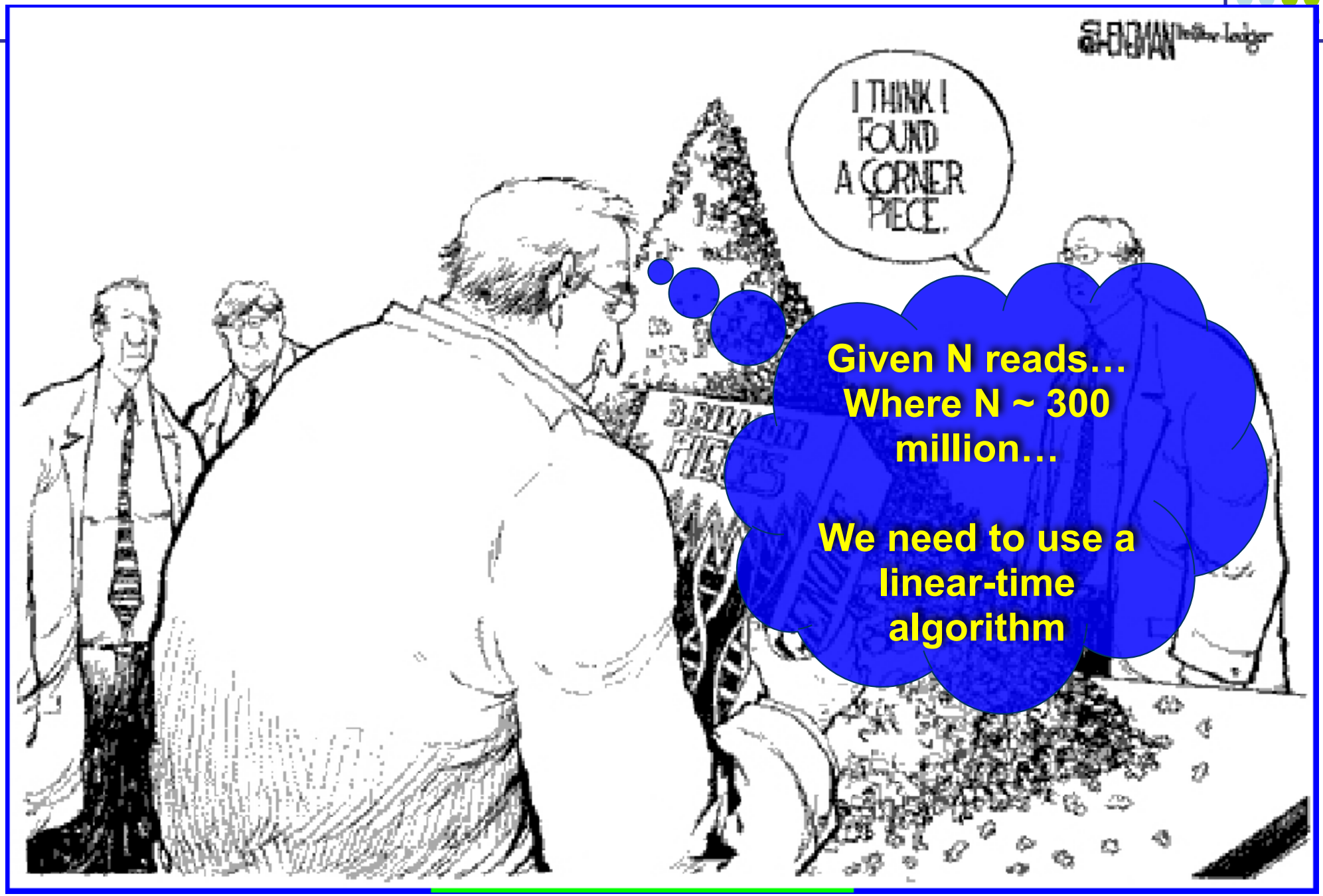




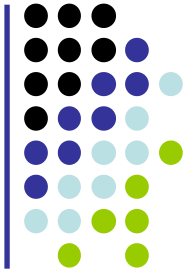
Fragment Assembly

(in whole-genome shotgun sequencing)

Fragment Assembly



Steps to Assemble a Genome



Some Terminology

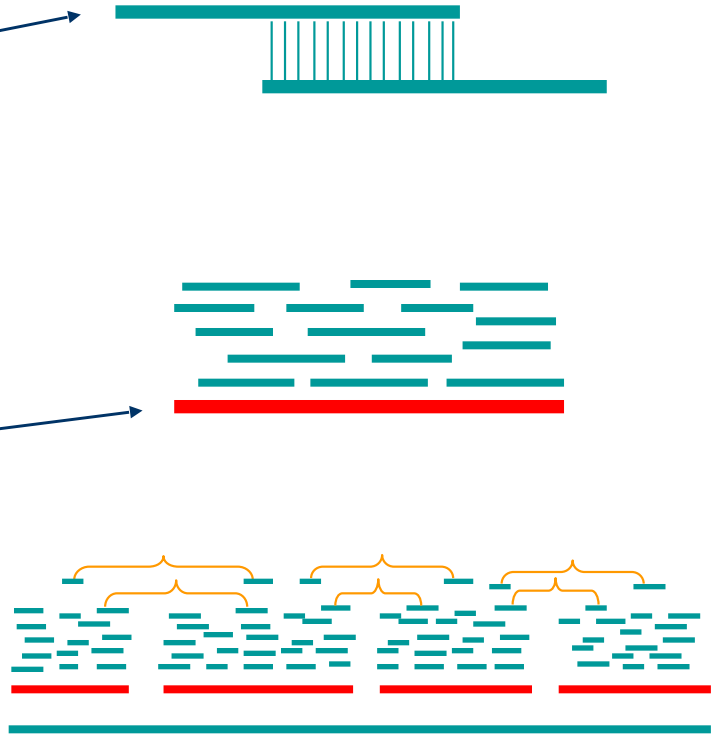
read a 500-900 long word that comes out of sequencer

mate pair a pair of reads from two ends of the same insert fragment

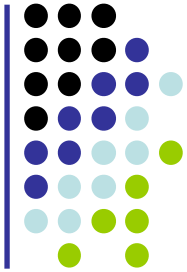
contig a contiguous sequence formed by several overlapping reads with no gaps

supercontig (scaffold) an ordered and oriented set of contigs, usually by mate pairs

consensus sequene sequence derived from the multiple alignment of reads in a contig



1. Find Overlapping Reads

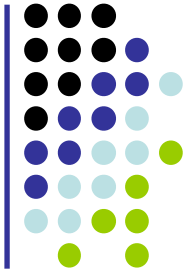


```
aaactgcagtacggatct
aaactgcag
 aactgcagt
…
          gtacggatct
           tacggatct
gggcccaaactgcagtac
gggcccaaa
 ggcccaaac
…
          actgcagta
           ctgcagtac
gtacggatctactacaca
gtacggatc
 tacggatct
…
          ctactacac
           tactacaca
```

