Genome assemblies

Isheng Jason Tsai

Introduction to NGS Data and Analysis
Lecture 4

Lecture outline

- Introduction
- Assembly algorithm overview
- Long read technologies
- Scaffolding
  - Chromosome conformation capture

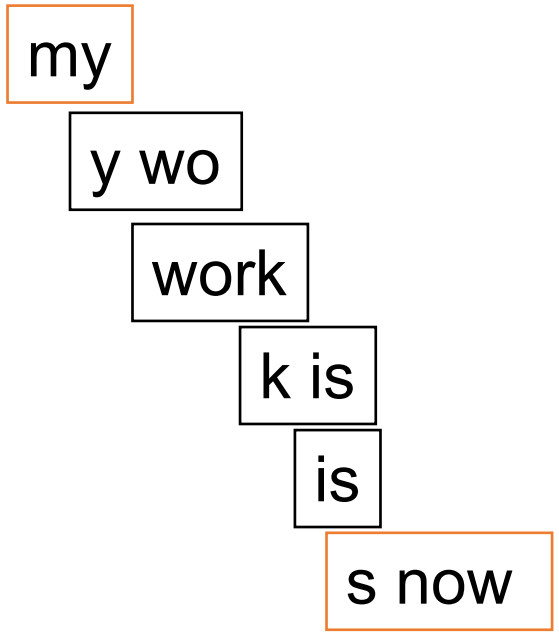
- Case studies

Problem

We accidently printed my five copies of “Origin of Species”, and shredded into pieces

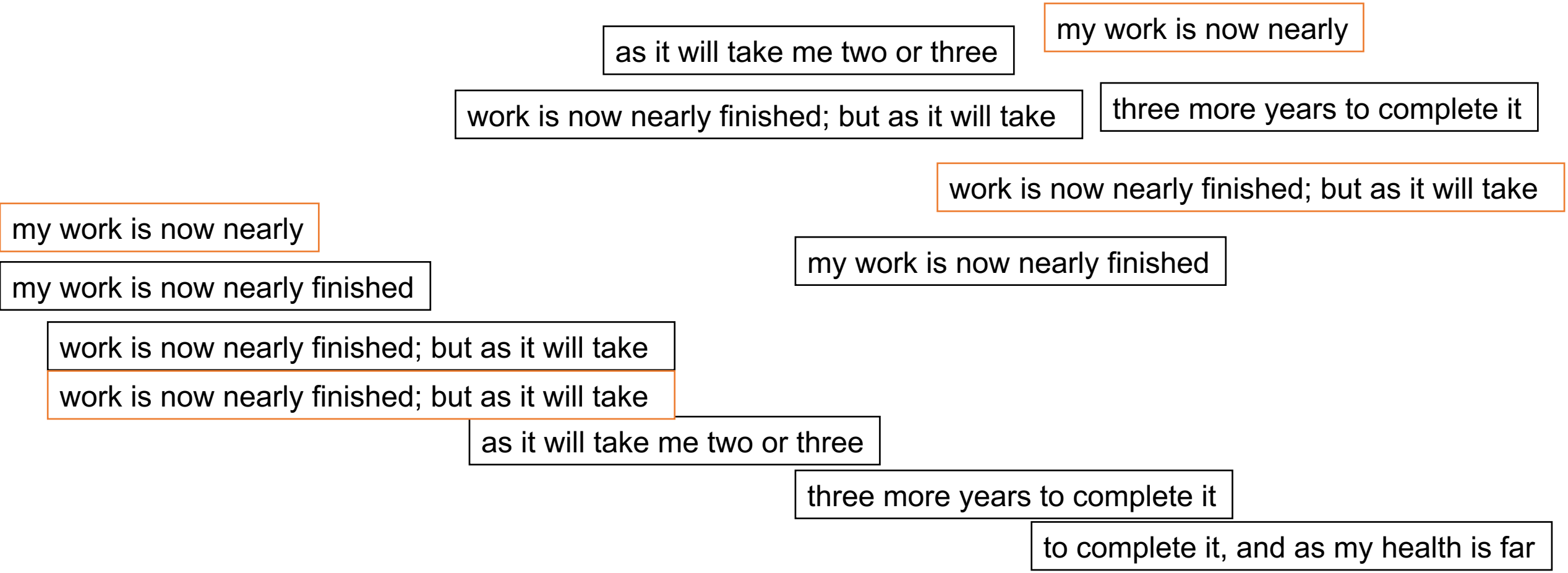
my work is now nearly finished; but as it will take me two or three more years to complete it, and as my health is far from strong, I have been urged to publish this Abstract. I have more especially been induced to do this, as Mr Wallace, who is now studying the natural history of the Malay archipelago, has arrived at almost exactly the same general conclusions that I have on the origin of species. Last year he sent to me a memoir on this subject, with a request that I would forward it to Sir Charles Lyell, who sent it to the Linnean Society, and it is published in the third volume of the journal of that Society. Sir C. Lyell and Dr Hooker, who both knew of my work -- the latter having read my sketch of 1844 -- honoured me by thinking it advisable to publish, with Mr Wallace's excellent memoir, some brief extracts from my manuscripts.

Assemby = Piecing the pieces together



my work is now

Long shredded pieces (read) = easier assembly

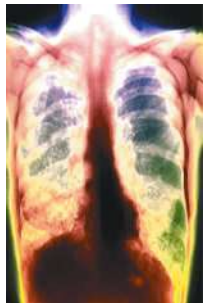


my work is now nearly finished; but as it will take me two or three more years to complete it, and as my health is far

Why sequence a genome?

Genomics advance our understanding of organisms across tree of life

pathogens



TB

model

C. elegans

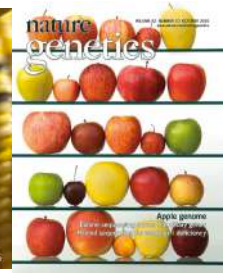
model

fruitfly

model

human

NGS



pathogens

Blood fluke

Tapeworms

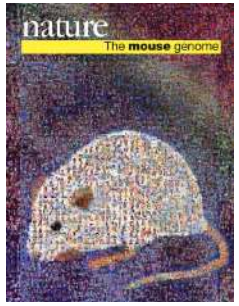
1996 1997 1998 1999 2000 2001 2002 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

S. cerevisiae

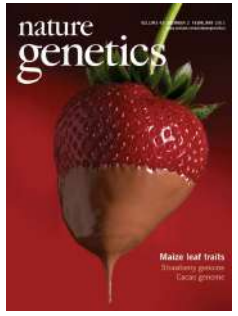
model

Plasmodium

pathogens



model



Phytophthora infestans
(potato blight)

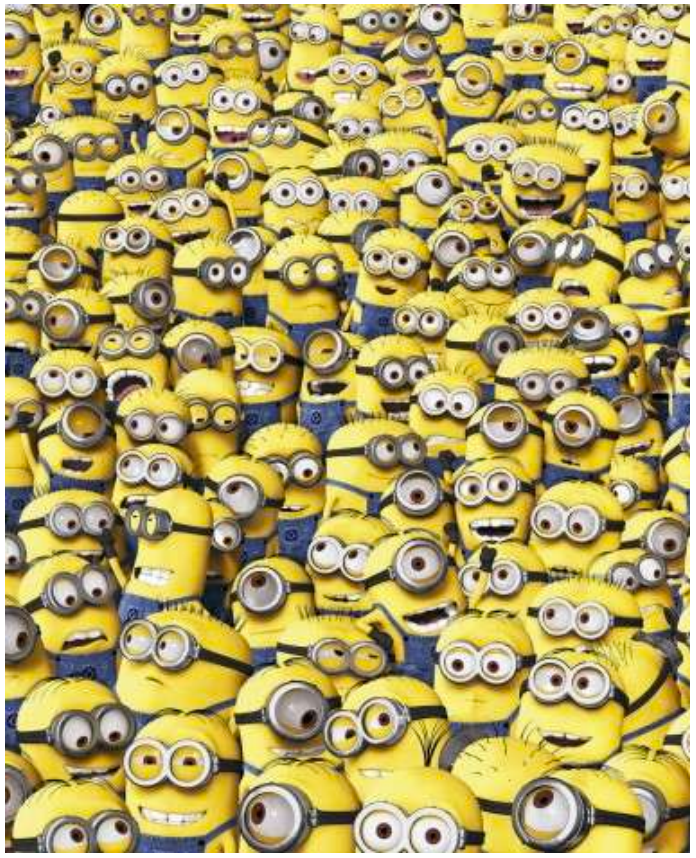
Black death

Smut fungi

pathogens

Assemble a genome is hard

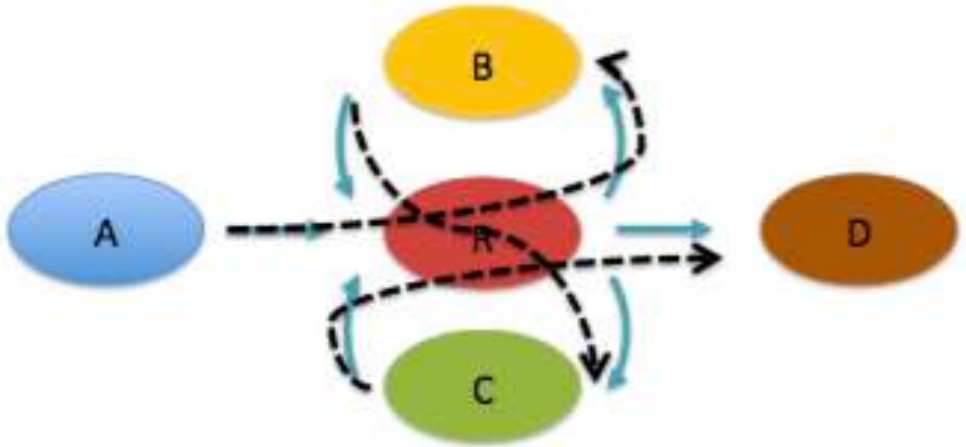
Less complicated



Complicated

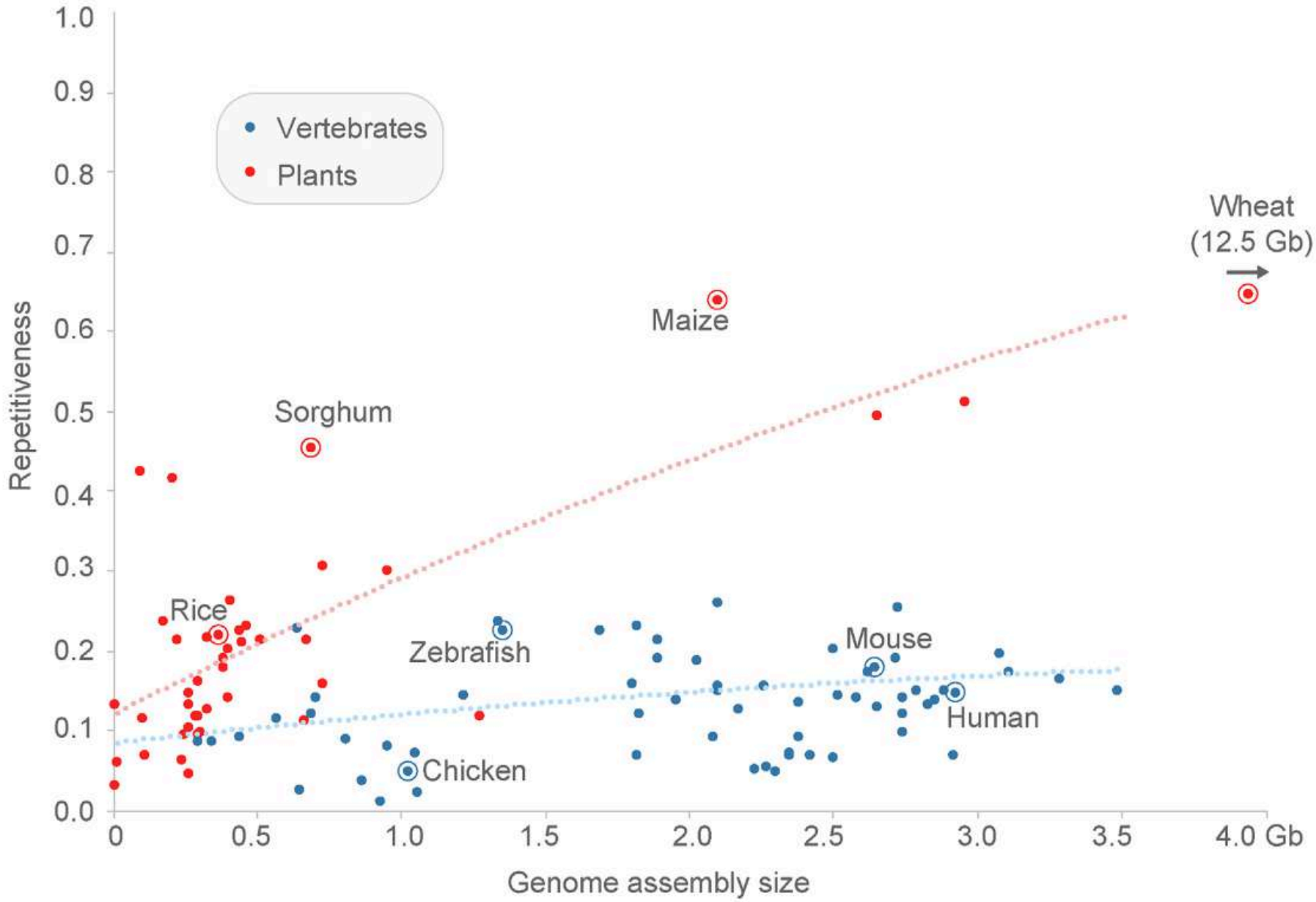
Repeats

You can't 100% resolve repeats repeats unless you have your sequence length > repeat



Credit: Michael Schatz

Repeat content of plant and vertebrate genomes



https://doi.org/10.1016/j.pbi.2017.02.002

Largest genome ever recorded: 149Gb

Organism Type	Organism Name	Approximate Genome size, in number of nucleotides ("letters")	Number of protein-coding genes
Bacterium	Nasuia deltocephalinicola, a tiny bacterium that lives inside an insect [3]	112,000 (0.112 million) * currently the smallest known bacterial genome	137
Bacterium	Escherichia coli [2]	4,600,000 (4.6 million)	5,000
Plant	Arabidopsis thaliana	135,000,000 (135 million)	27,416
Mammal	Homo sapiens, Humans	3,000,000,000 (3 billion)	20,000 [5]
Plant	Norway Spruce	19,000,000,000 (19 billion)	28,000
Plant	Paris japonica, a rare Japanese flower [4]	149,000,000,000 (149 billion) * currently the largest known genome	unknown

http://thatslifesci.com/2019-05-06-Biggest-genome-of-them-all-JBarnett/

QC and understand your data before assembly

Conversation:

Jason, we don’t know why our assembly is not as good as yours. We have about 90X of Mate pair data..

Email 2 weeks later:

Dear Jason,

I run the Trimmomatic analysis for the raw data of mate-pair libraries (10kL2_1, 10kL7_1, 15kL3_1 and 15kL8_4 as examples) with a custom adapter file containing mate-pair adapter sequences (junction and external adaptors, you may find them in the attached technote pdf file) and found **that over 80% reads of the library were dropped out.**

Always be careful of contamination

PNAS

Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade

Thomas C. Boothby[a,1], Jennifer R. Tenlen[a,2], Frank W. Smith[a], Jeremy R. Wang[a,b], Kiera A. Patanella[a], Erin Osborne Nishimura[a], Sophia C. Tintori[a], Qing Li[c], Corbin D. Jones[a], Mark Yandell[c], David N. Messina[d], Jarret Glasscock[d], and Bob Goldstein[a]

[a]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; [b]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599; [c]Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112; and [d]Cofactor Genomics, St. Louis, MO 63110

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved September 28, 2015 (received for review May 28, 2015)

17.5 % HGT

PNAS

No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*

Georgios Koutsovoulos[a], Sujai Kumar[a], Dominik R. Laetsch[a,b], Lewis Stevens[a], Jennifer Daub[a], Claire Conlon[a], Habib Maroon[a], Fran Thomas[a], Aziz A. Aboobaker[c], and Mark Blaxter[a,1]

[a]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom; [b]The James Hutton Institute, Dundee DD2 5DA, United Kingdom; and [c]Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved March 1, 2016 (received for review January 8, 2016)

0.4 % HGT

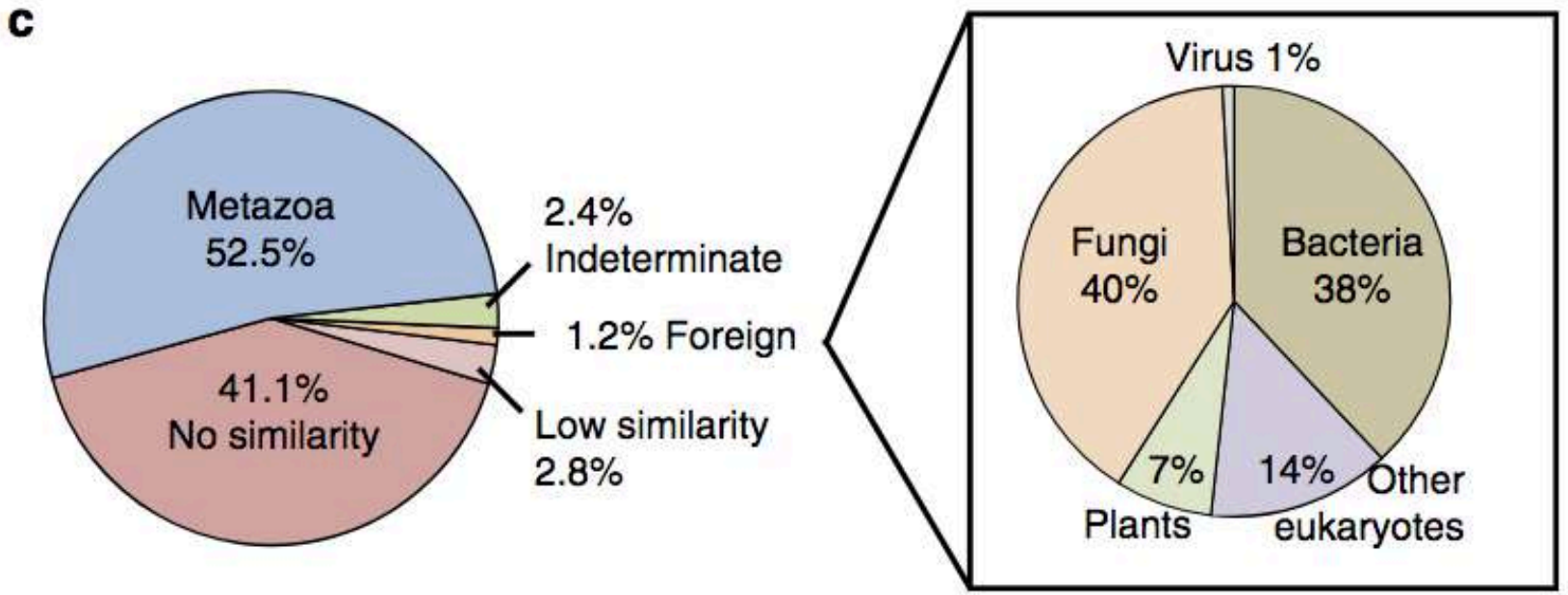
Always be careful of contamination

ARTICLE

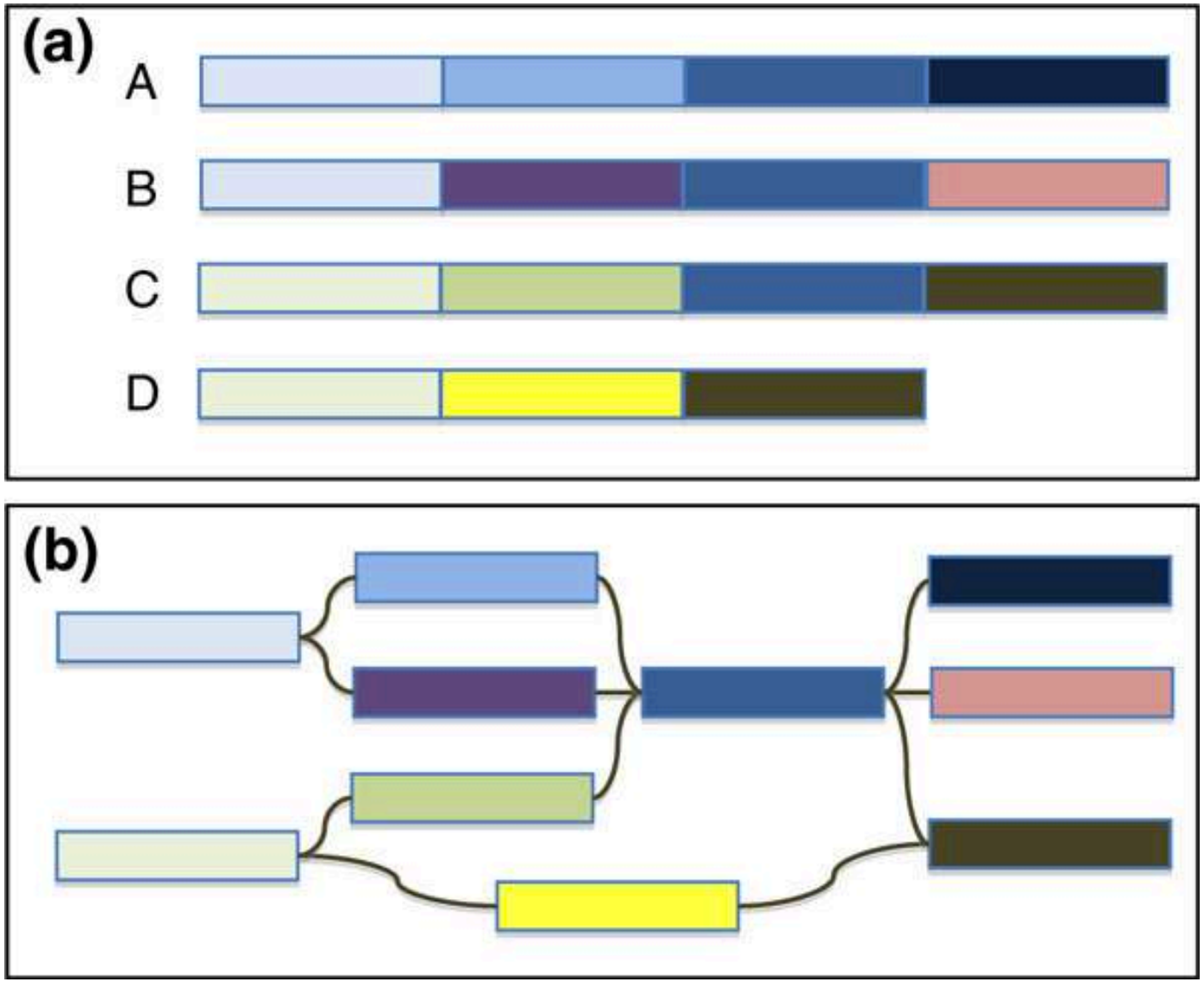
Received 21 Jun 2015 | Accepted 3 Aug 2016 | Published 20 Sep 2016

DOI: 10.1038/ncomms12808 OPEN

Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein



Ploidy, heterozygosity and the assembly graph



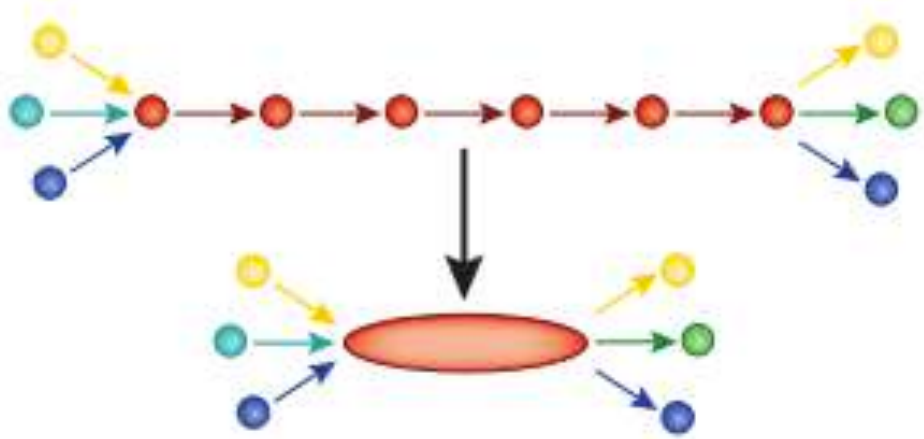
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3446297/

Assembly process

1. Fragment DNA and sequence

2. **Find overlaps** between reads

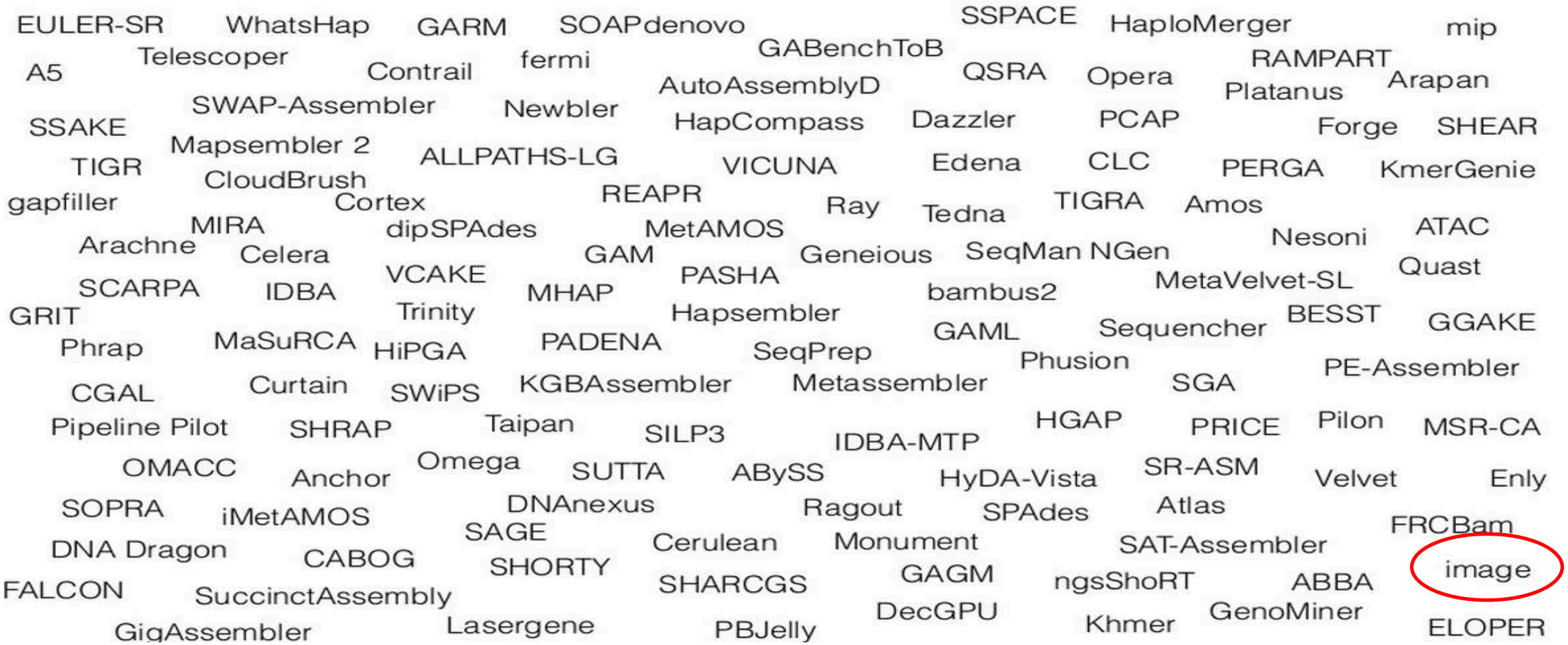
3. **Assemble overlaps** into **contigs**



4. Assemble **contigs** into **scaffolds**

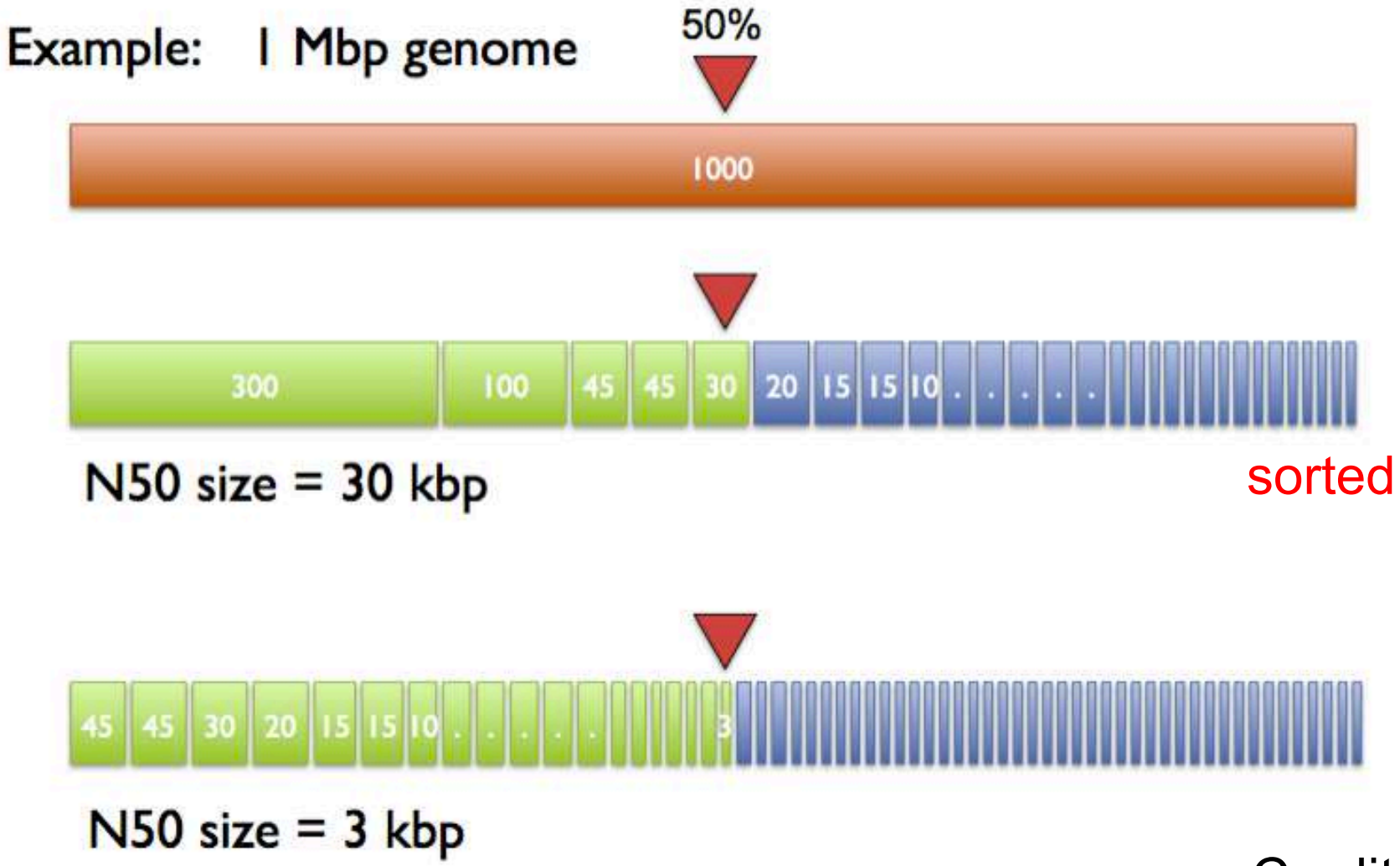
http://www.nature.com/nmeth/journal/v9/n4/pdf/nmeth.1935.pdf

Which program to choose?



Contiguity (good) is a genome? **N50**

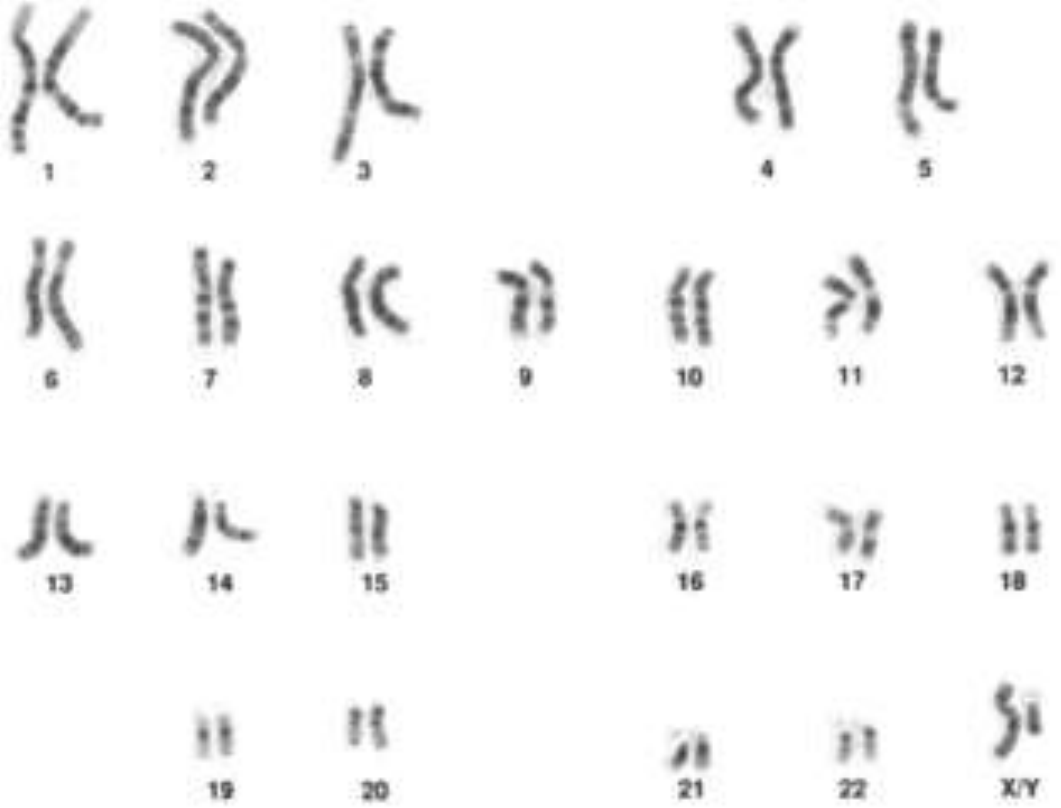
Definition: 50% of genome in contigs/scaffolds of length **N50 bp** or greater



Credit: Michael Schatz

Most assemblies are fragmented

!=



Contigs

chromosomes

Statistics for current GenBank assemblies.

a

Most genomes in this N50 range

Contig N50

Number of vertebrate genomes		Less than 10 kb	10 kb to 100 kb	100 kb to 1 Mb	1 Mb to 10 Mb	Greater than 10 Mb
Diploid human	21	–	8	6	3	4
Non-human mammal	196	34	34	14	1	3
Non-mammal	193	43	133	14	3	0
Total	410	77	285	34	7	7

Genomes with highest N50

doi:10.1038/ng.3824

Why do we need a good assembly?

It is easier to analyse

10000 pieces vs. 23 pieces (chromosomes)

Allows more accurate representation of genes locations on genome

Is gene A close to gene B? On the same chromosome?

Transposon dynamics

Missing in most assemblies (located in NNNNNNNNNN gaps)

Responsibility to contribute to your community

Do you want others to work on the same genome / species as well?

Bottom line:

It’s really no point to do one if you can’t produce an accurate and useful assembly

Assembly qualities

	Whole Genome Representation	Sequence Status	Genes	Usability
1	Incomplete for non-repetitive regions	Small scaffolds and contigs	Incomplete genes	Markers development
2	Complete for non-repetitive regions	Medium scaffolds and contigs	Complete but 1-2 genes/contig	Gene mining
3	Complete for non-repetitive regions	Large scaffolds and contigs	Several dozens of genes/contig	Microsynteny
4	Complete for almost the whole genome	Pseudomolecules	Hundreds of genes/ contig	Any (Synteny, Candidate gene by QTLs)
5	Complete genome	Pseudomolecules	Thousands of genes/ contig	

Credit: Aureliano Bombarely

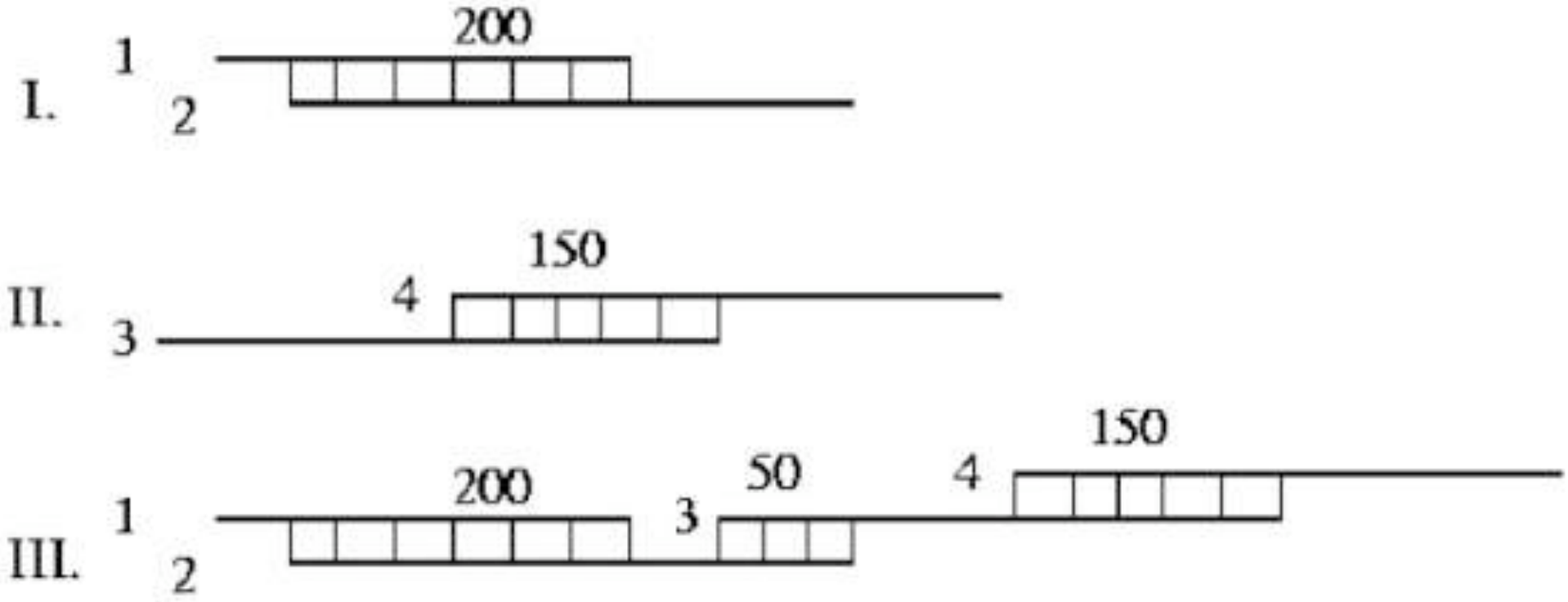
Assembly algorithms

Approaches

- Greedy extension
  - only mentioned for historical reasons
- Overlap – Layout – Consensus (**OLC**) assembly: ’traditional’ and well established method, but challenging to implement at each stage
  - Most ”old” and “newest” assemblies were produced using this approach
- de Bruijn graph (**DBG**) assembly