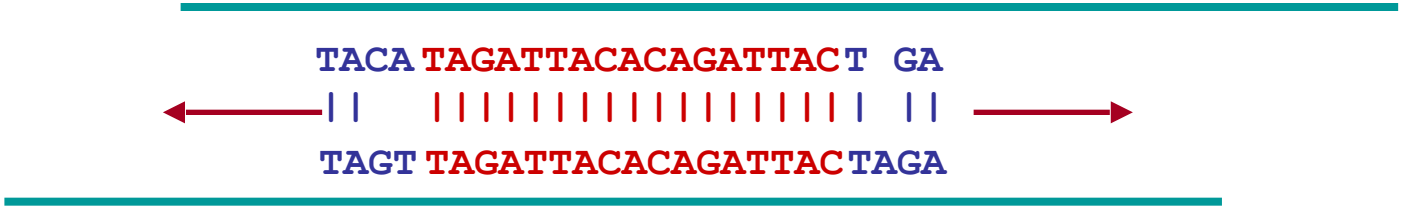
```
(read, pos., word, orient.)
aaactgcag
aactgcagt
actgcagta
…
gtacggatc
tacggatct
gggcccaaa
ggcccaaac
gcccaaact
…
actgcagta
ctgcagtac
gtacggatc
tacggatct
acggatcta
…
ctactacac
tactacaca
```

```
(word, read, orient., pos.)
aaactgcag
aactgcagt
acggatcta
actgcagta
actgcagta
cccaaactg
cggatctac
ctactacac
ctgcagtac
ctgcagtac
gcccaaact
ggcccaaac
gggcccaaa
gtacggatc
gtacggatc
tacggatct
tacggatct
tactacaca
```

1. Find Overlapping Reads

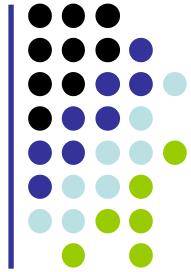


- Find pairs of reads sharing a k-mer, k ~ 24
- Extend to full alignment – throw away if not >98% similar



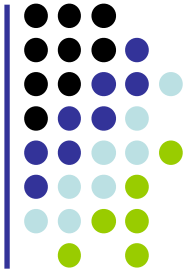
- Caveat: repeats
  - A k-mer that occurs N times, causes $O(N^2)$ read/read comparisons
  - ALU k-mers could cause up to $1{,}000{,}000^2$ comparisons
- Solution:
  - Discard all k-mers that occur “too often”
    - Set cutoff to balance sensitivity/speed tradeoff, according to genome at hand and computing resources available

1. Find Overlapping Reads



Create local multiple alignments from the overlapping reads

1. Find Overlapping Reads



- Correct errors using multiple alignment

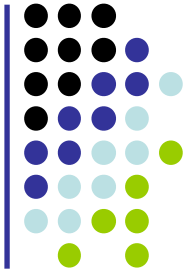
insert *A*

replace T with *C*

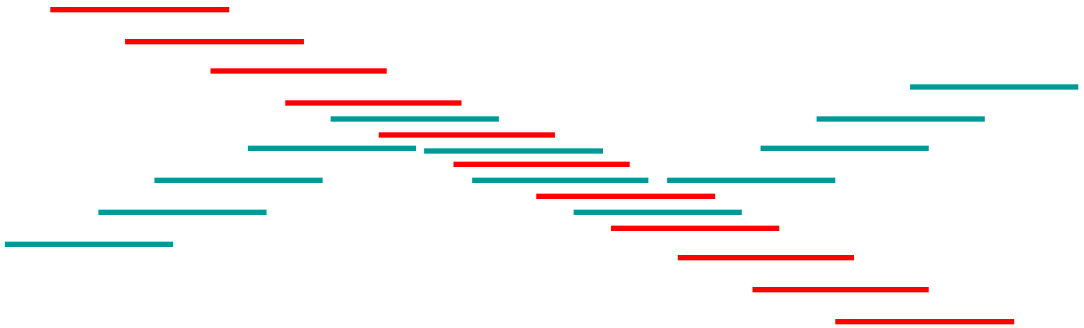
correlated errors—
probably caused by repeats
⇒ disentangle overlaps

In practice, error correction removes up to 98% of the errors

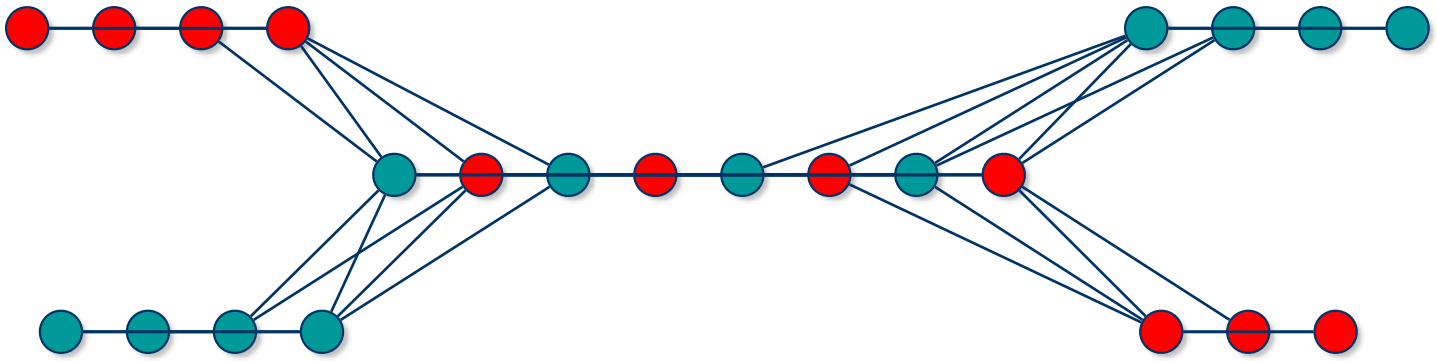
2. Merge Reads into Contigs



- Overlap graph:
  - Nodes: reads r_1…..r_n
  - Edges: overlaps (r_i, r_j, shift, orientation, score)