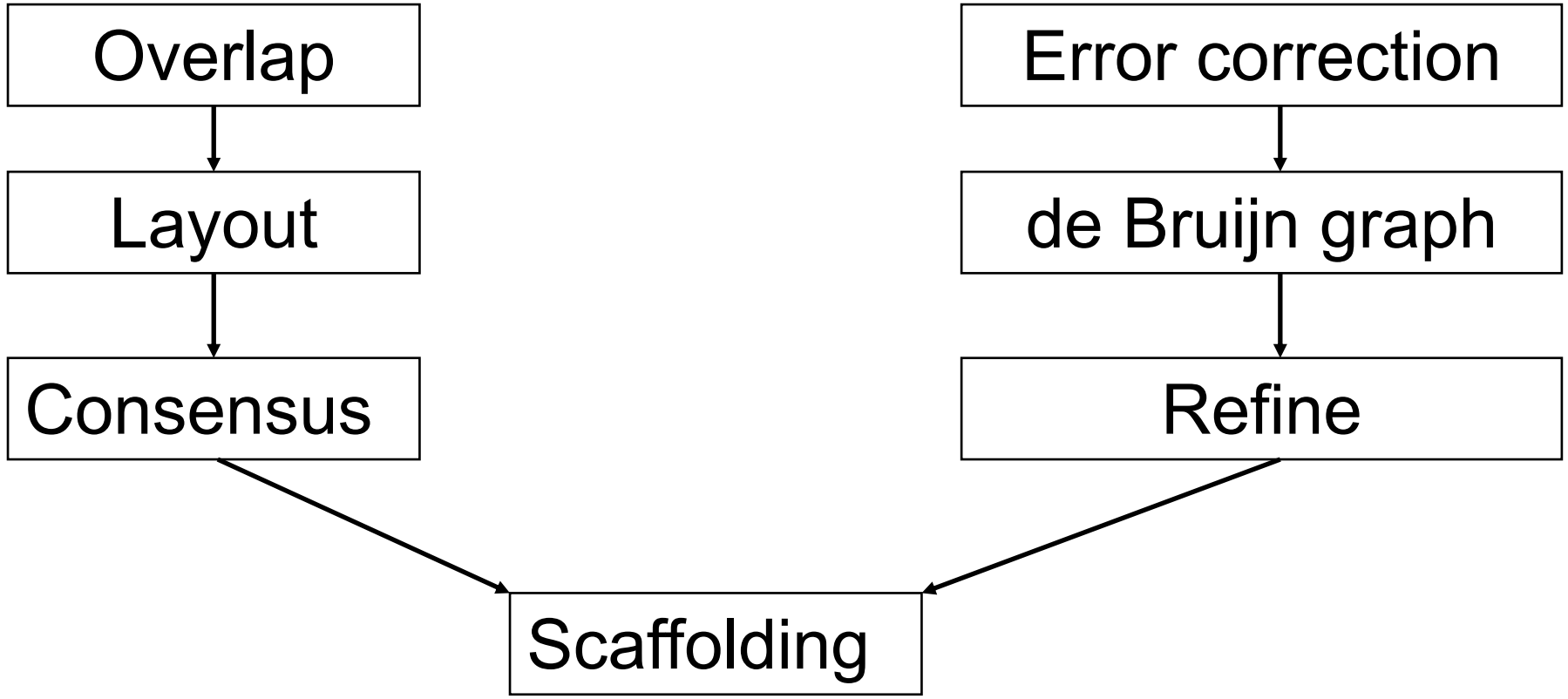
Greedy extension

- Oldest and not really useful in most cases



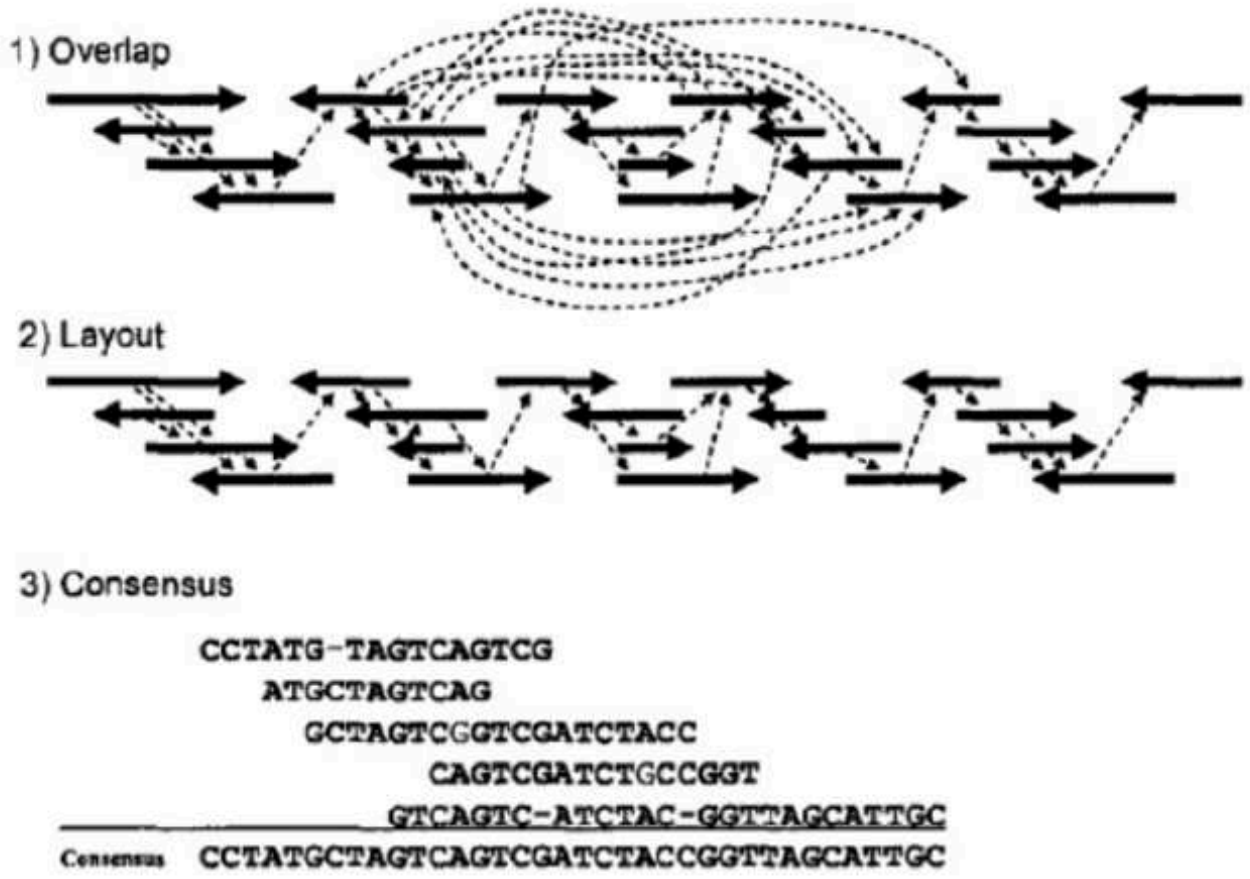
http://www.cbcb.umd.edu/research/assembly_primer.shtml

OLC and DBG assemblers



OLC approach

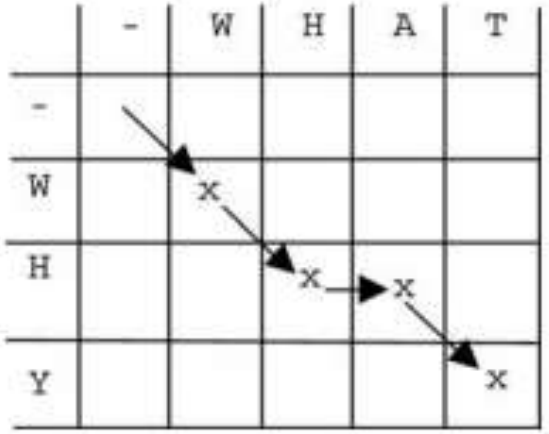
- Pairwise alignment of all reads to find **overlaps**
- **Layout** the reads to decide which read align to which
- Get **consensus** by joining join the read sequences, merging overlaps
- All three are challenging

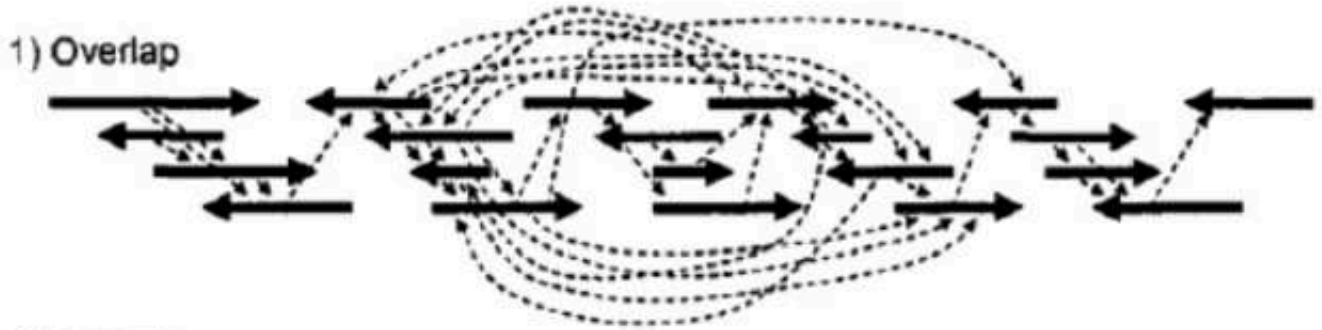


Overlap

- All vs. all pairwise alignment
  - Smith-Waterman? Blast? Kmer based? Suffix tree? Dynamic programming?

- Computationally very intensive but can be parallelised
  - Need lots of CPUs!
  - Batch1 align Batch 1 in 1st CPU
  - Batch1 align Batch 2 in 2nd CPU





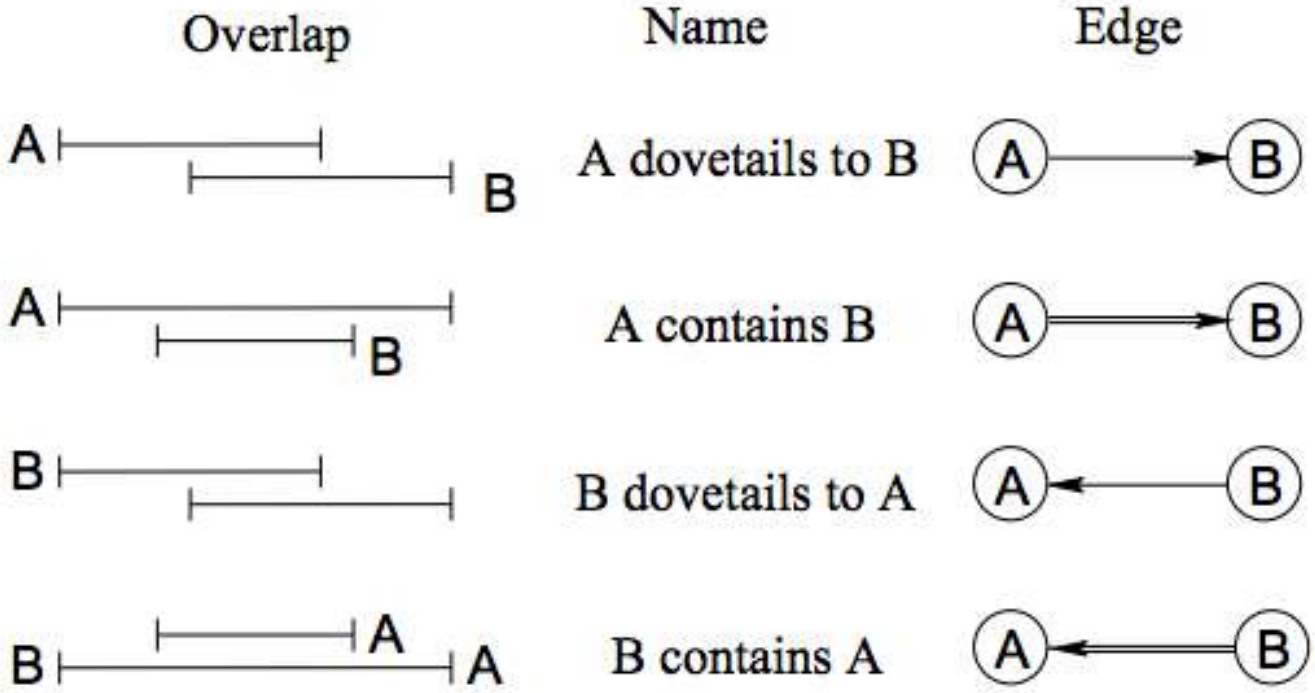
Computational Example 8.1: Pseudocode for overlap alignment

Input sequences A, B
Set $O_{i,0} = O_{0,j} = 0$ for all i, j
for $i = 1$ to n
    for $j = 1$ to m
        $O_{i,j} = \max\{O_{i-1,j} - \delta, O_{i-1,j-1} + s(a_i, b_j), O_{i,j-1} - \delta\}$
    end
end
Best overlap $= \max\{O_{i,m}, O_{n,j}; 1 \leq i \leq n, 1 \leq j \leq m\}$

Deonier *et al.,* Computational Genome Analysis

Build overlap graph

- It's common practice to represent them in **graphs**
- The actual overlaps are the edges
- Now we create the genome assembly graph



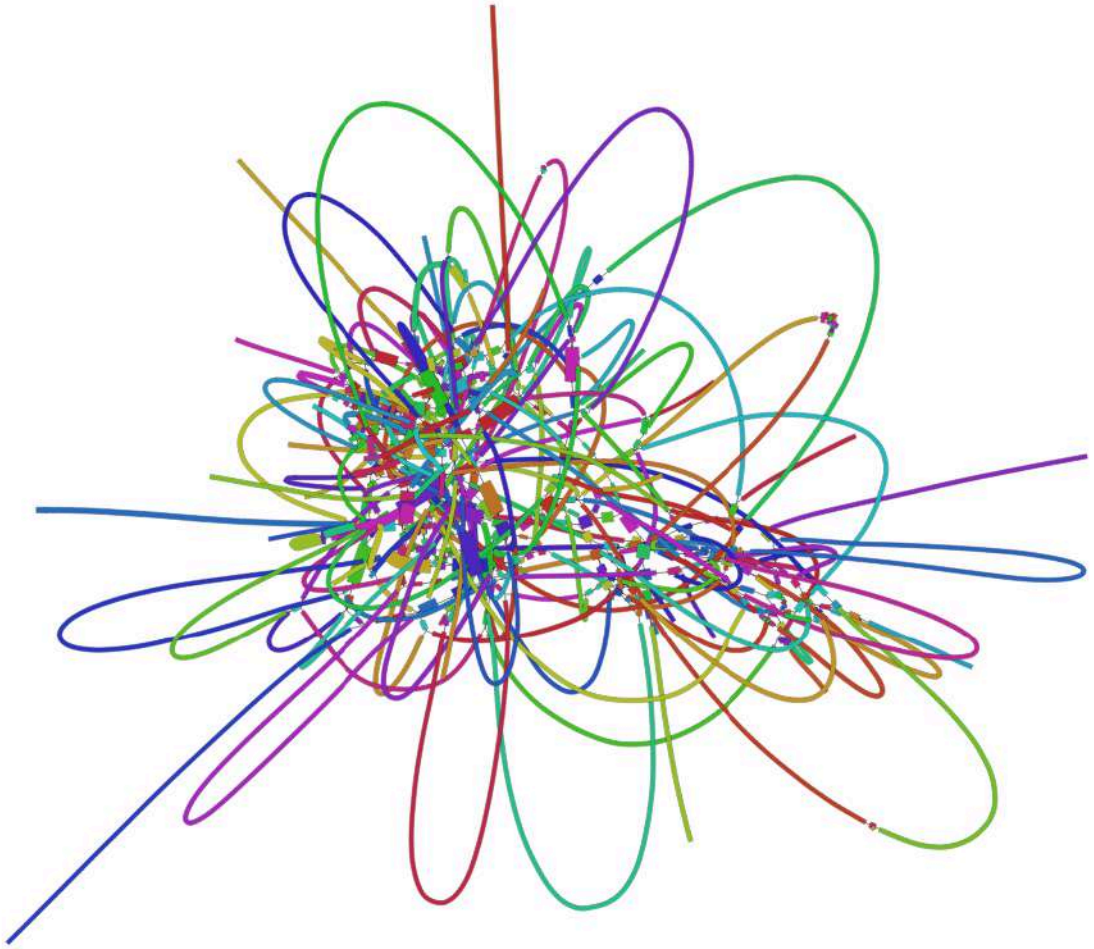
my work is now nearly finished

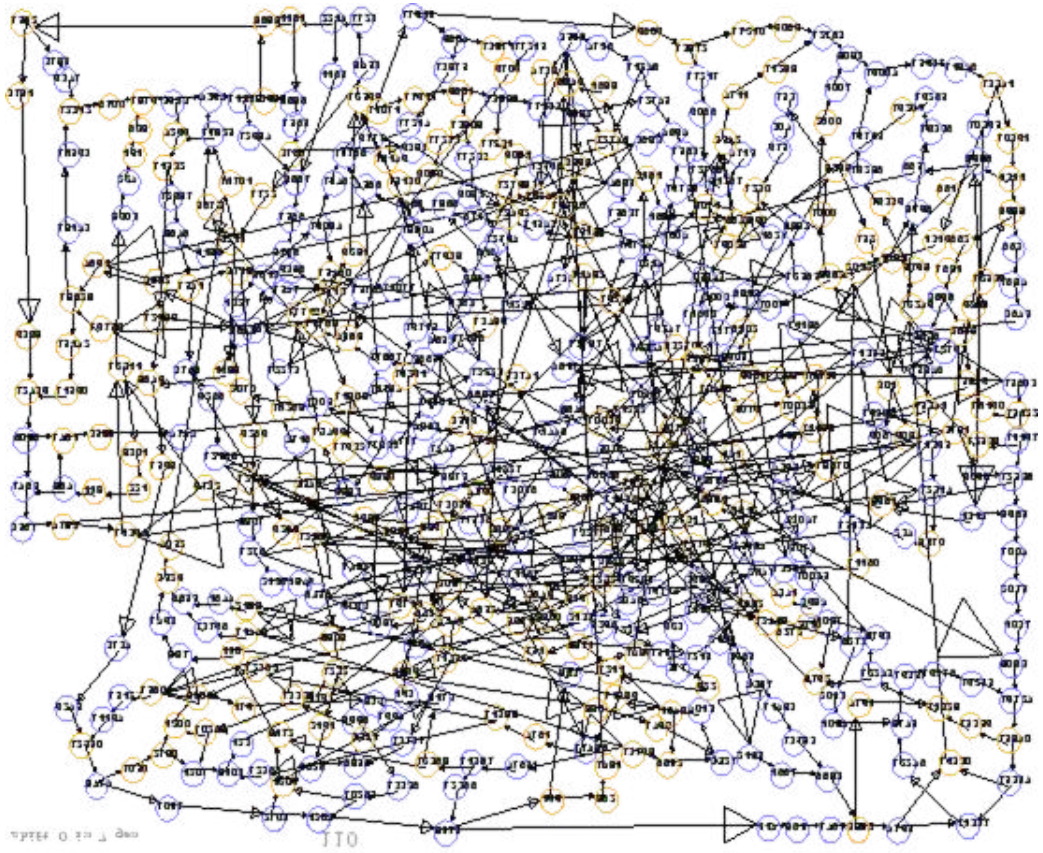
work is now nearly finished; but as it will take

my work is now nearly finished → work is now nearly finished; but as it will take

Kececioglu *et al.,* 1993

Some assembly graph can be complicated…





http://rrwick.github.io/Bandage/images/screenshots/screenshot02.png

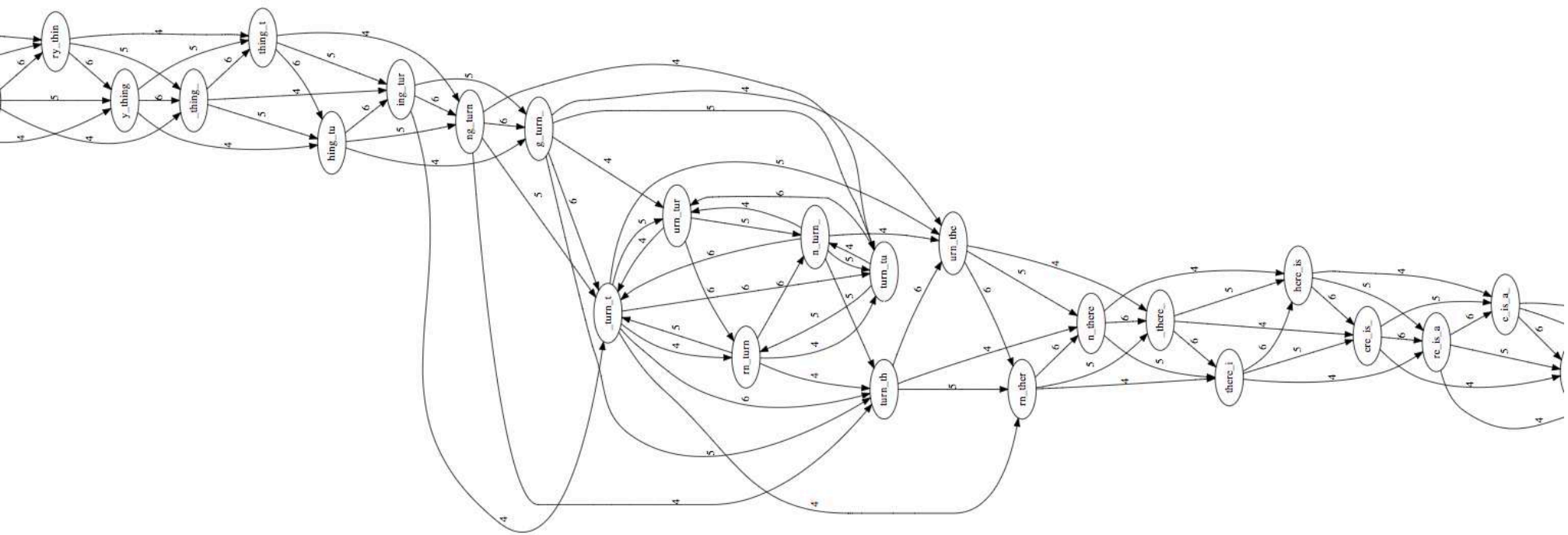
Layout

Overlap graph is big and messy. Contigs don't "pop out" at us.

Below: part of the overlap graph for

`to_every_thing_turn_turn_turn_there_is_a_season`

$l = 4, k = 7$



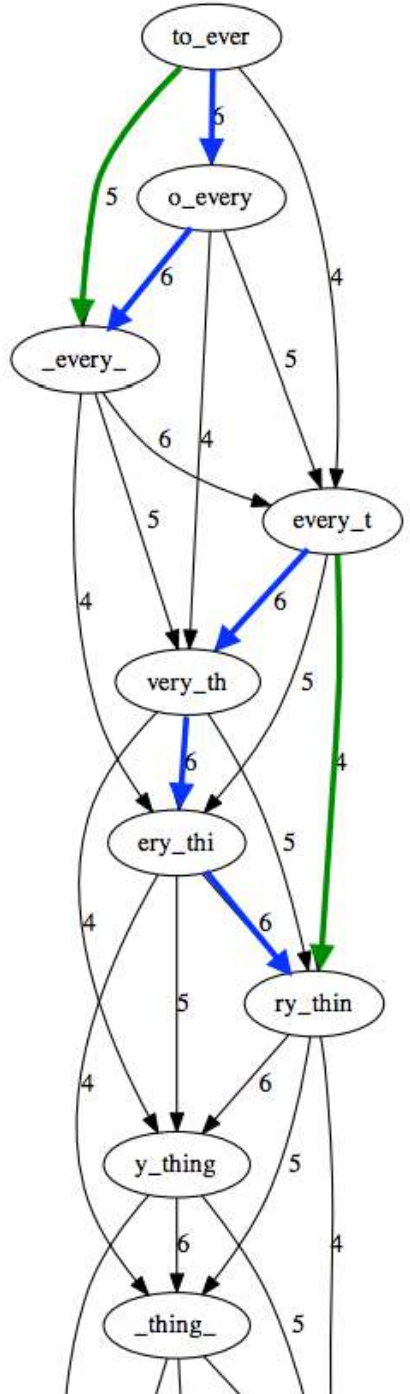
Layout

Order the reads into a consistent manner

Anything redundant about this part of overlap graph?

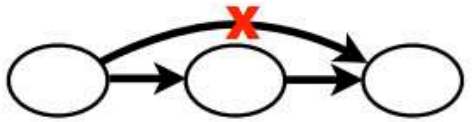
Some edges can be inferred from other edges

E.g., green edge can be inferred from blue

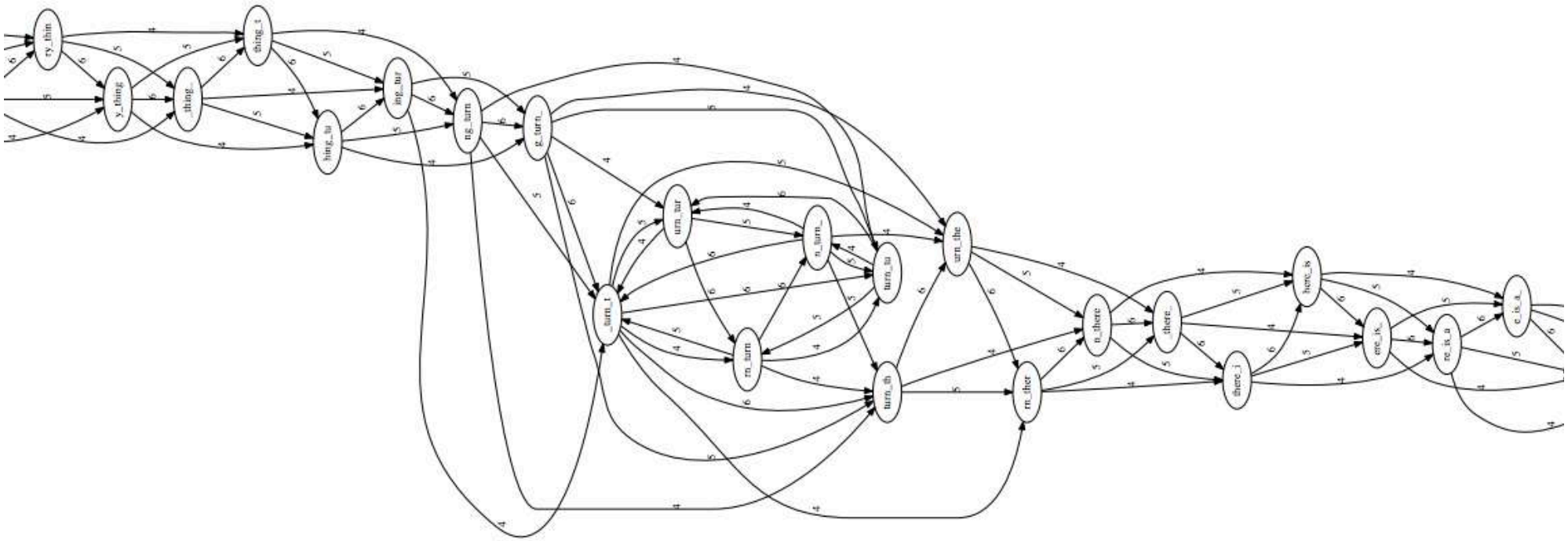


Layout

Remove transitively-inferrible edges, starting with edges that skip one node:



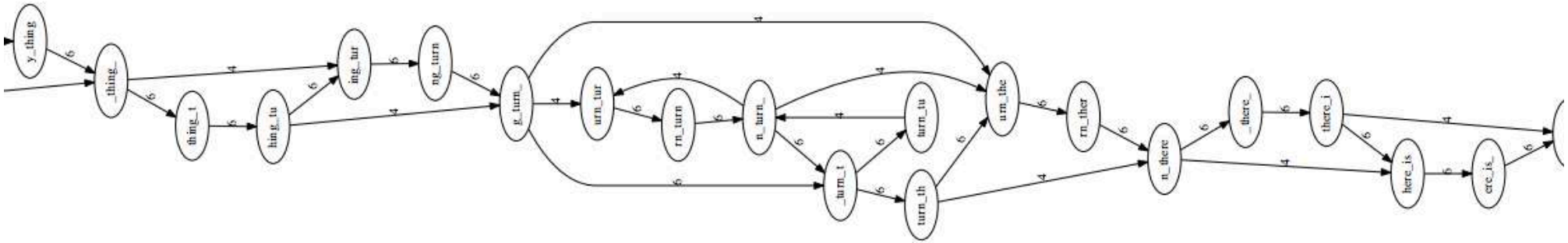
Before:



Layout

Remove transitively-inferrible edges, starting with edges that skip one node:

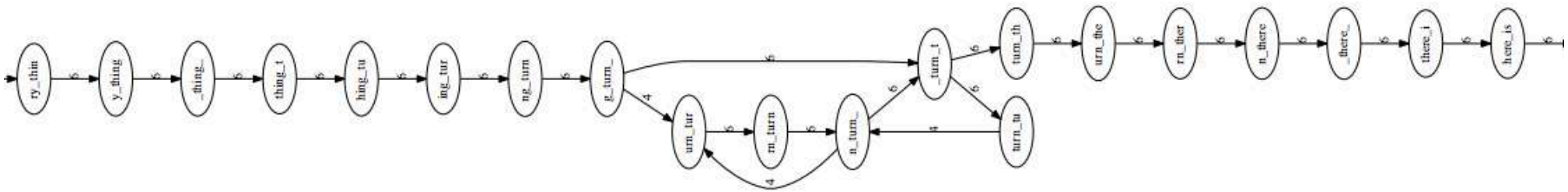
After:



Layout

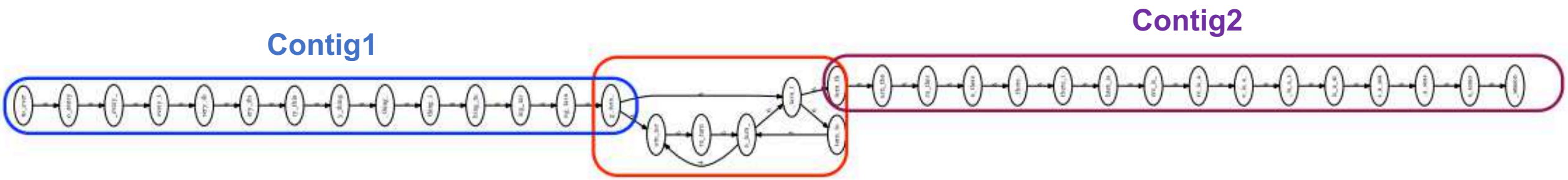
Remove transitively-inferrible edges, starting with edges that skip one *or two* nodes:

After:



Even simpler

Layout

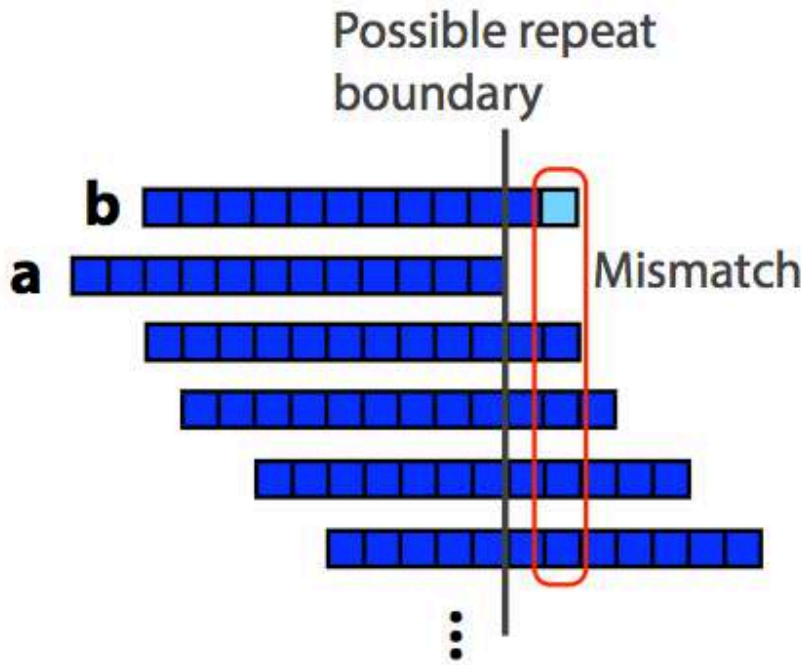


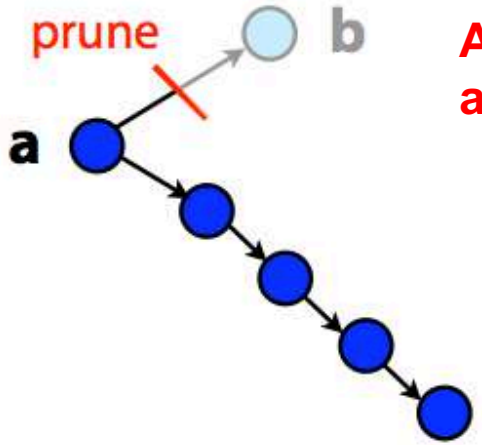
Unresolvable repeat

Depending on assemblers, they may result in 1 or 2 contigs

Layout – can we do more?

In practice, layout step also has to deal with spurious subgraphs, e.g. because of sequencing error



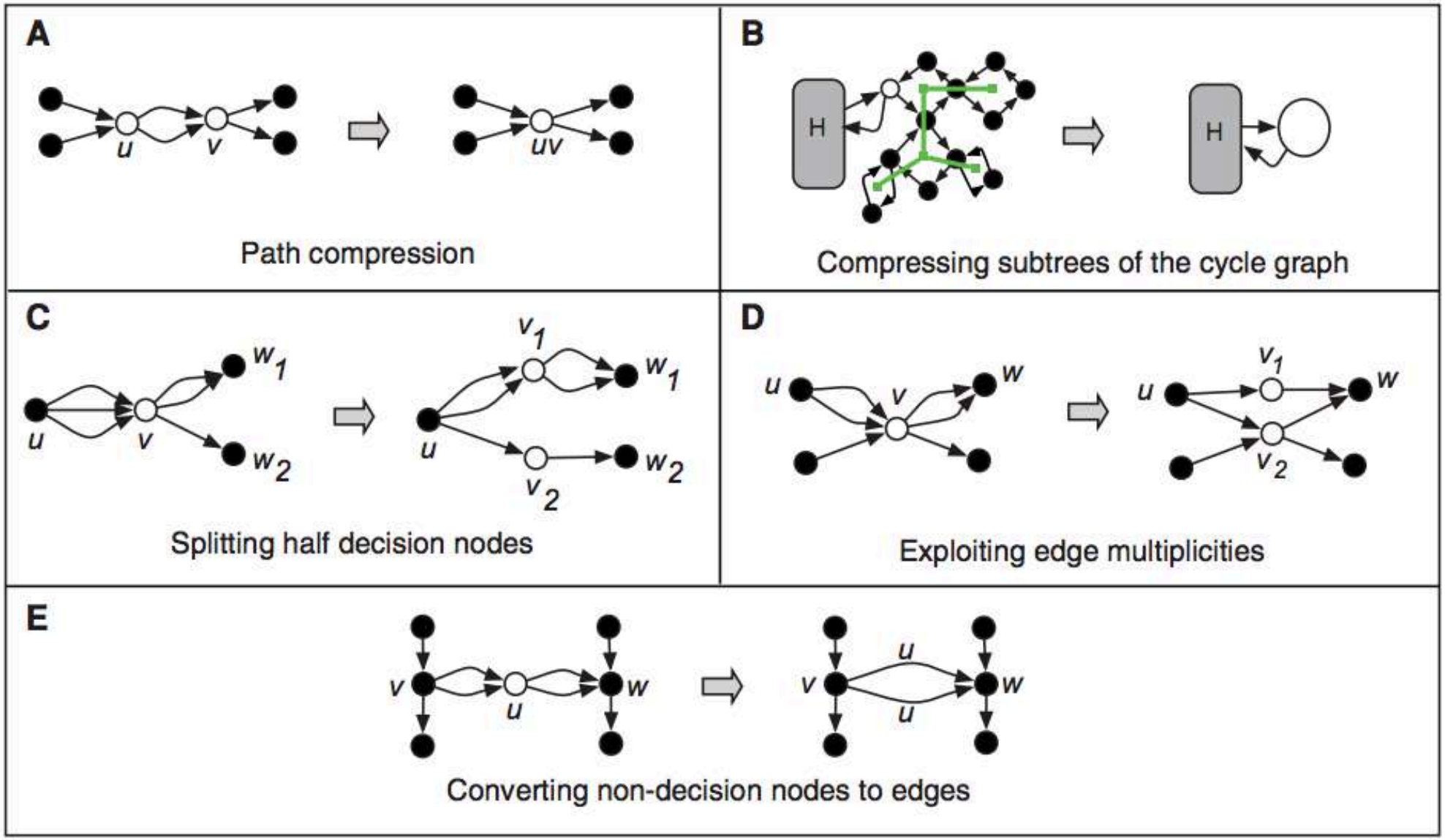


Also often referred as trim the edges

Mismatch could be due to sequencing error or repeat. Since the path through **b** ends abruptly we might conclude it's an error and prune **b**.

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_olc.pdf

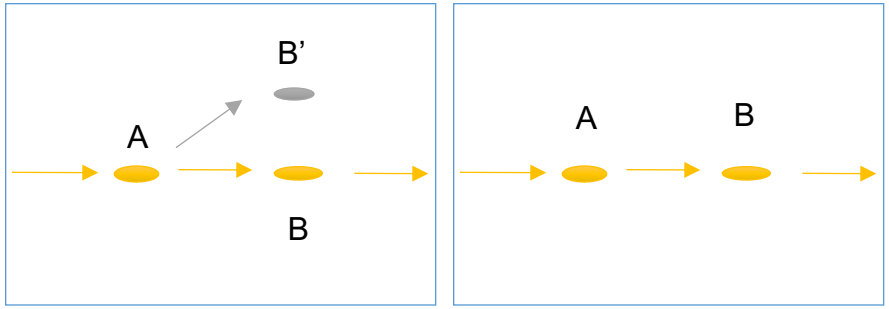
Usefulness of graph transformation



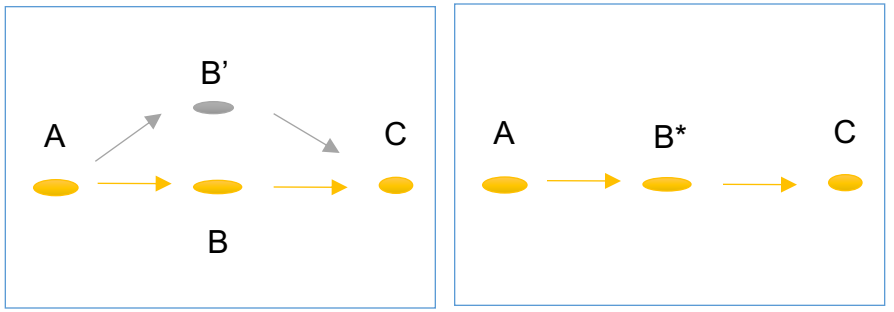
Kingsford *et al.,* (2010)

Error correction in graph

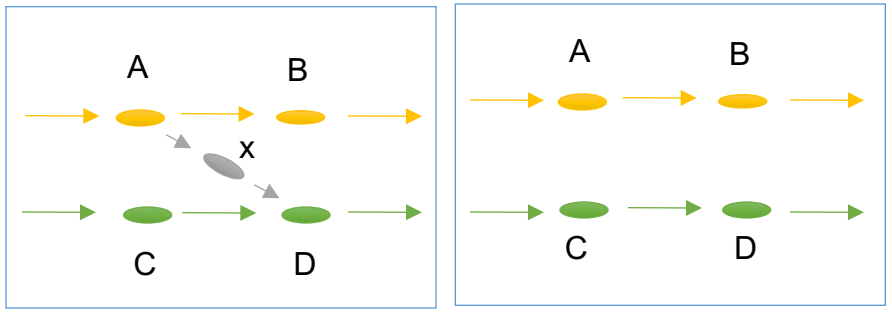
- Errors at end of read
  - Trim off ‘dead-end’ tips



- Errors in middle of read
  - Pop Bubbles



- Chimeric Edges
  - Clip short, low coverage nodes



Credit: Michael Schatz

Consensus

At each position, ask: what nucleotide (and/or gap) is here?