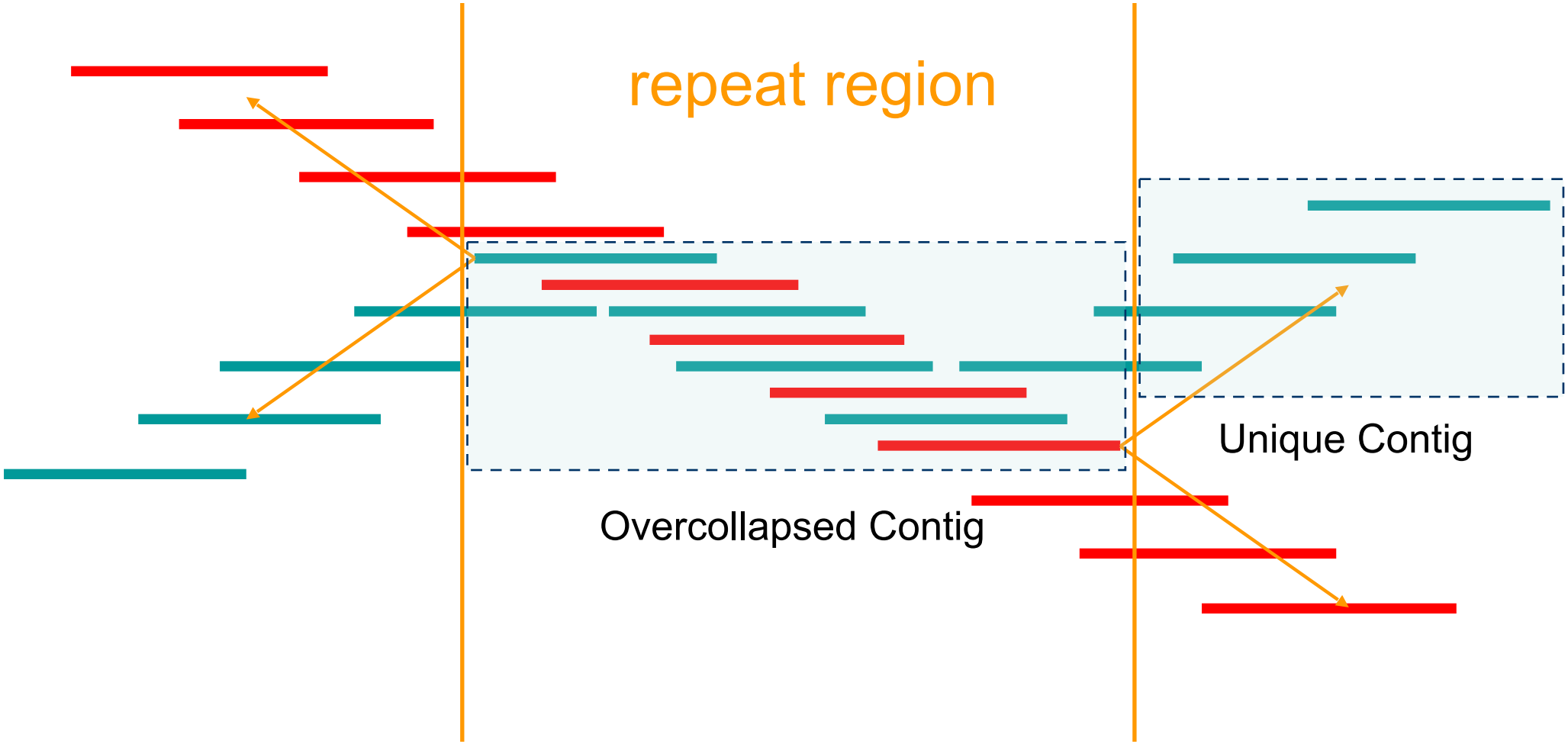


Reads that come from two regions of the genome (blue and red) that contain the same repeat



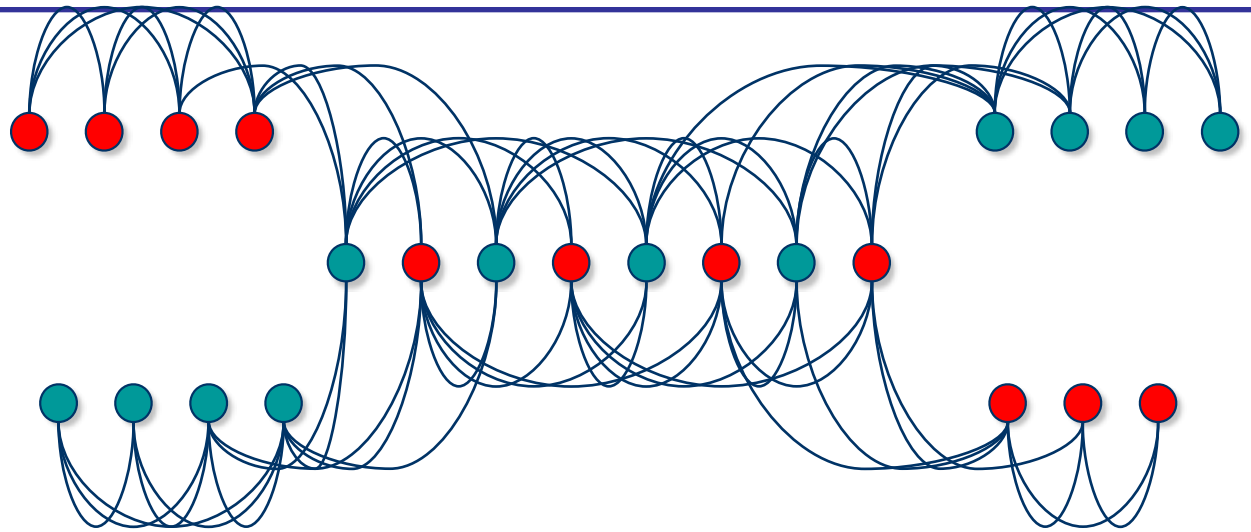
Note:
of course, we don't know the "color" of these nodes

2. Merge Reads into Contigs

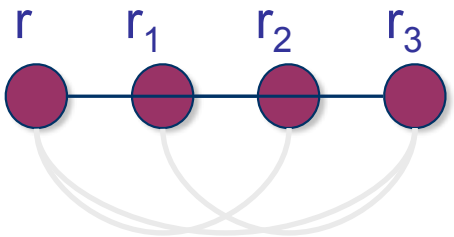


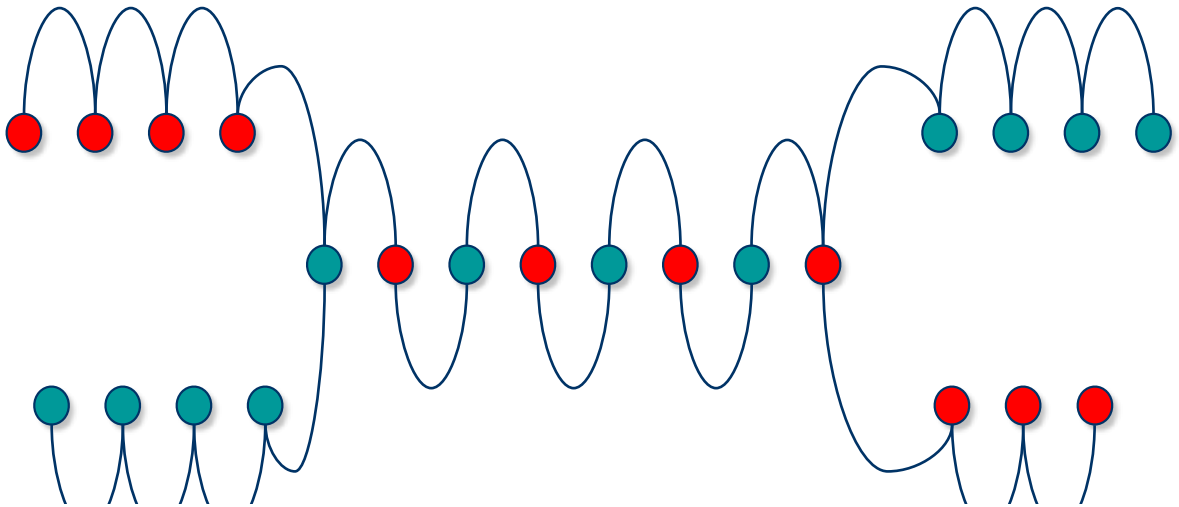
We want to merge reads up to potential repeat boundaries