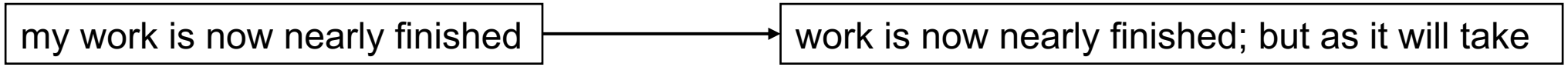
Complications: (a) sequencing error, (b) ploidy

Say the true genotype is AG, but we have a high sequencing error rate and only about 6 reads covering the position.

OLC Assemblers

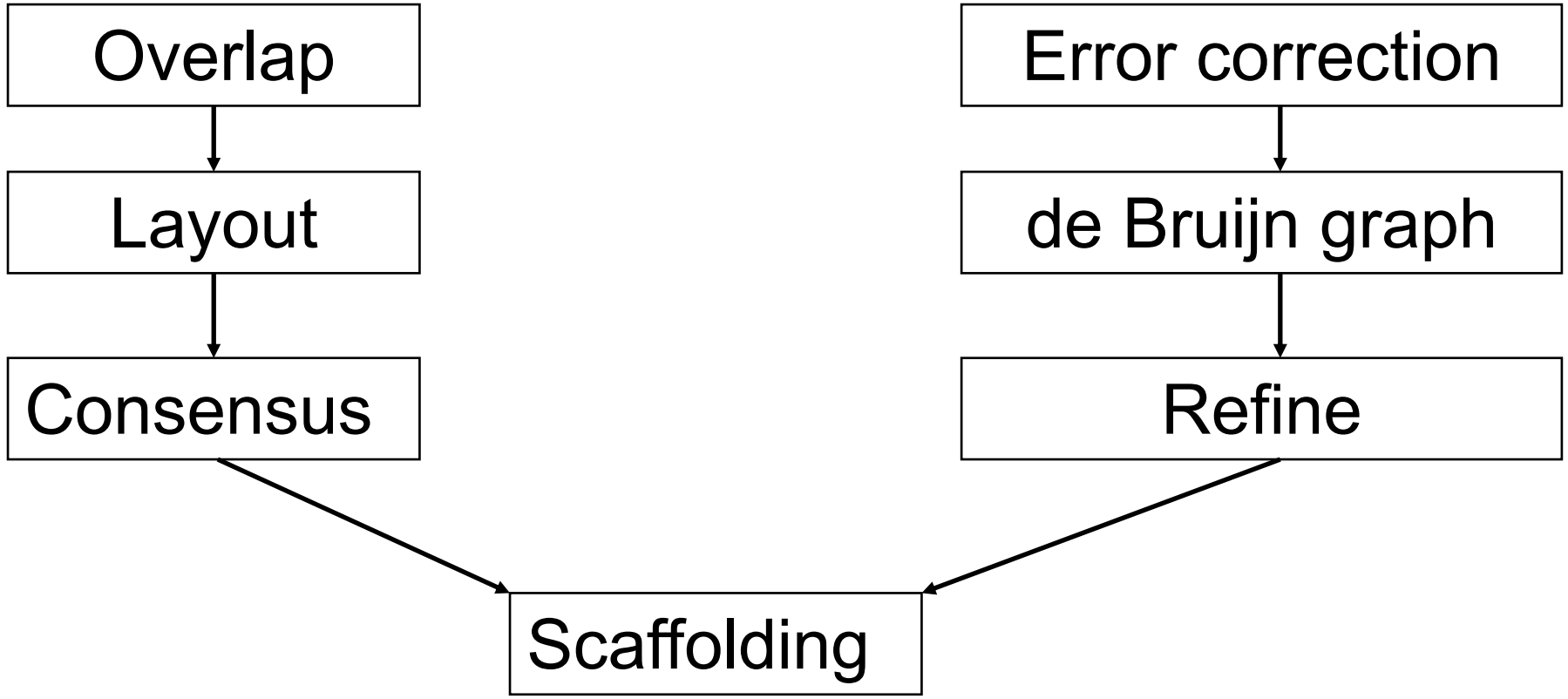
- Mostly used in the Sanger sequencing era
  - Celera, Phusion, PCAP, Arachne

- Disadvantages of OLC
  - Computing overlaps is slow
  - 5 billion reads -> takes **400 years** to compute overlaps if 1 million overlap per second
  - Overlap graph is big and complicated
    - One node per read
    - Number of edges grows superlinearly with number of reads



- When 2nd generation dataset first arrived
  - Millions and millions of reads
  - Short read length – difficult to build sufficient overlap

OLC and DBG assemblers



k-mer

"*k*-mer" is a substring of length *k*

S: GGCGATTCATCG

A 4-mer of *S*: ATTC

All 3-mers of *S*:

GGC
GCG
CGA
GAT
ATT
TTC
TCA
CAT
ATC
TCG

mer: from Greek meaning "part"

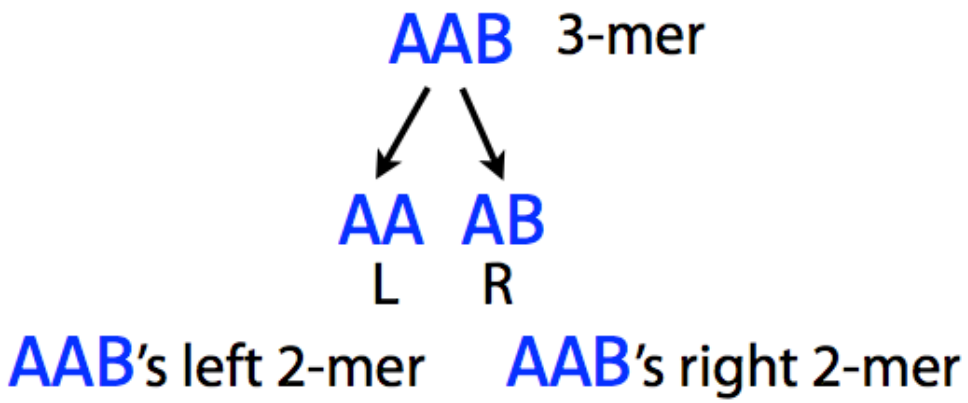
http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

De Bruijn graph

We start with a collection of reads of **3bp** from the reference genome **AAABBBA**

AAA, AAB, ABB, BBB, BBA

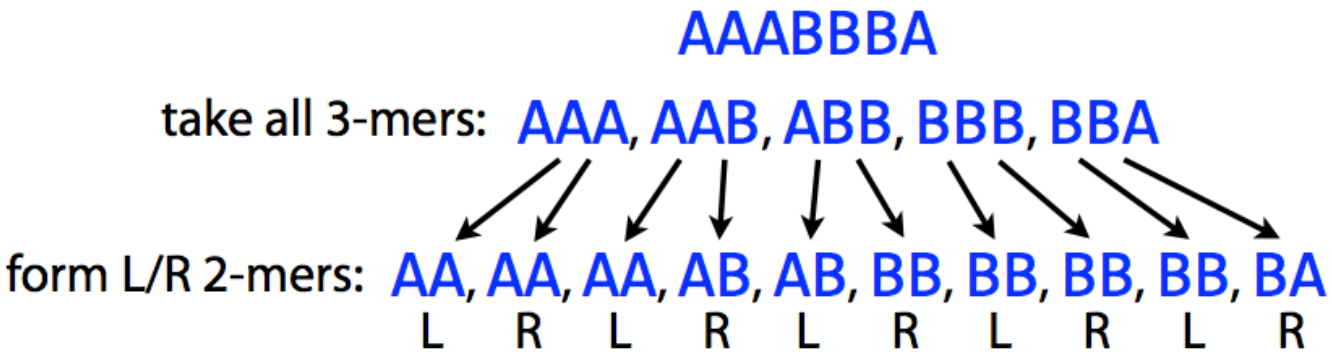
AAB is a *k*-mer (*k* = 3). AA is its *left* *k*-1-mer, and AB is its right *k*-1-mer.



http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

De Bruijn graph

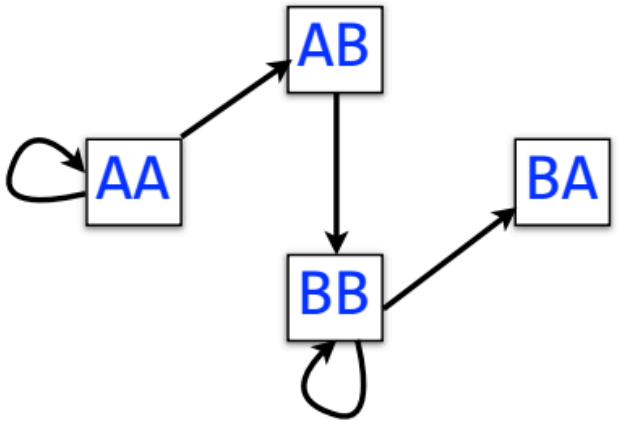
Take each length-3 input string and split it into two overlapping substrings of length 2. Call these the *left* and *right 2-mers*.



From these 2-mers, only
AA, AB, BA, BA are present
(they will be nodes)

Let 2-mers be nodes in a new graph. Draw a directed edge from each left 2-mer to corresponding right 2-mer:

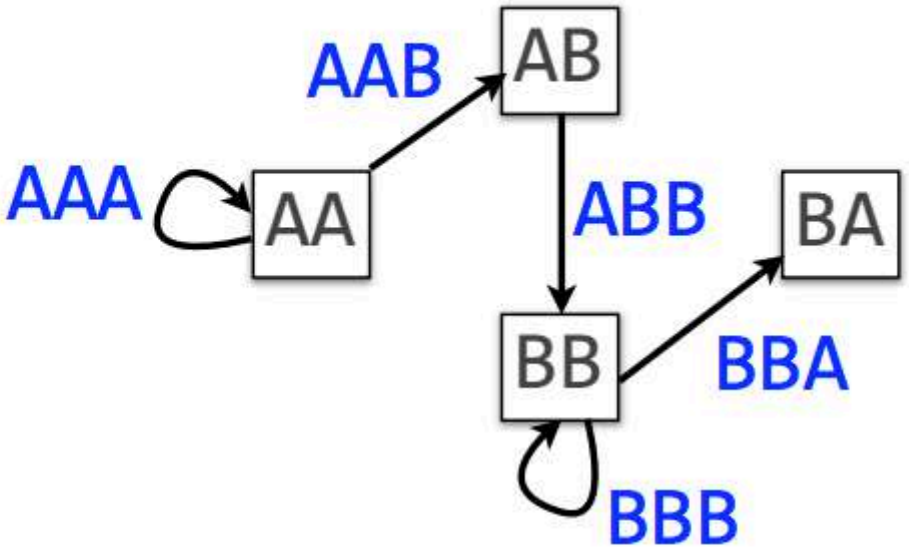
So AAB will be AA → AB



Each *edge* in this graph corresponds to a length-3 input string

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

De Bruijn graph



How do we get contigs from the graph?

Intuitively we walk and visited all edges and node of the graph, but how?

An edge corresponds to an overlap (of length *k*-2) between two *k*-1 mers.
More precisely, it corresponds to a *k*-mer from the input.

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

De Bruijn graph is a directed multigraph

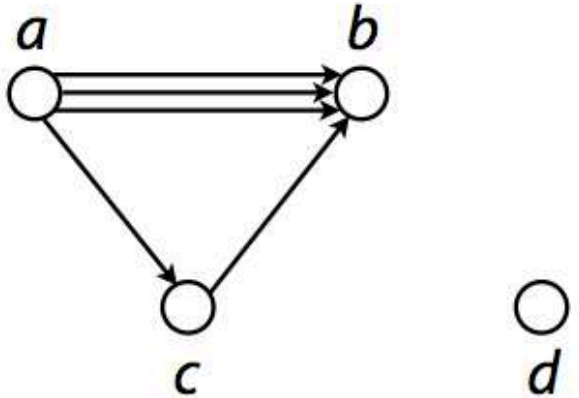
Directed **multigraph** $G(V, E)$ consists of set of *vertices*, V and **multiset** of *directed edges*, E

Otherwise, like a directed graph

Node's *indegree* = # incoming edges

Node's *outdegree* = # outgoing edges

De Bruijn graph is a directed multigraph



$V = \{ a, b, c, d \}$

$E = \{ (a, b), (a, b), (a, b), (a, c), (c, b) \}$

Repeated

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

Eulerian walk definitions and statement

1. A **directed graph** is a *graph* in which each edge has a direction, usually represented as an arrow from a node *v* to a node *w*.

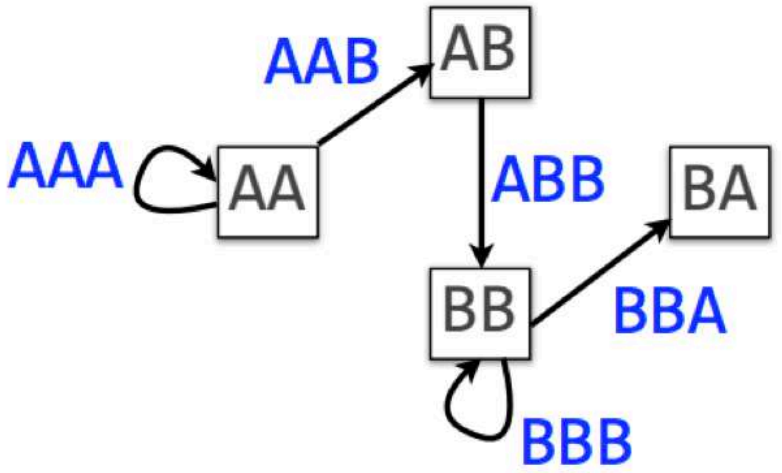
Yes

2. Graph is connected if each node can be reached by some other node.

Yes

3. Node is **balanced** if indegree equals outdegree Node is **semi-balanced** if indegree differs from outdegree by 1

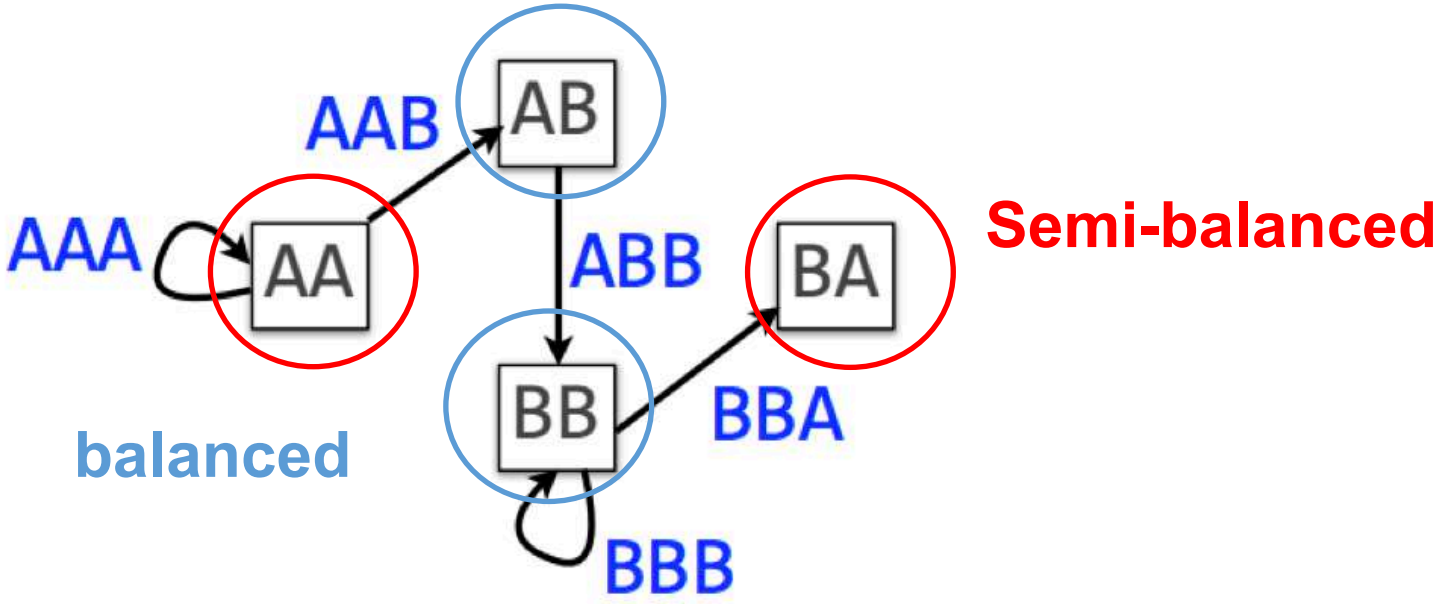
4. A directed, connected graph is Eulerian if and only if it has **at most 2 semi-balanced nodes** and all other nodes are balanced



https://en.wikipedia.org/wiki/Eulerian_path

Eulerian walk definitions and statement

A **directed, connected graph** is Eulerian if and only if **it has at most 2 semi-balanced nodes and all other nodes are balanced**



Is it Eulerian? Yes

Argument 1: AA → AA → AB → BB → BB → BA

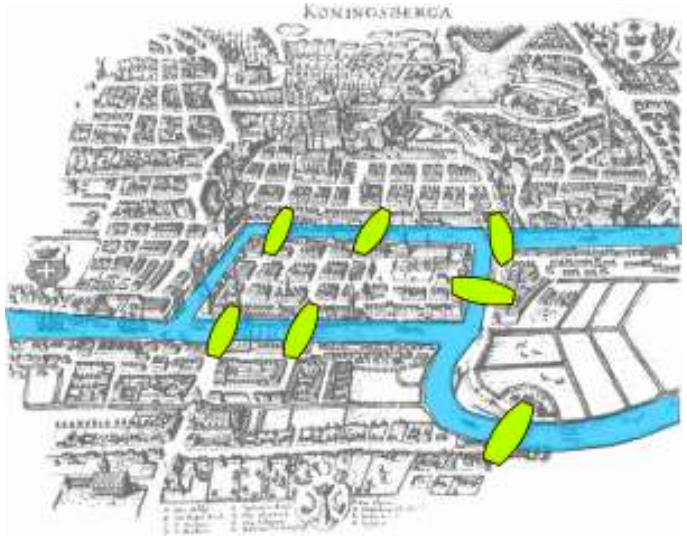
Argument 2: AA and BA are semi-balanced, AB and BB are balanced

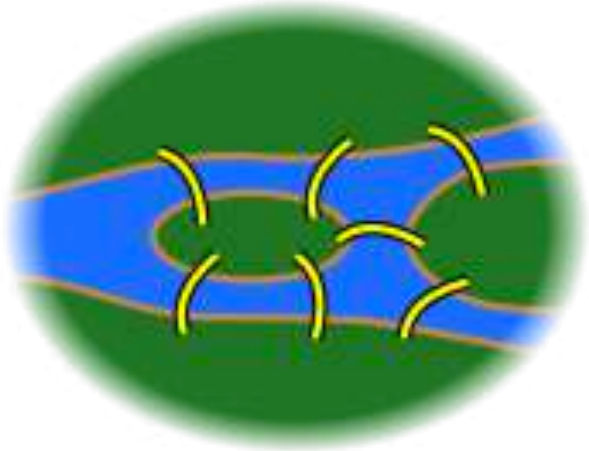
Then we can search for Eulerian path from the graph:
Eulerian walk visits each edge exactly once ****

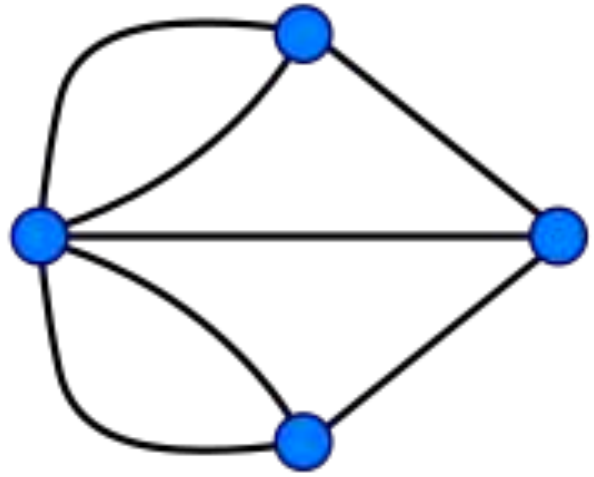
Eulerian walk

In graph theory, an **Eulerian trail** (or **Eulerian path**) is a trail in a graph which visits every edge exactly once. Similarly, an **Eulerian circuit** or **Eulerian cycle** is an Eulerian trail which starts and ends on the same vertex. They were first discussed by Leonhard Euler while solving the famous Seven Bridges of Königsberg problem in 1736. Mathematically the problem can be stated like this:

Given the graph in the image, is it possible to construct a path (or a cycle, i.e. a path starting and ending on the same vertex) which visits each edge exactly once?