2. Merge Reads into Contigs

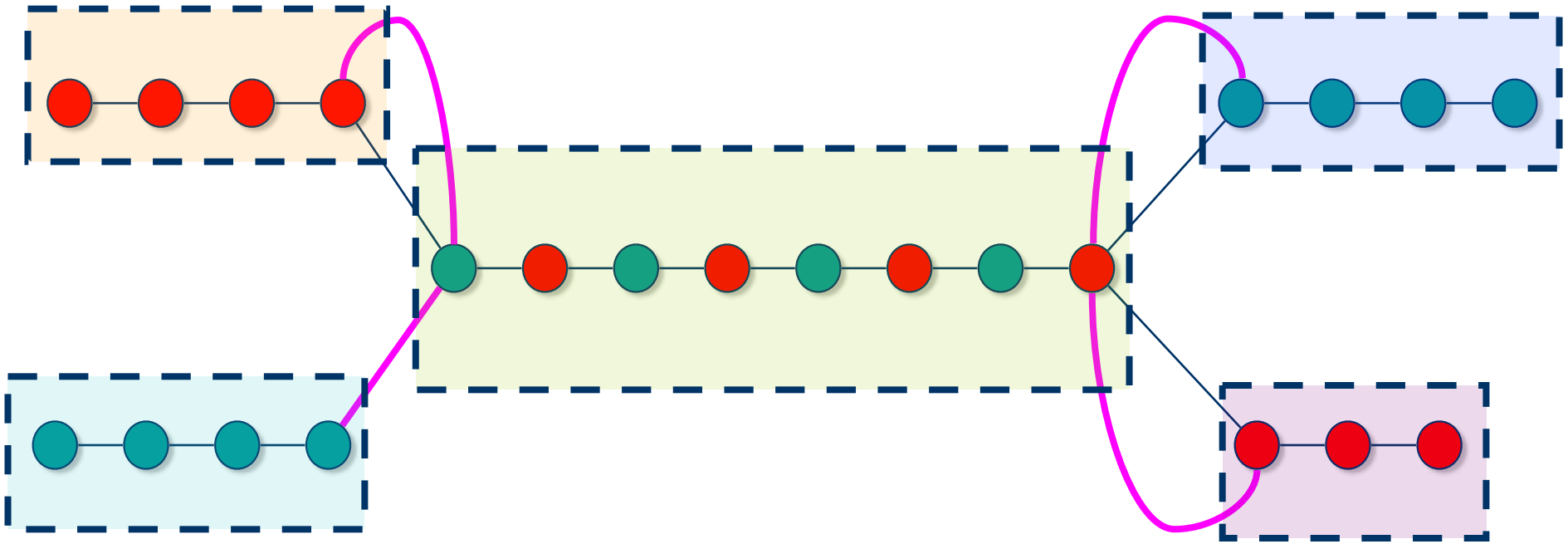


- Remove transitively inferable overlaps
  - If read r overlaps to the right reads r_1, r_2, and r_1 overlaps r_2, then (r, r_2) can be inferred by (r, r_1) and (r_1, r_2)

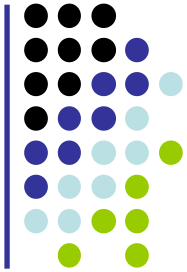




2. Merge Reads into Contigs



Repeats, errors, and contig lengths



- Repeats shorter than read length are easily resolved
  - Read that spans across a repeat disambiguates order of flanking regions
- Repeats with more base pair diffs than sequencing error rate are OK
  - We throw overlaps between two reads in different copies of the repeat
- To make the genome **appear** less repetitive, try to:
  - Increase read length
  - Decrease sequencing error rate

Role of error correction:

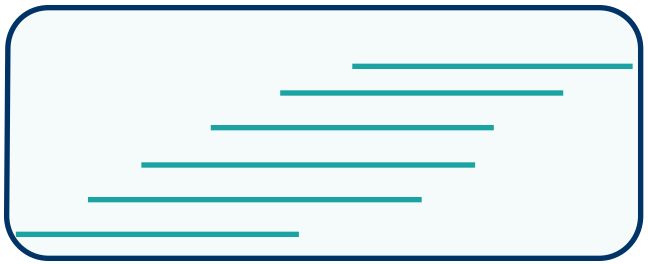
Discards up to 98% of single-letter sequencing errors

decreases error rate

⇒ decreases effective repeat content

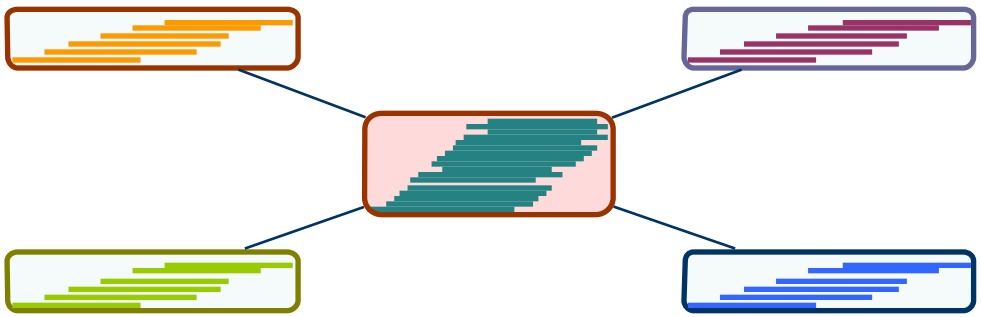
⇒ increases contig length

3. Link Contigs into Supercontigs



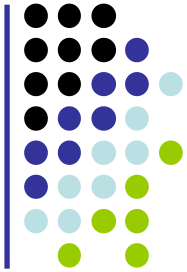
Normal density

Too dense
⇒ Overcollapsed



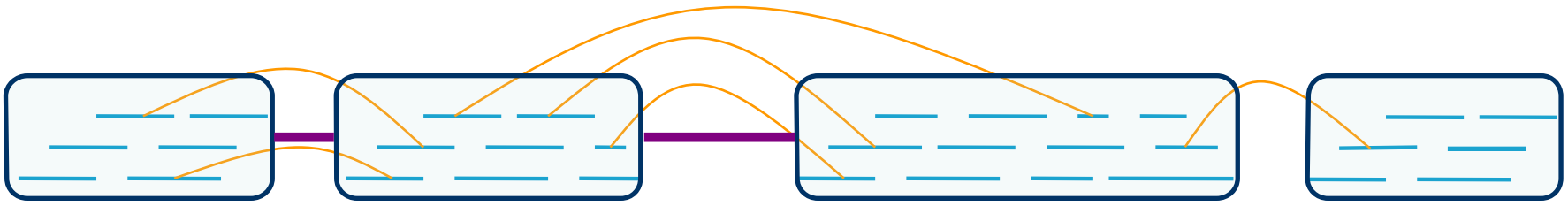
Inconsistent links
⇒ Overcollapsed?

3. Link Contigs into Supercontigs



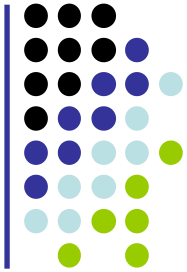
Find all links between unique contigs

Connect contigs incrementally, if ≥ 2 forward-reverse links



supercontig
(aka *scaffold*)

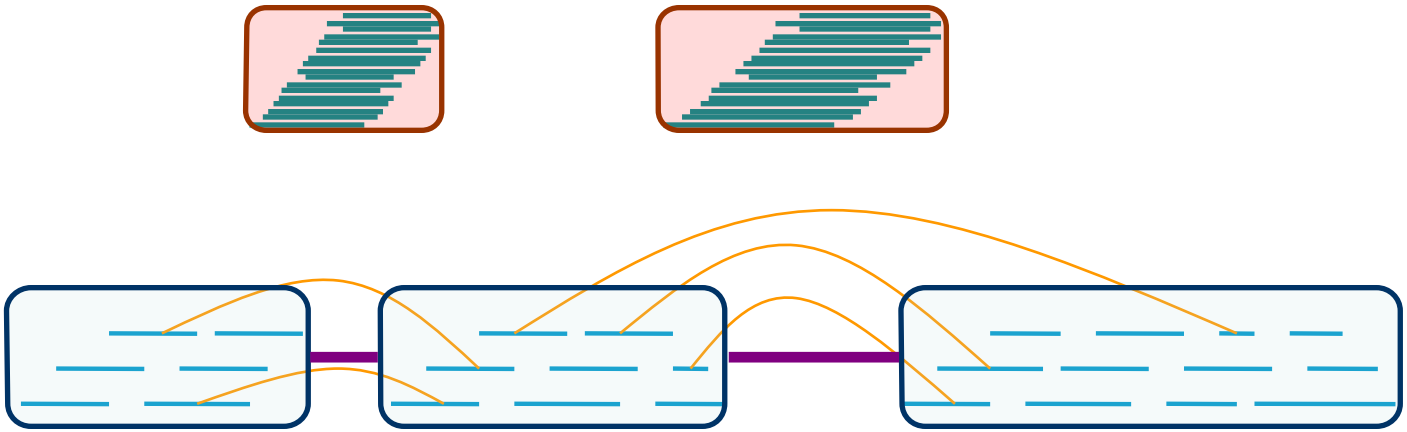
3. Link Contigs into Supercontigs



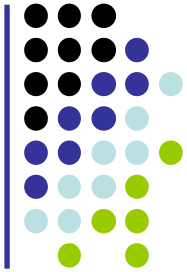
Fill gaps in supercontigs with paths of repeat contigs

Complex algorithmic step

- Exponential number of paths
- Forward-reverse links



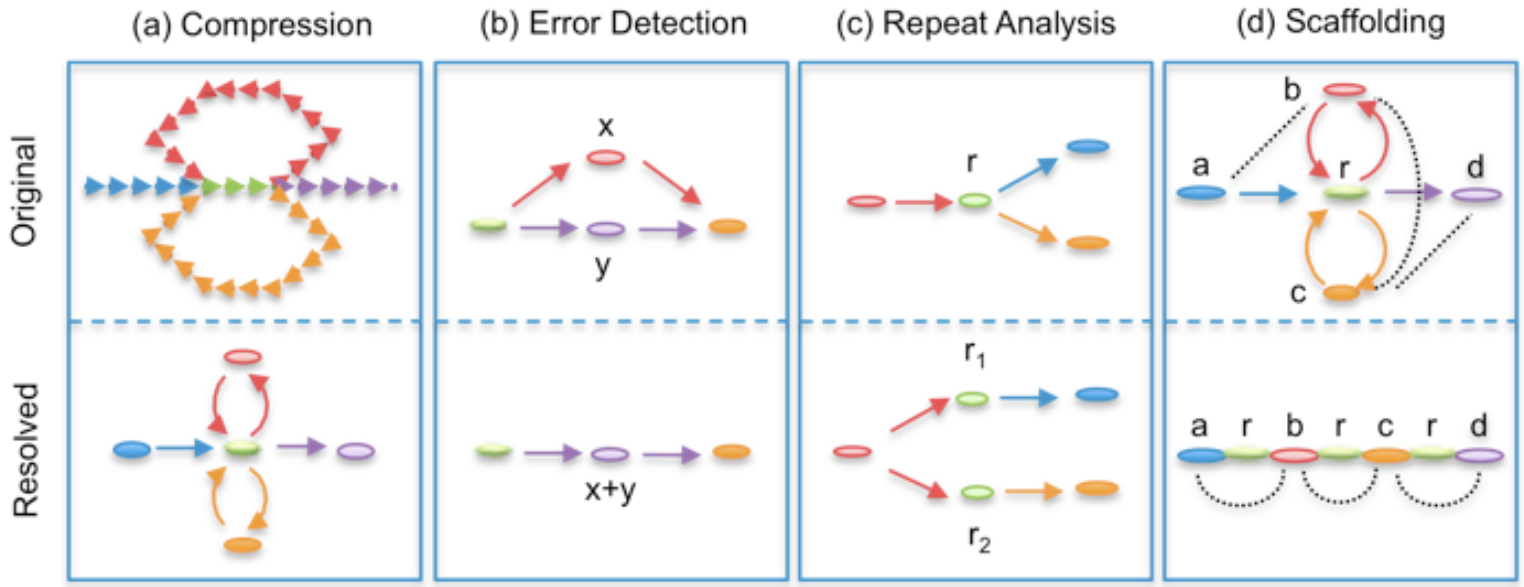
De Brujin Graph formulation



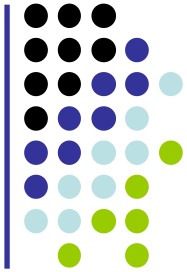
- Given sequence $x_1 \ldots x_N$, k-mer length k,

  Graph of 4^k vertices,

  Edges between words with (k-1)-long overlap



4. Derive Consensus Sequence



Derive multiple alignment from pairwise read alignments

Derive each consensus base by weighted voting

(Alternative: take maximum-quality letter)

Quality of assemblies—mouse

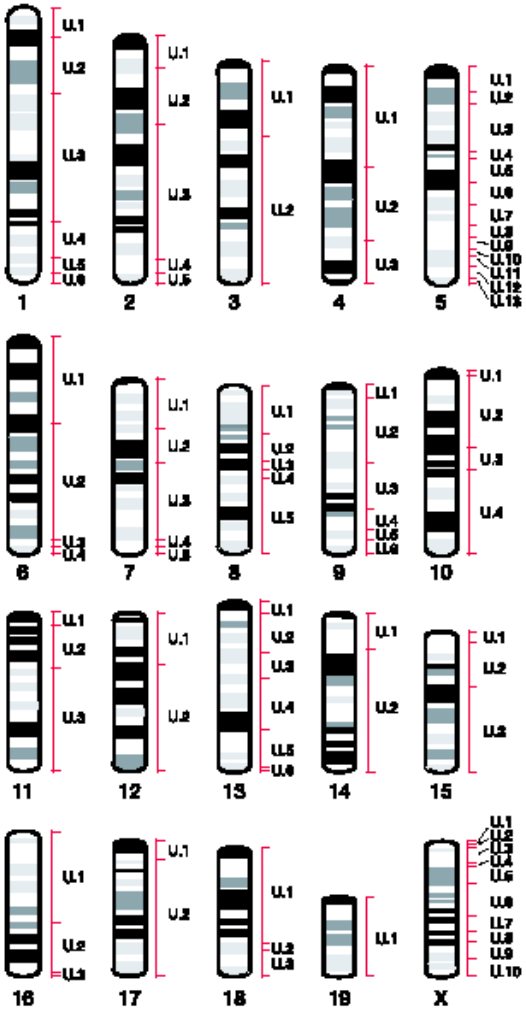


Figure 1 The mouse genome in 88 sequence-based ultracontigs. The position and extent of the 88 ultracontigs of the MGSCv3 assembly are shown adjacent to ideograms of the mouse chromosomes. All mouse chromosomes are acrocentric, with the centromeric end at the top of each chromosome. The supercontigs of the sequence assembly were anchored to the mouse chromosomes using the MIT genetic map. Neighbouring supercontigs were linked together into ultracontigs using information from single BAC links and the fingerprint and radiation-hybrid maps, resulting in 88 ultracontigs containing 95% of the bases in the euchromatic genome.

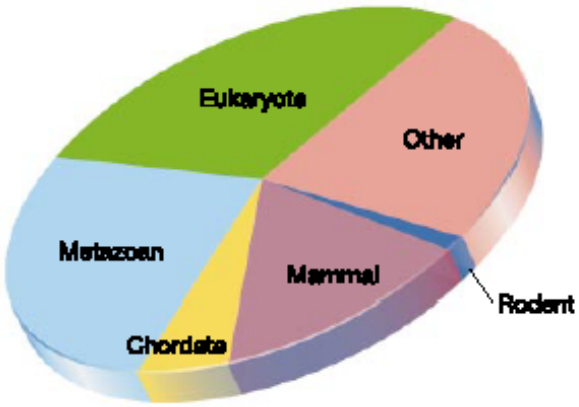
N50 length (kb)*	Bases (Gb)	Bases plus gaps (Gb)	Percentage of genome
25.9	2.372	2.372	94.9
18,600	2.372	2.477	99.1
50,600	2.372	2.493	99.7
2.3	0.106	0.106	–
18,700	2.352	2.455	98.2
22,900	1.955	2.039	81.6

ides spanned gaps.
ercontigs with an N50 value of 3.4 kb. The N50 value for all contigs is 24.8 kb, and for all supercontigs is 16,900 kb (excluding gaps to gaps in the ultracontigs and are thus accounted for in the 'bases plus gaps' estimate.