PS. This graph is
Not Eulerian

https://en.wikipedia.org/wiki/Eulerian_path

De Bruijn graph

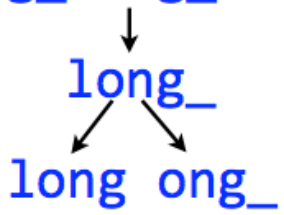
A procedure for making a De Bruijn graph
for a genome

Assume *perfect sequencing* where each length-*k*
substring is sequenced exactly once with no errors

Pick a substring length *k*: 5

Start with an input string: a_long_long_long_time

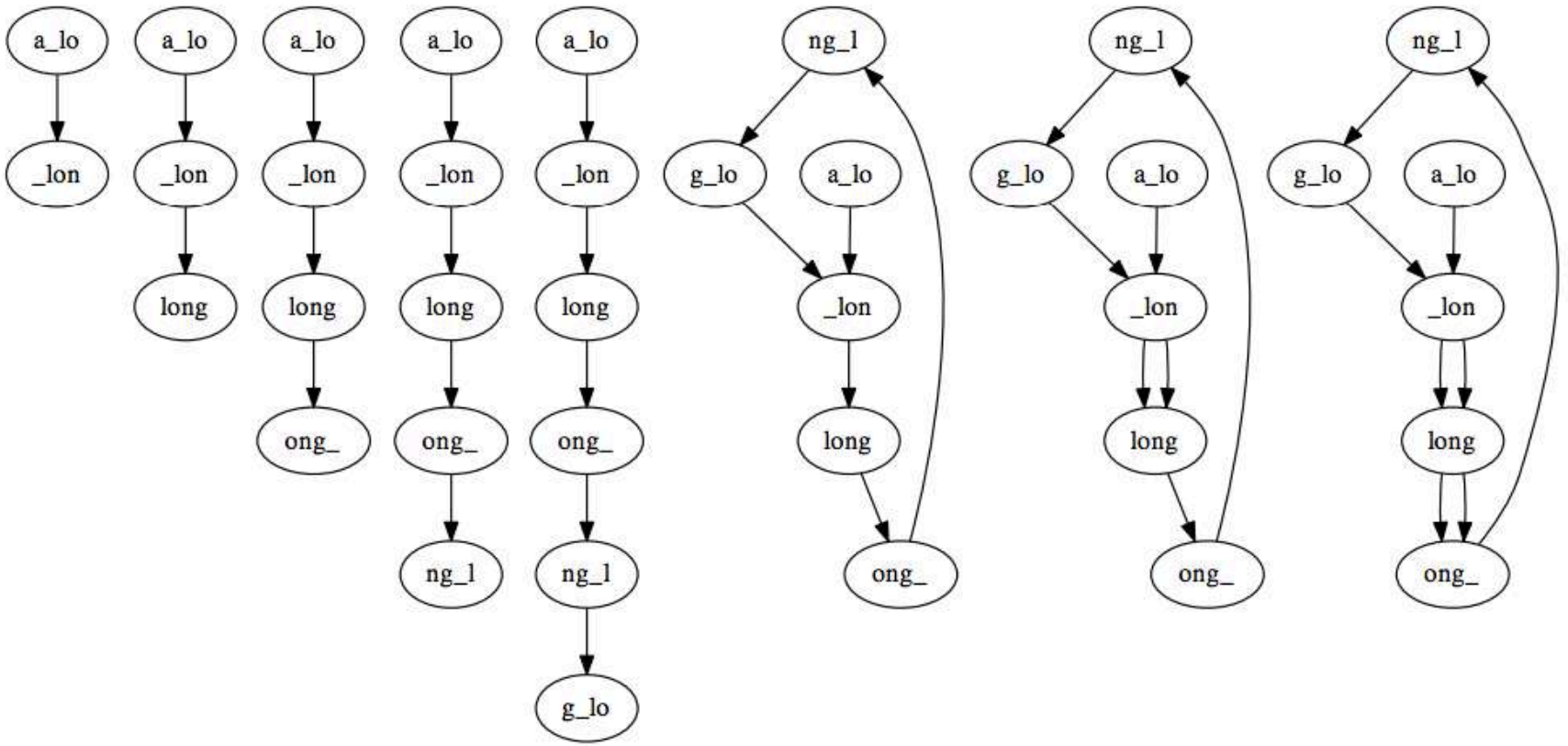
Take each *k* mer and split
into left and right *k*-1 mers



Add k-1 mers as nodes to De Bruijn graph
(if not already there), add edge from left *k*-1
mer to right *k*-1 mer

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

De Bruijn graph

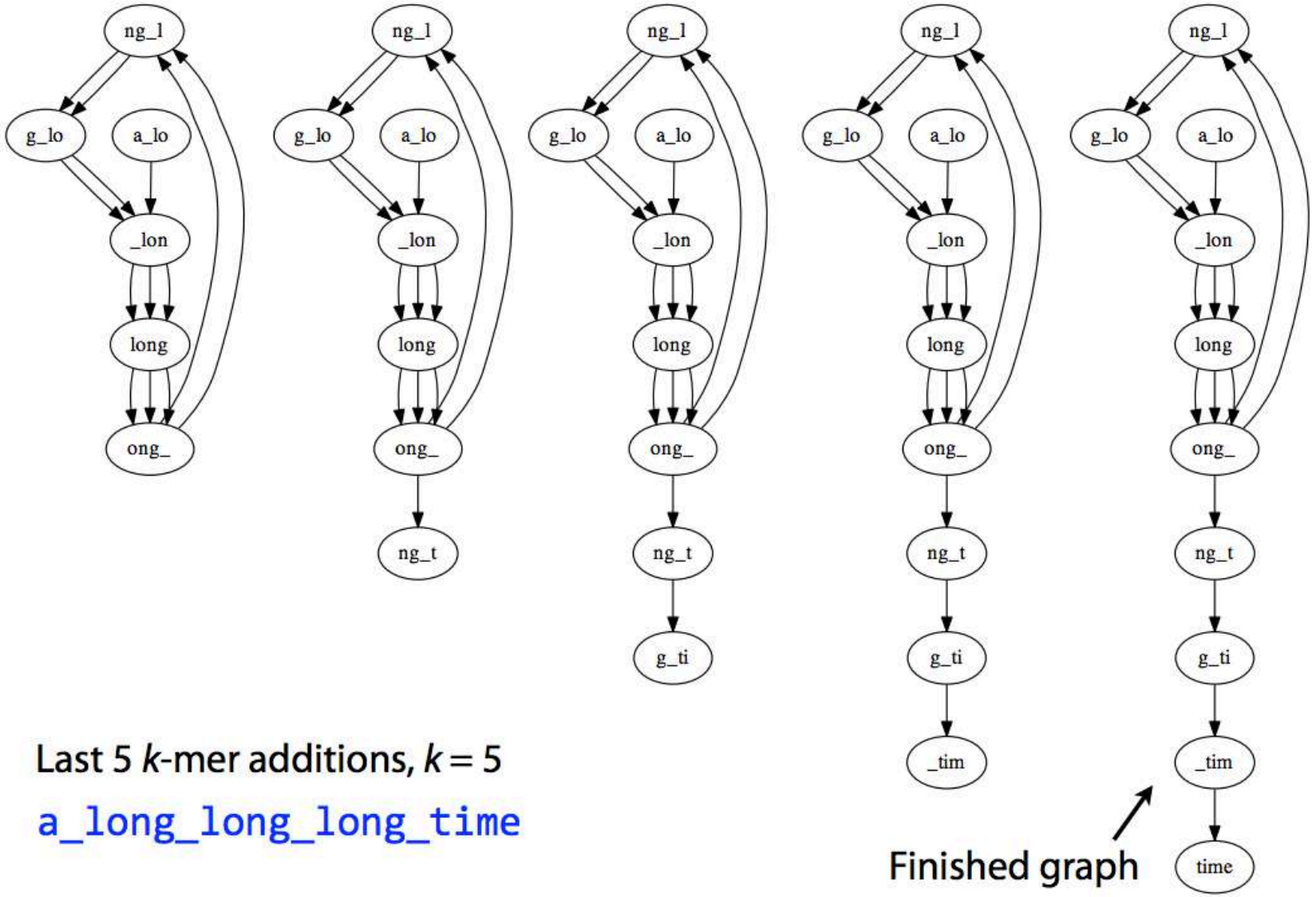


First 8 *k*-mer additions, $k = 5$

```
a_long_long_long_time
```

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

De Bruijn graph



http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

De Bruijn graph

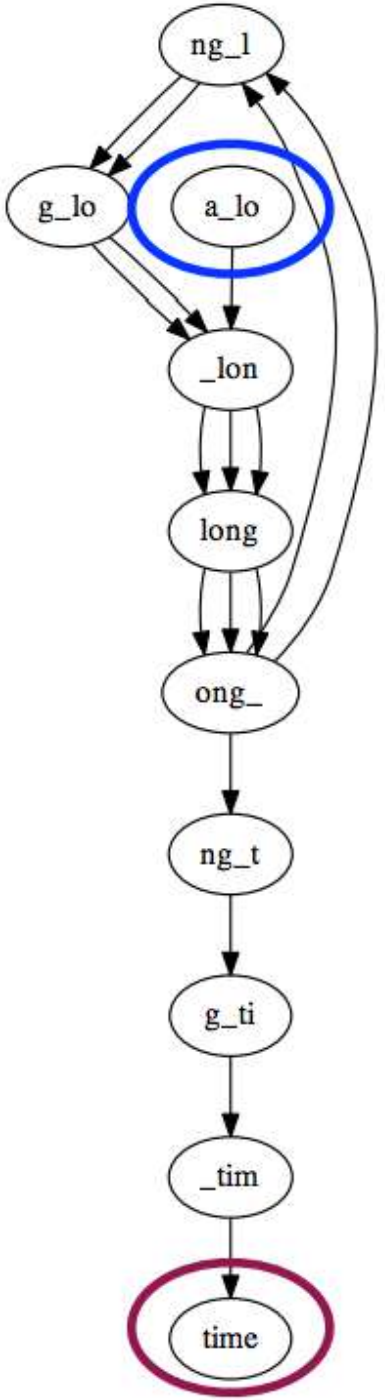
With perfect sequencing, this procedure always yields an Eulerian graph. Why?

Node for *k*-1-mer from left end is semi-balanced with one more outgoing edge than incoming *

Node for *k*-1-mer at right end is semi-balanced with one more incoming than outgoing *

Other nodes are balanced since # times *k*-1-mer occurs as a left *k*-1-mer = # times it occurs as a right *k*-1-mer

* Unless genome is circular

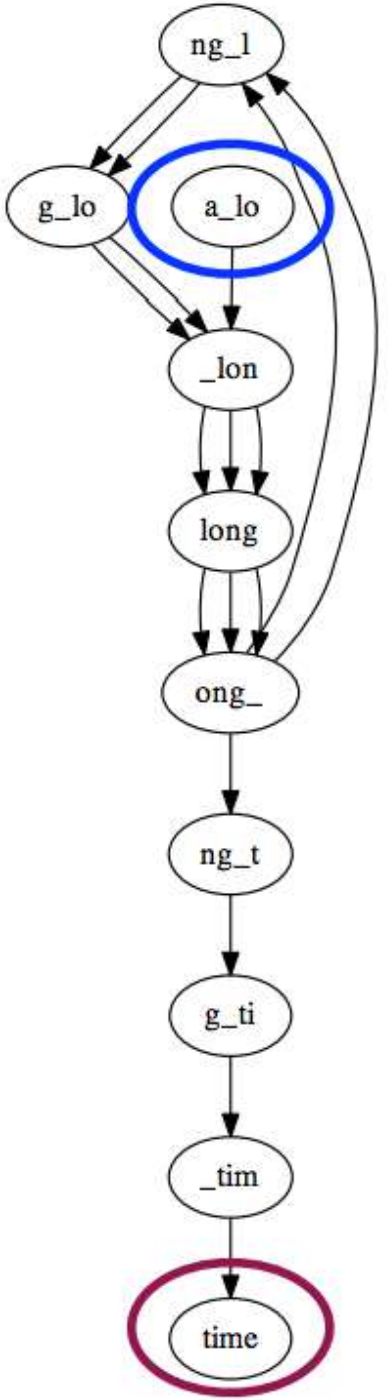


http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

De Bruijn graph with actual data

Assuming perfect sequencing, procedure yields graph with Eulerian walk that can be found efficiently.

We saw cases where Eulerian walk corresponds to the original superstring. Is this always the case?



http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

When k-mer is repeat

No: graph can have multiple Eulerian walks, only one of which corresponds to original superstring

Right: graph for ZABCDABEFABY, $k = 3$

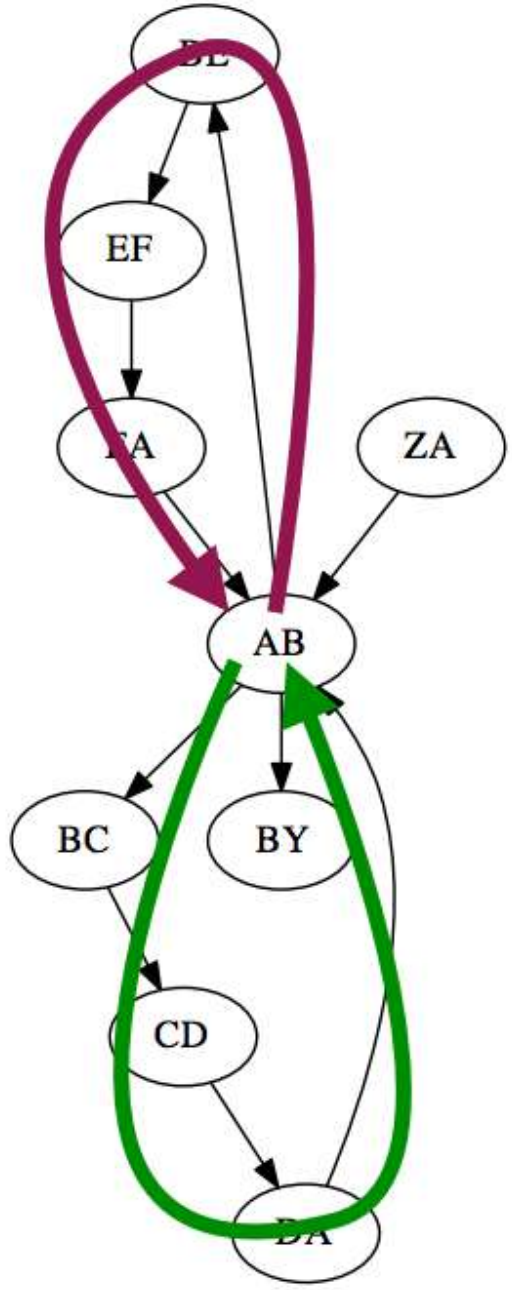
Alternative Eulerian walks:

ZA → AB → BE → EF → FA → AB → BC → CD → DA → AB → BY

ZA → AB → BC → CD → DA → AB → BE → EF → FA → AB → BY

These correspond to two edge-disjoint directed cycles joined by node AB

AB is a repeat: ZABCDABEFABY



http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

When k-mer is repeat (in practice)