Terminology: ***N50 contig length***

If we sort contigs from largest to smallest, and start Covering the genome in that order, N50 is the length Of the contig that just covers the 50th percentile.

7.7X sequence coverage



Panda Genome

Table 1 | Summary of the panda genome sequencing and assembly

Step	Paired-end insert size (bp)*	Sequence coverage (×)†	Physical coverage (×)†	N50 (bp) ‡	N90 (bp) ‡	Total length (bp)
Initial contig	110–230; 380–570	38.5	96	1,483	224	2,021,639,596
Scaffold 1				32,648	7,780	2,213,848,409
Scaffold 2	Add 1,700–2,800	8.4	151	229,150	45,240	2,250,442,210
Scaffold 3	Add 3,700–7,500	6.5	450	581,933	127,336	2,297,100,301
Scaffold 4	Add 9,200–12,300	2.6	373	1,281,781	312,670	2,299,498,912
Final contig	All	56.0	1,070	39,886	9,848	2,245,302,481

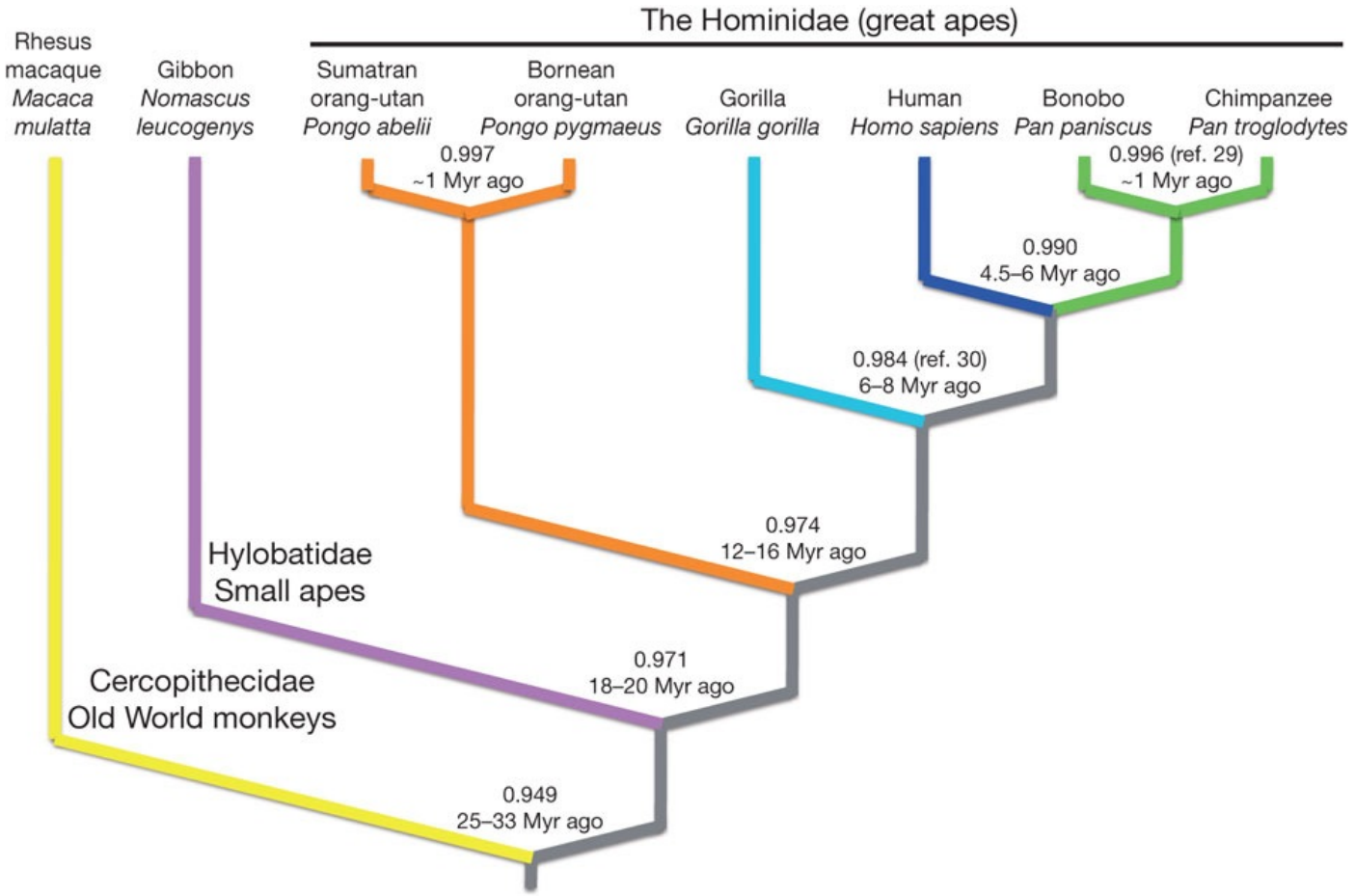
Add denotes accumulative; for example, scaffold 2 uses data of 110–230, 380–570 and 1,700–2,800.

* Approximate average insert size of Illumina Genome Analyser sequencing libraries. The sizes were estimated by mapping the reads onto the assembled genome sequences.

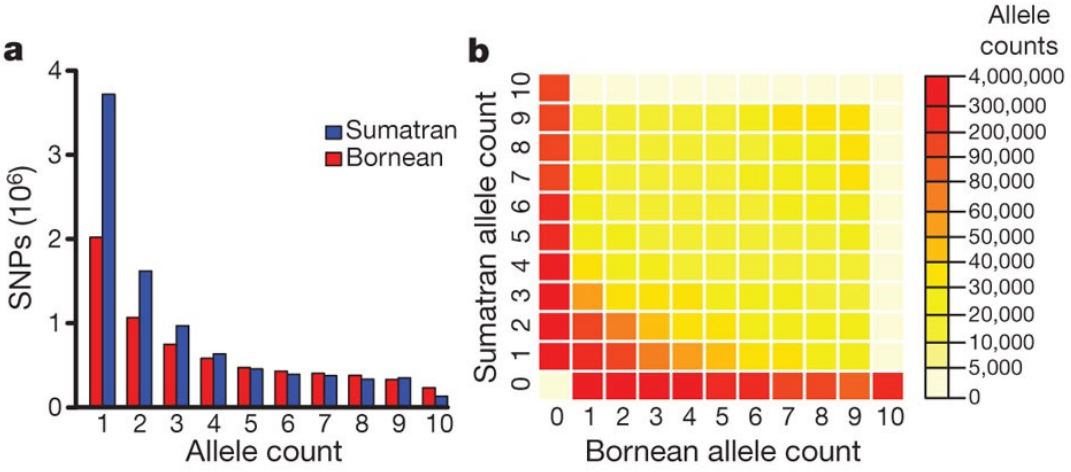
† High-quality read sequences that were used in assembly. Coverage was estimated assuming a genome size of 2.4 Gb. Sequence coverage refers to the total length of generated reads, and physical coverage refers to the total length of sequenced clones of the libraries.

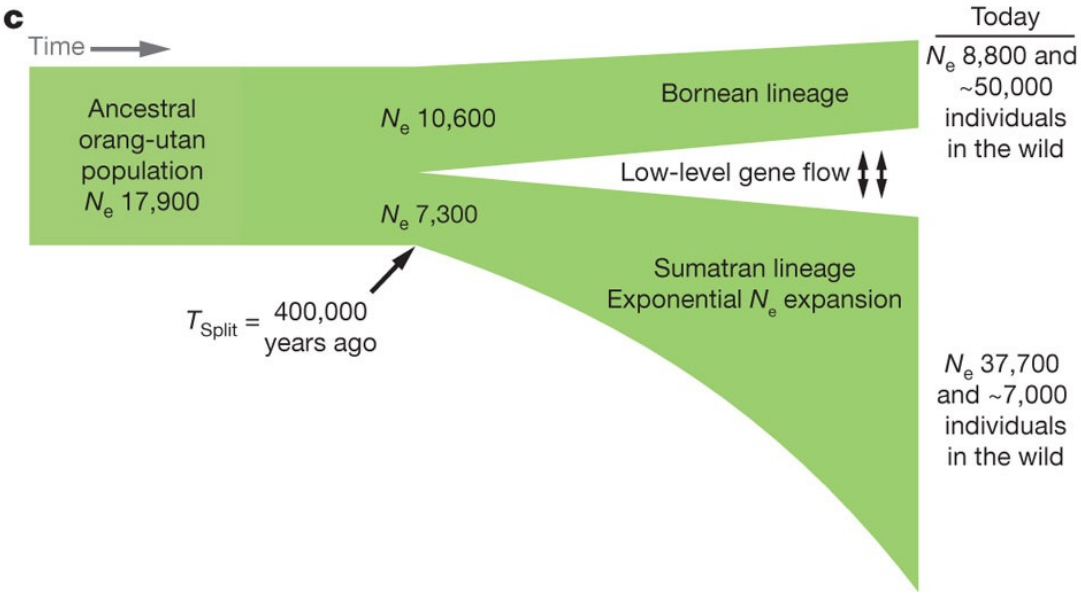
‡ N50 size of contigs or scaffolds was calculated by ordering all sequences then adding the lengths from longest to shortest until the summed length exceeded 50% of the total length of all sequences. N90 is similarly defined.

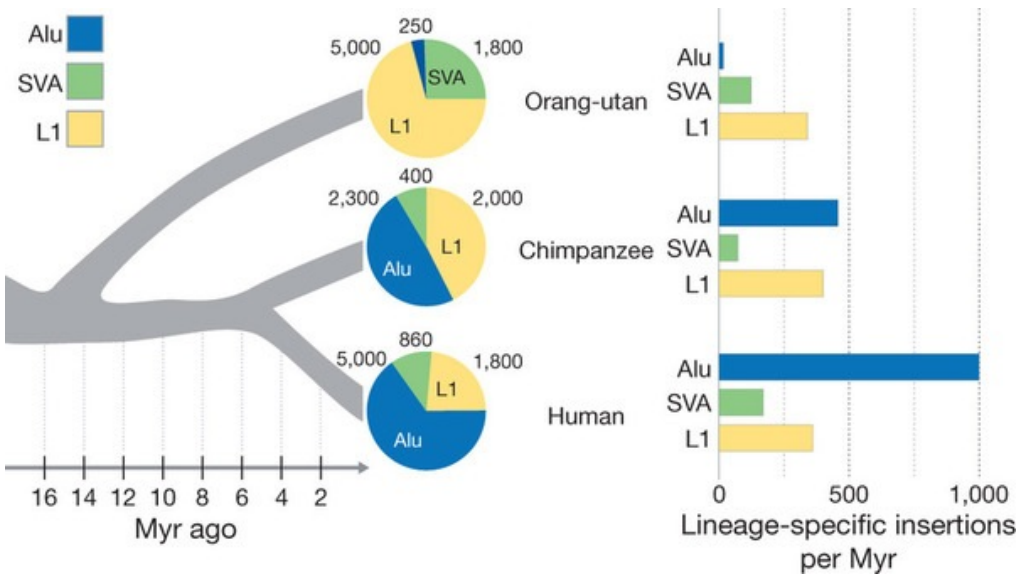
Hominid lineage



Orangutan genome







Assemblathon

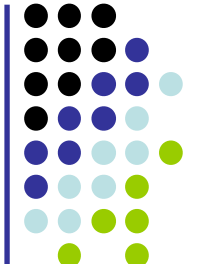


Table 1 Assemblathon 2 participating team details

Team name	Team identifier	Number of assemblies submitted			Sequence data used for bird assembly	Institutional affiliations	Principal assembly software used
		Bird	Fish	Snake			
ABL	ABL	1	0	0	4 + I	Wayne State University	HyDA
ABySS	ABYSS	0	1	1		Genome Sciences Centre, British Columbia Cancer Agency	ABySS and Anchor
Allpaths	ALLP	1	1	0	I	Broad Institute	ALLPATHS-LG
BCM-HGSC	BCM	2	1	1	4 + I + P[1]	Baylor College of Medicine Human Genome Sequencing Center	SeqPrep, KmerFreq, Quake, BWA, Newbler, ALLPATHS-LG, Atlas-Link, Atlas-GapFill, Phrap, CrossMatch, Velvet, BLAST, and BLASR
CBCB	CBCB	1	0	0	4 + I + P	University of Maryland, National Biodefense Analysis and Countermeasures Center	Celera assembler and PacBio Corrected Reads (PBcR)
CoBiG[2]	COBIG	1	0	0	4	University of Lisbon	4Pipe4 pipeline, Seqclean, Mira, Bambus2
CRACS	CRACS	0	0	1		Institute for Systems and Computer Engineering of Porto TEC, European Bioinformatics Institute	ABySS, SSPACE, Bowtie, and FASTX
CSHL	CSHL	0	3	0		Cold Spring Harbor Laboratory, Yale University, University of Notre Dame	Metassembler, ALLPATHS, SOAPdenovo
CTD	CTD	0	3	0		National Research University of Information Technologies, Mechanics, and Optics	Unspecified
Curtain	CURT	0	0	1		European Bioinformatics Institute	SOAPdenovo, fastx_toolkit, bwa, samtools, velvet, and curtain
GAM	GAM	0	0	1		Institute of Applied Genomics, University of Udine, KTH Royal Institute of Technology	GAM, CLC and ABySS
IOBUGA	IOB	0	2	0		University of Georgia, Institute of Aging Research	ALLPATHS-LG and SOAPdenovo
MLK Group	MLK	1	0	0	I	UC Berkeley	ABySS
Meraculous	MERAC	1	1	1	I	DOE Joint Genome Institute, UC Berkeley	meraculous
Newbler-454	NEWB	1	0	0	4	454 Life Sciences	Newbler
Phusion	PHUS	1	0	1	I	Wellcome Trust Sanger Institute	Phusion2, SOAPdenovo, SSPACE
PRICE	PRICE	0	0	1		UC San Francisco	PRICE
Ray	RAY	1	1	1	I	CHUQ Research Center, Laval University	Ray
SGA	SGA	1	1	1	I	Wellcome Trust Sanger Institute	SGA
SOAPdenovo	SOAP	3	1	1	I[2]	BGI-Shenzhen, HKU-BGI	SOAPdenovo
Symbiose	SYMB	0	1	1		ENS Cachan/IRISA, INRIA, CNRS/ Symbiose	Monument, SSPACE, SuperScaffolder, and GapCloser

Table 2 Overview of sequencing data provided for Assemblathon 2 participants

Species	Estimated genome size	Illumina	Roche 454	Pacific biosciences
Bird (*Melopsittacus undulatus*)	1.2 Gbp	285x coverage from 14 libraries (mate pair and paired-end)	16x coverage from 3 library types (single end and paired-end)	10x coverage from 2 libraries
Fish (*Maylandia zebra*)*	1.0 Gbp	192x coverage from 8 libraries (mate pair and paired-end)	NA	NA
Snake (*Boa constrictor constrictor*)	1.6 Gbp	125x coverage from 4 libraries (mate pair and paired-end)	NA	NA