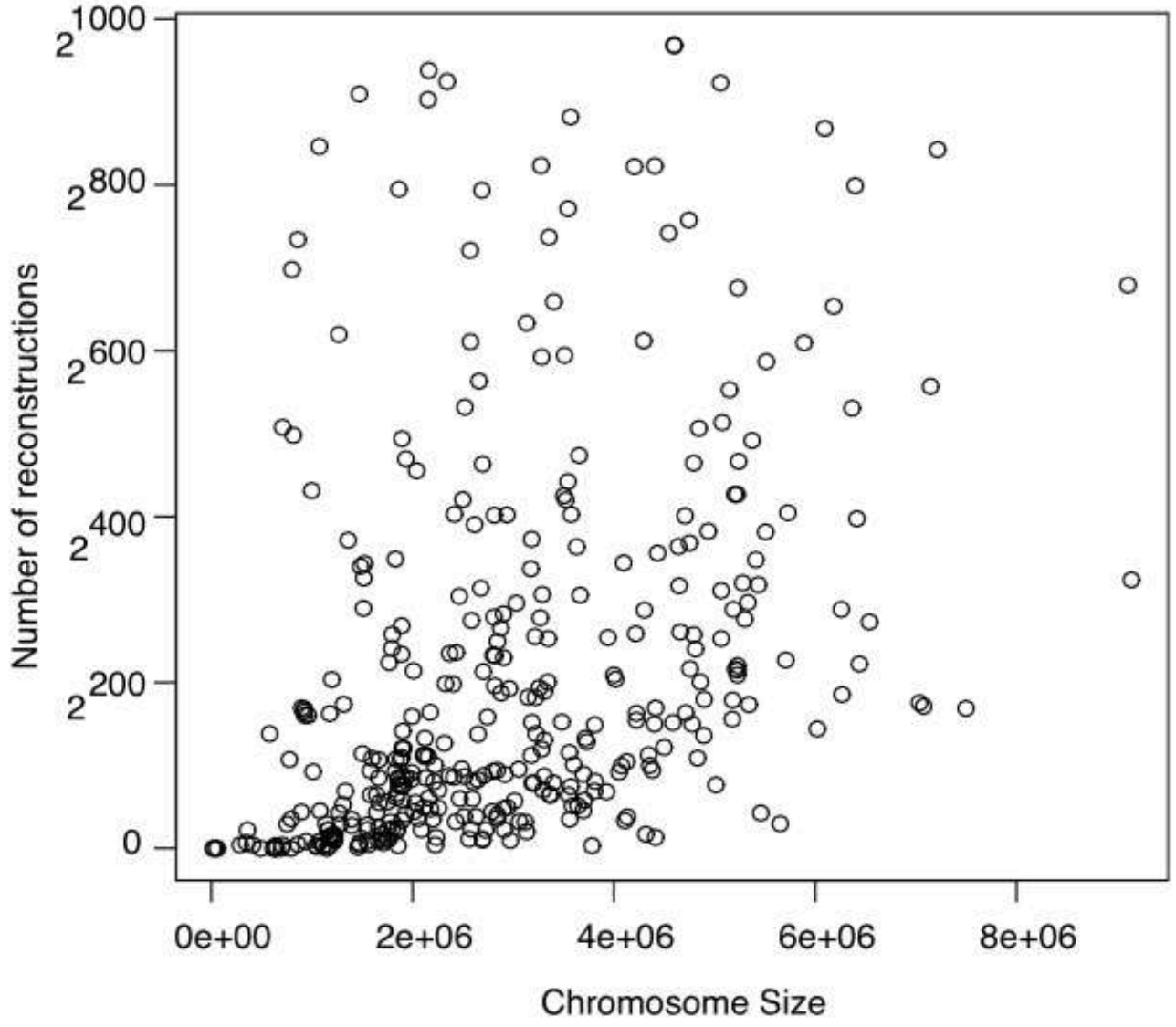


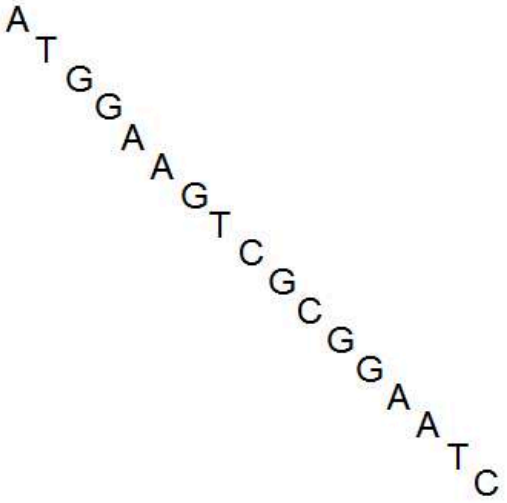
Figure 2 Number of words consistent with genome graphs. The

Kingsford *et al.*, 2010

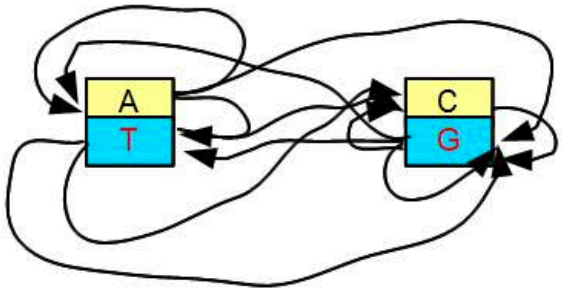
Impact of changing kmer size

sequence ATGGAAGTCGCGGAATC

1mers



de Bruijn graph



Example of a kmer of 1 basically means A,C,T,G are all repeats..

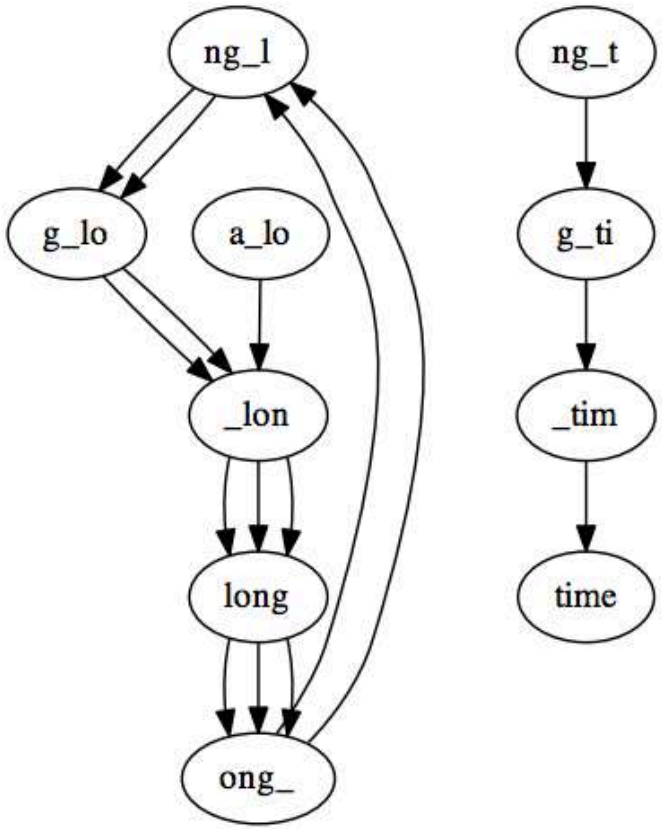
Larger kmer will span more small repeat less than kmer size, but likely to have less overlap

http://www.homolog.us/Tutorials/index.php?p=2.4&s=1

Low coverage = disconnected graph

Gaps in coverage can lead to *disconnected* graph

Graph for `a_long_long_long_time`, $k = 5$ but *omitting* `ong_t` :



Connected components are individually Eulerian, overall graph is not

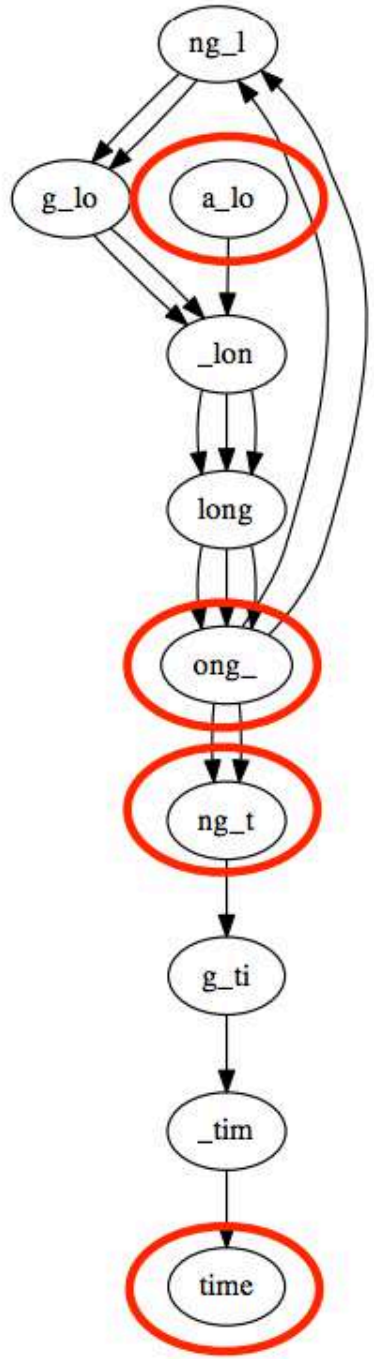
http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

Coverage difference = not Eulerian

Differences in coverage also lead to non-Eulerian graph

Graph for `a_long_long_long_time`, $k = 5$ but with *extra copy* of `ong_t` :

Graph has 4 semi-balanced nodes, isn't Eulerian



http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

De Bruijn graph

Gaining assembly as Eulerian walk is appealing, not many practical cases impede this:

- Uneven coverage, sequencing errors, make graph non-Eulerian
- Repeats produces many possible walks

But there is one major advantage of De Bruijn graph over OLC

- Computationally efficient

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

Efficiency

Assume you have 5 billion reads to assemble

- Not uncommon nowadays in some plant species

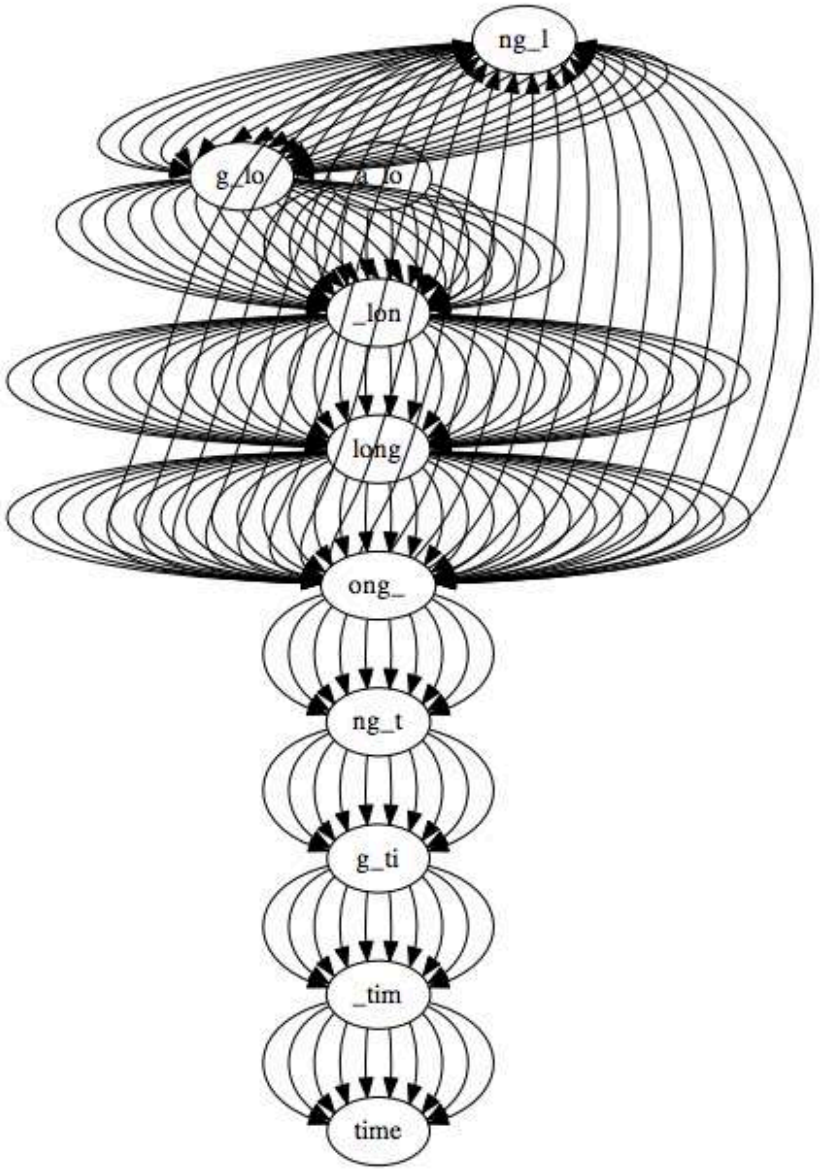
OLC: 12.5 quadrillion overlaps to compute first
Even 1 million overlap per sec will equate to **400 years**

DBG: depends on genome size

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

Size of De Bruijn graph depends on genome size

Usually ~50X paired end reads in coverage



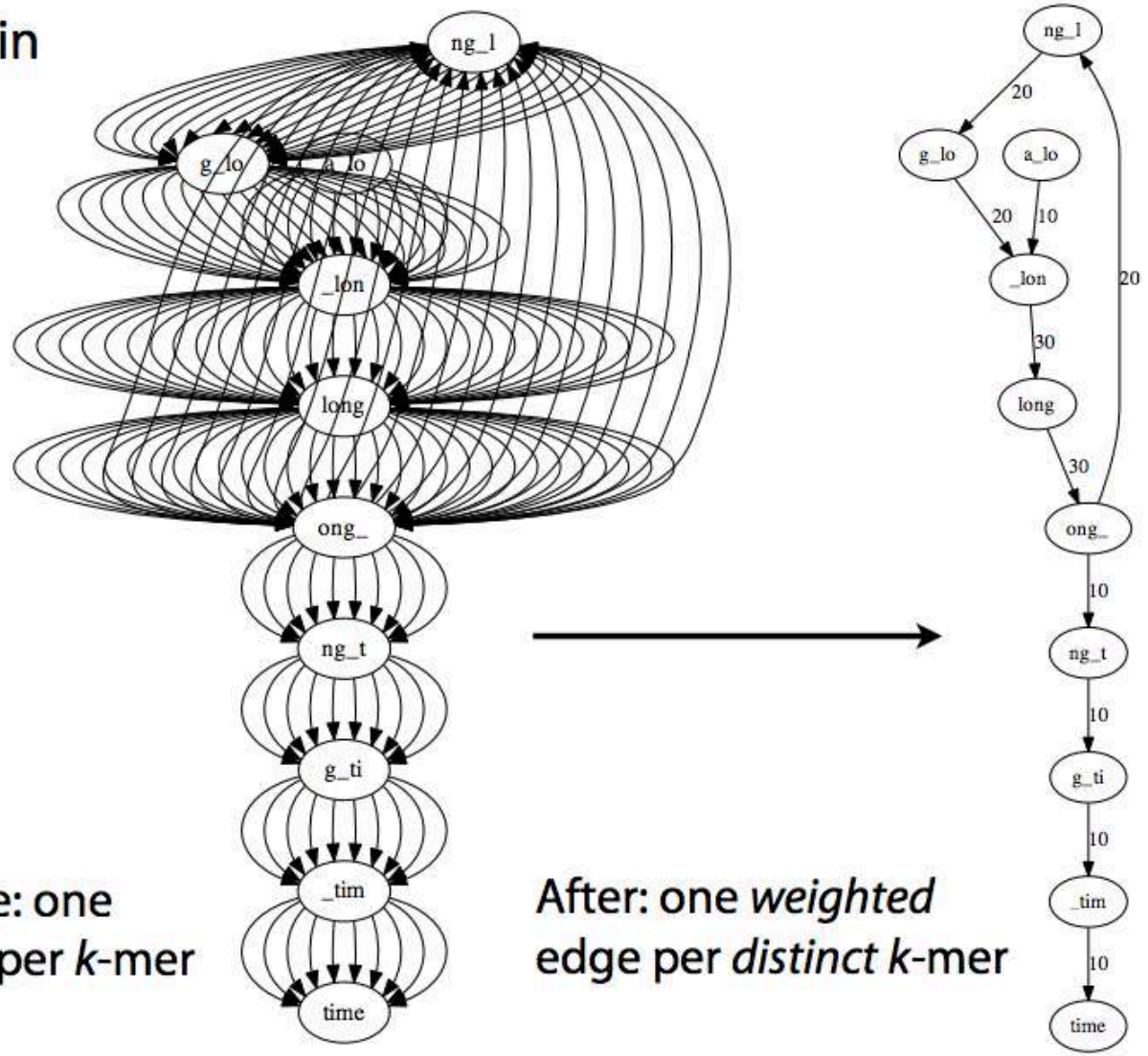
http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

Size of De Bruijn graph depends on genome size

Same edge might appear in dozens of copies; let's use edge *weights* instead

Weight = # times *k*-mer occurs

Using weights, there's one *weighted* edge for each *distinct k*-mer

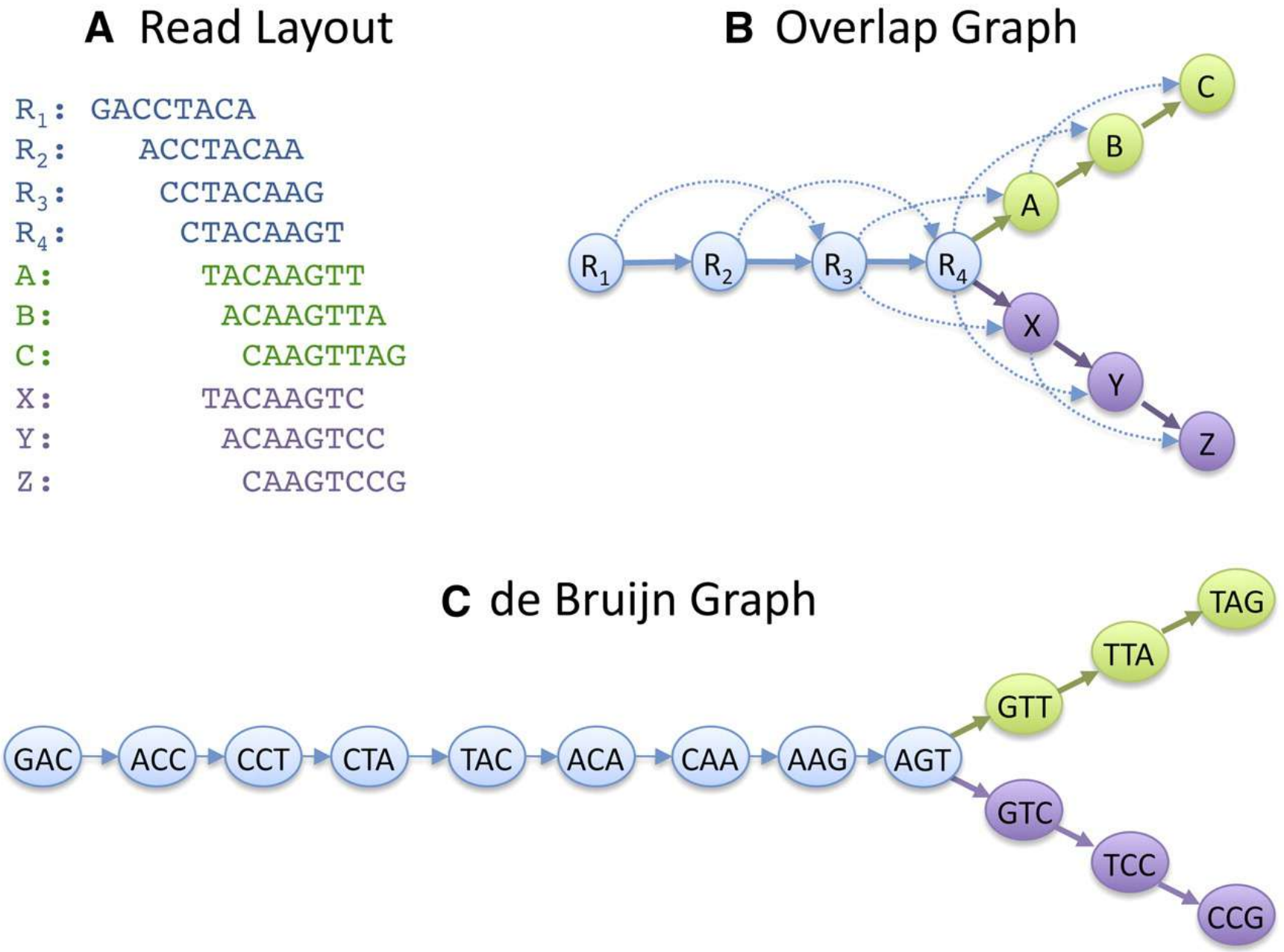


Before: one edge per *k*-mer

After: one *weighted* edge per *distinct k*-mer

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

Summary



Schatz M C et al. Genome Res. 2010;20:1165-1173

Summary

Advantage of DBG:

Time to build based on Genome size (G) or total length of reads (N)

For OLC: time to build overlap graph is based on number of reads

But:

DBG not flexible: only overlap of **fixed length** (=kmer)

can't solve repeat with repeat > kmer

Read information is lost: All reads are split into kmers.