Assemblathon

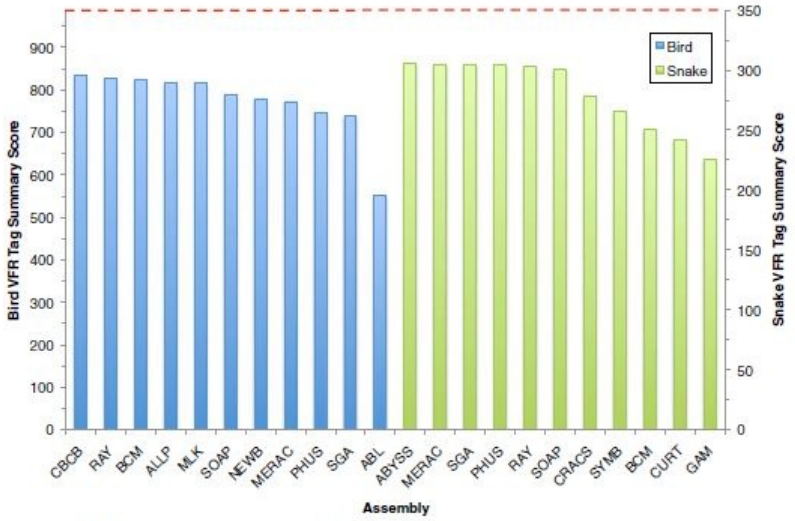


Figure 12 Short-range scaffold accuracy assessment via Validated Fosmid Regions. First, validated Fosmid regions (VFRs) were identified (86 in bird and 56 in snake, see text). Then VFRs were divided into non-overlapping 1,000 nt fragments and pairs of 100 nt 'tags' were extracted from ends of each fragment and searched (using BLAST) against all scaffolds from each assembly. A summary score for each assembly was calculated as the product of a) the number of pairs of tags that both matched the same scaffold in an assembly (at any distance apart) and b) the percentage of only the uniquely matching tag pairs that matched at the expected distance (± 2 nt). Theoretical maximum scores, which assume that all tag-pairs would map uniquely to a single scaffold, are indicated by red dashed line (988 for bird and 350 for snake).

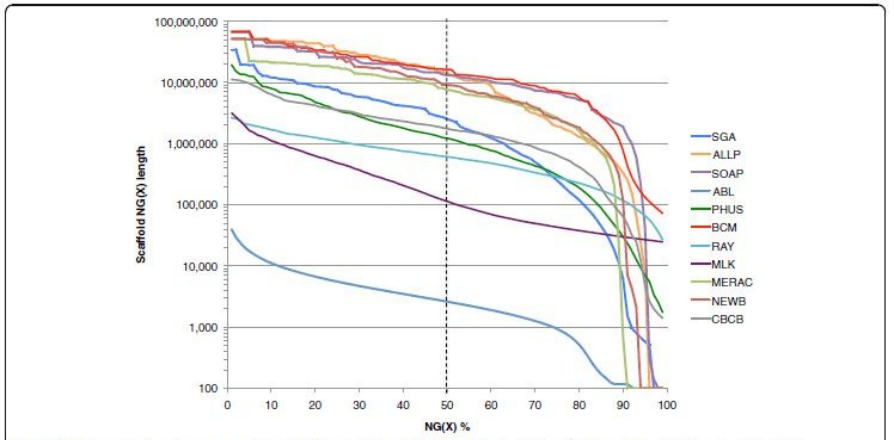


Figure 1 NG graph showing an overview of bird assembly scaffold lengths. The NG scaffold length (see text) is calculated at integer thresholds (1% to 100%) and the scaffold length (in bp) for that particular threshold is shown on the y-axis. The dotted vertical line indicates the NG50 scaffold length: if all scaffold lengths are summed from longest to the shortest, this is the length at which the sum length accounts for 50% of the estimated genome size. Y-axis is plotted on a log scale. Bird estimated genome size = ~1.2 Gbp.

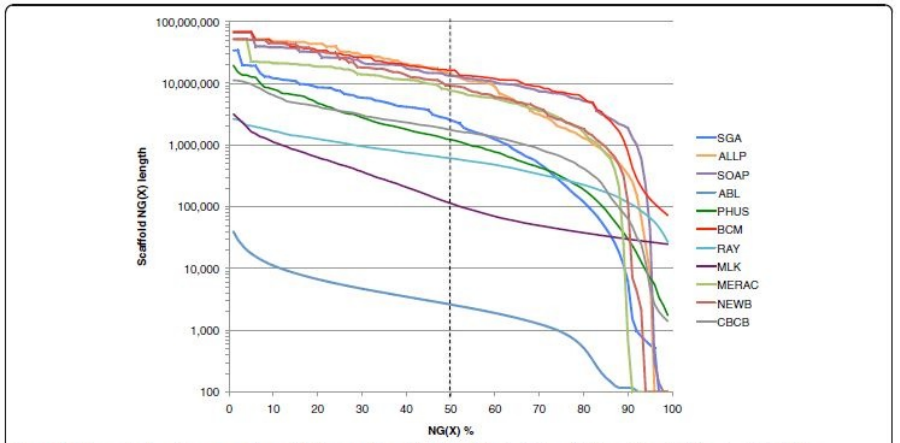


Figure 1 NG graph showing an overview of bird assembly scaffold lengths. The NG scaffold length (see text) is calculated at integer thresholds (1% to 100%) and the scaffold length (in bp) for that particular threshold is shown on the y-axis. The dotted vertical line indicates the NG50 scaffold length: if all scaffold lengths are summed from longest to the shortest, this is the length at which the sum length accounts for 50% of the estimated genome size. Y-axis is plotted on a log scale. Bird estimated genome size = ~1.2 Gbp.

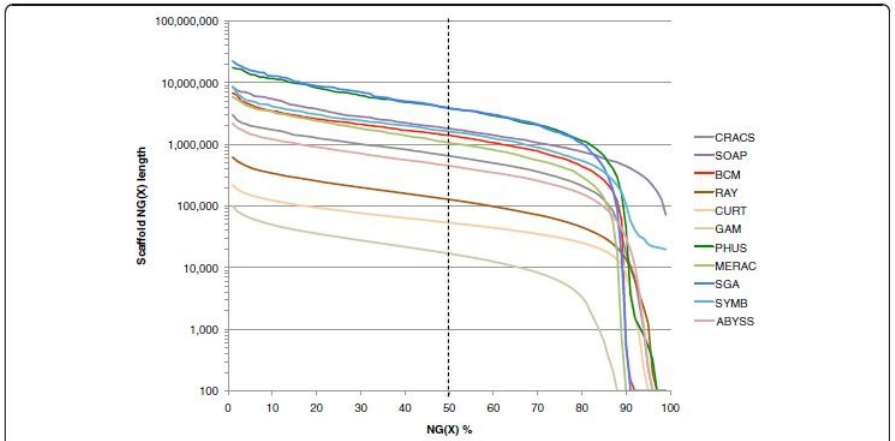


Figure 3 NG graph showing an overview of snake assembly scaffold lengths. The NG scaffold length (see text) is calculated at integer thresholds (1% to 100%) and the scaffold length (in bp) for that particular threshold is shown on the y-axis. The dotted vertical line indicates the NG50 scaffold length: if all scaffold lengths are summed from longest to the shortest, this is the length at which the sum length accounts for 50% of the estimated genome size. Y-axis is plotted on a log scale. Snake estimated genome size = ~1.0 Gbp.

History of WGA

1997

Let's sequence the human genome with the shotgun strategy

That is impossible, and a bad idea anyway

Gene Myers

Phil Green