(A lot of work on later DBG assemblers are put in this)

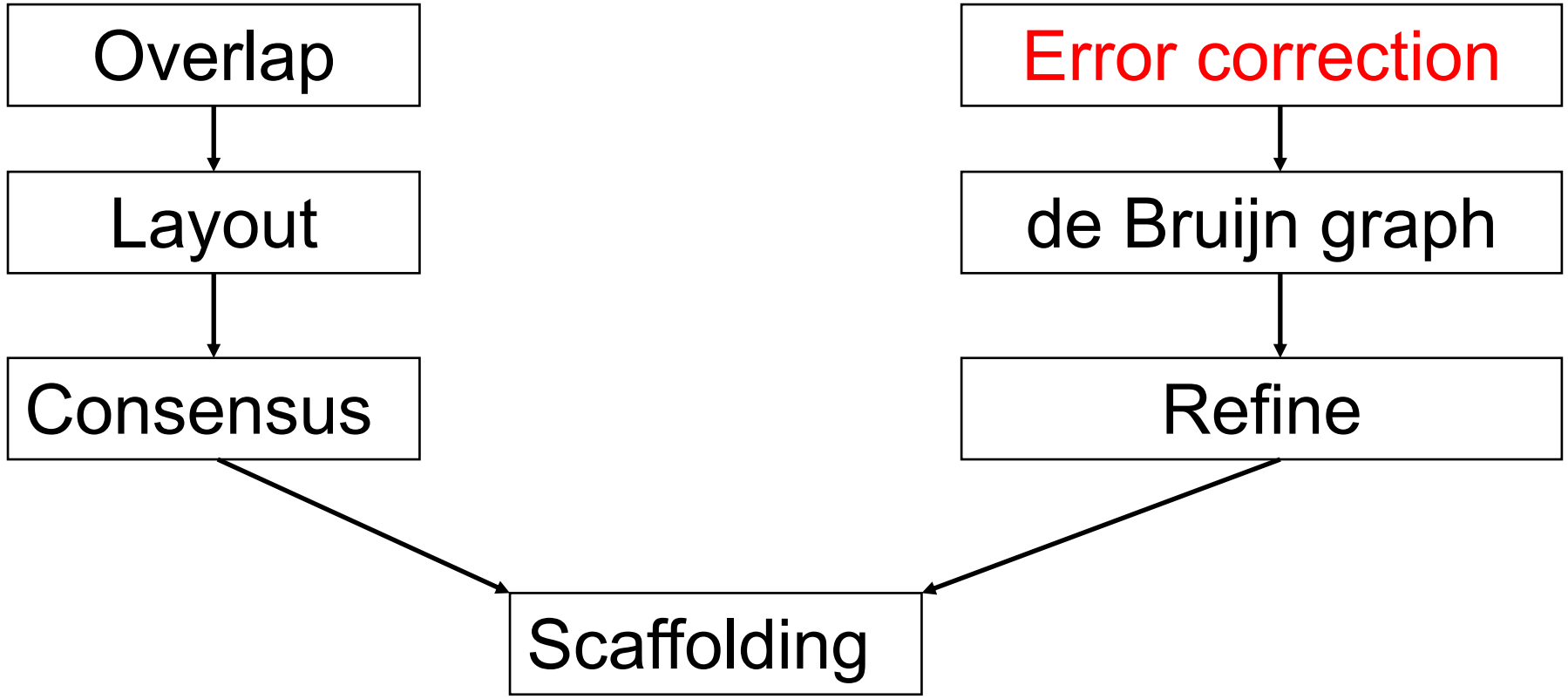
Tradeoff between DBG and OLC needed

Some improve over existing approach: Spades

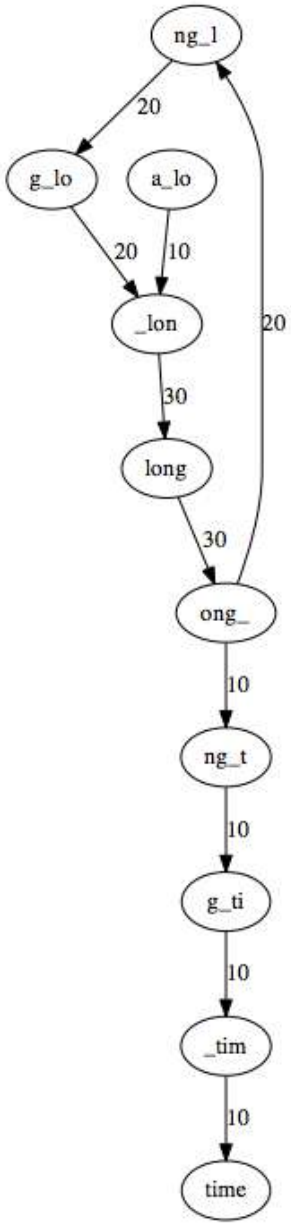
Some combine both: Masurca

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

OLC and DBG assemblers



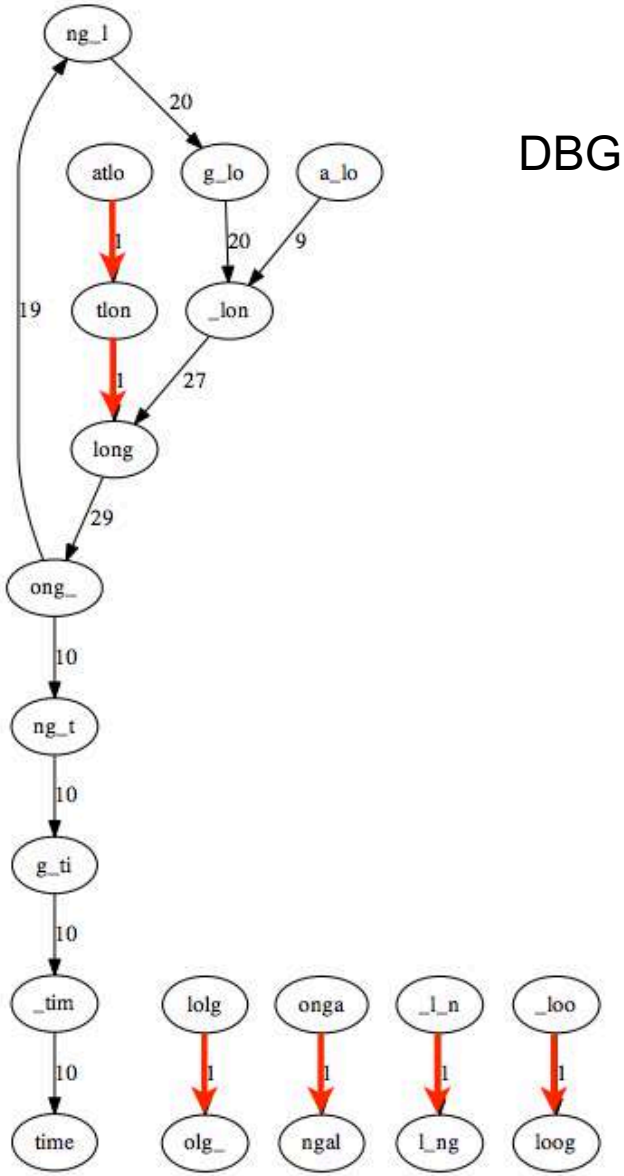
Error in graphs



DBG from perfect reads (10X)

As you can see all weighted **edges** have high coverage

Some higher than the other; this is obviously repeat



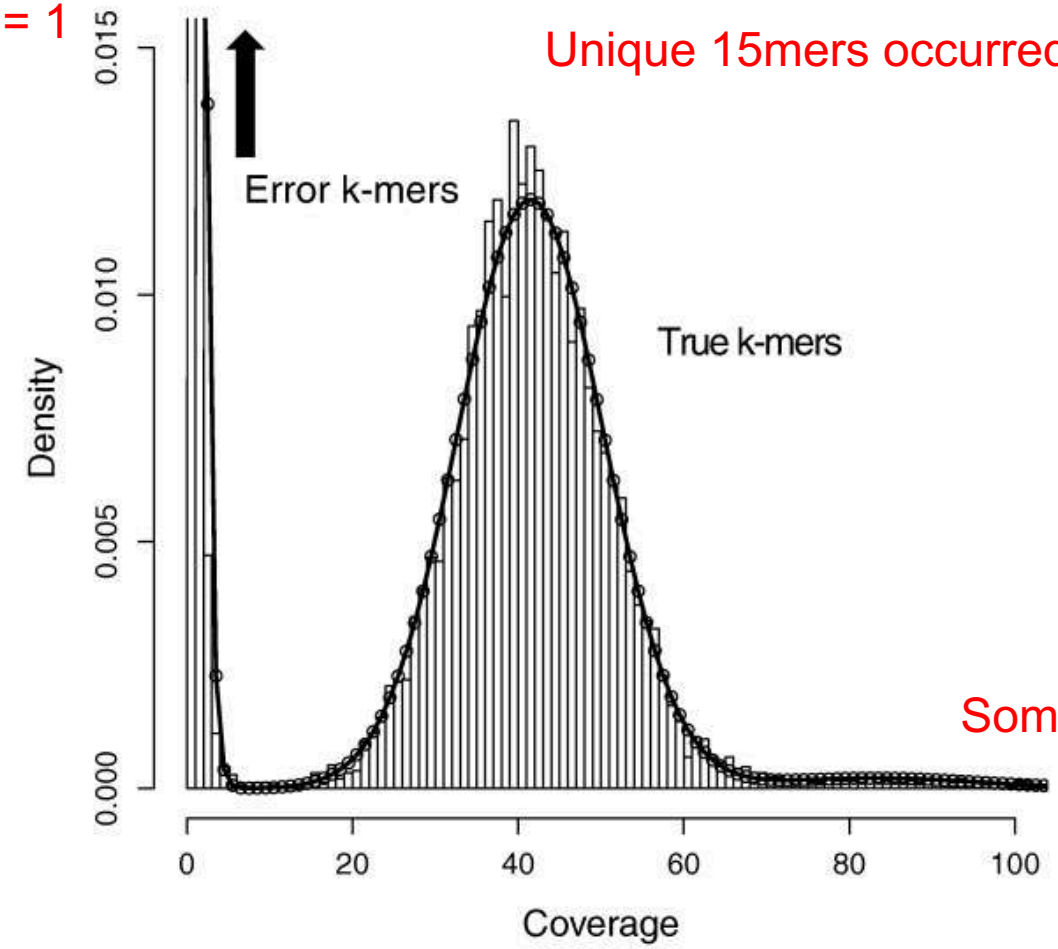
DBG from reads with some errors

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/error_correction.pdf

Kmer coverage

Indeed most erroneous kmer has kmer = 1

Unique 15mers occurred exactly ~41X



Some 15mer occurred at >80X ; repeat

Variance = 77X
Almost twice of mean suggest it's Poisson distribution

Kelley *et al.*, (2010)

Choosing the right kmer

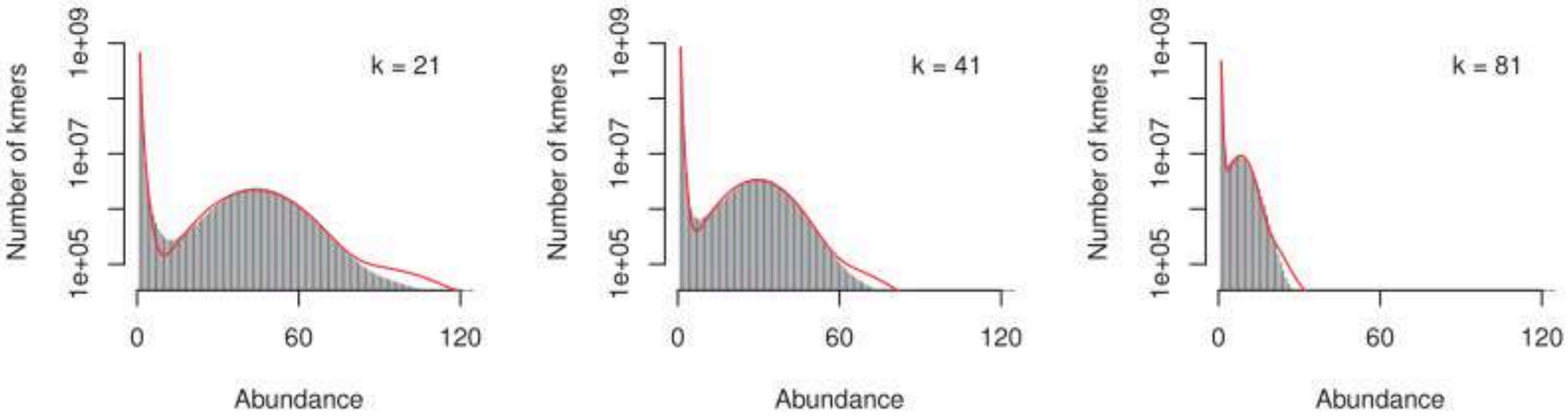


Fig. 2. The abundance histograms for *chr14* with *k* values of 21, 41 and 81 (on a y log scale). Each plot also shows a curve corresponding to the optimized statistical model (haploid)

Chikhi and Medvedev *et al.,* (2014)

Error correction: rationale

Idea: errors tend to turn frequent *k*-mers to infrequent *k*-mers, so corrections should do the reverse

Say we have a collection of reads where each distinct 8-mer occurs an average of ~10 times, and we have the following read:

Read: GCGTATTACGCGTCTGGCCT (20 nt)

8-mers:

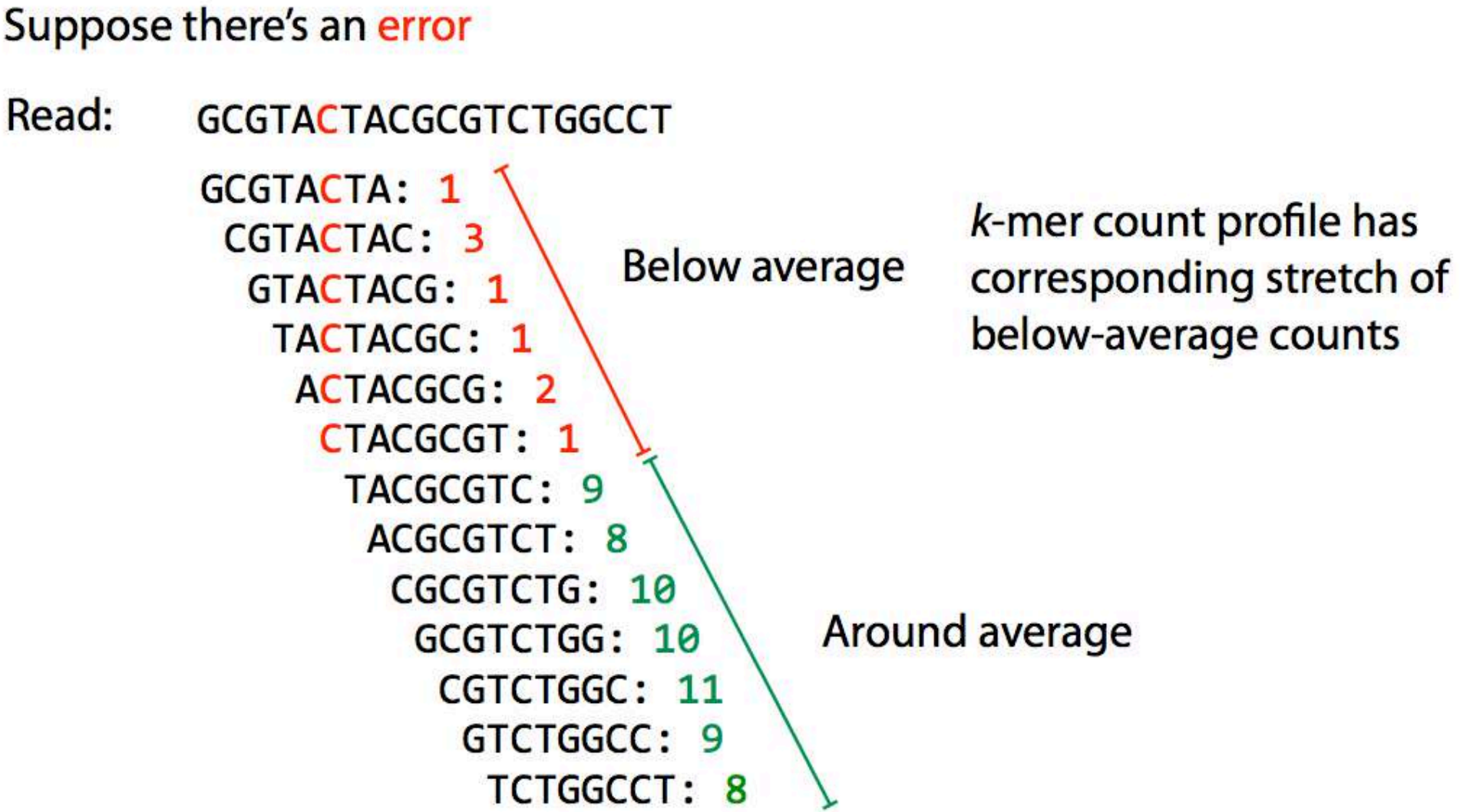
```
GCGTATTA: 8
 CGTATTAC: 8
  GTATTACG: 9
   TATTACGC: 9
    ATTACGCG: 9
     TTACGCGT: 12
      TACGCGTC: 9
       ACGCGTCT: 8
        CGCGTCTG: 10
         GCGTCTGG: 10
          CGTCTGGC: 11
           GTCTGGCC: 9
            TCTGGCCT: 8
```

times each 8-mer occurs in the dataset. "*k*-mer count profile"

All 8-mer counts are around the average, suggesting read is error-free

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/error_correction.pdf

Error correction: rationale

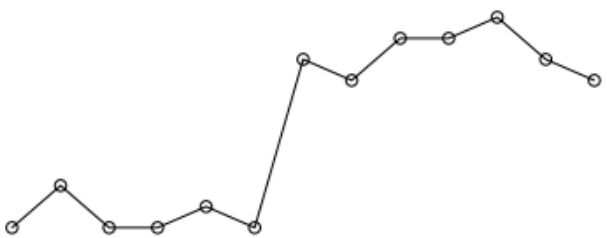


http://www.cs.jhu.edu/~langmea/resources/lecture_notes/error_correction.pdf

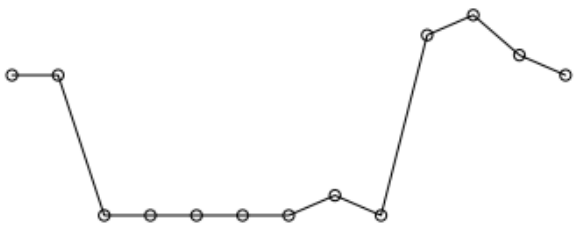
Error correction: rationale

k-mer count profiles when errors are in different parts of the read:

```
GCGTACTACGCGTCTGGCCT
GCGTACTA: 1
 CGTACTAC: 3
  GTACTACG: 1
   TACTACGC: 1
    ACTACGCG: 2
     CTACGCGT: 1
      TACGCGTC: 9
       ACGCGTCT: 8
        CGCGTCTG: 10
         GCGTCTGG: 10
          CGTCTGGC: 11
           GTCTGGCC: 9
            TCTGGCCT: 8
```



```
GCGTATTACACGTCTGGCCT
GCGTATTA: 8
 CGTATTAC: 8
  GTATTACA: 1
   TATTACAC: 1
    ATTACACG: 1
     TTACACGT: 1
      TACACGTC: 1
       ACACGTCT: 2
        CACGTCTG: 1
         GCGTCTGG: 10
          CGTCTGGC: 11
           GTCTGGCC: 9
            TCTGGCCT: 8
```



```
GCGTATTACGCGTCTGGTCT
GCGTATTA: 8
 CGTATTAC: 8
  GTATTACG: 9
   TATTACGC: 9
    ATTACGCG: 9
     TTACGCGT: 12
      TACGCGTC: 9
       ACGCGTCT: 8
        CGCGTCTG: 10
         GCGTCTGG: 10
          CGTCTGGT: 1
           GTCTGGTC: 2
            TCTGGTCT: 1
```

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/error_correction.pdf

Localize error and correct

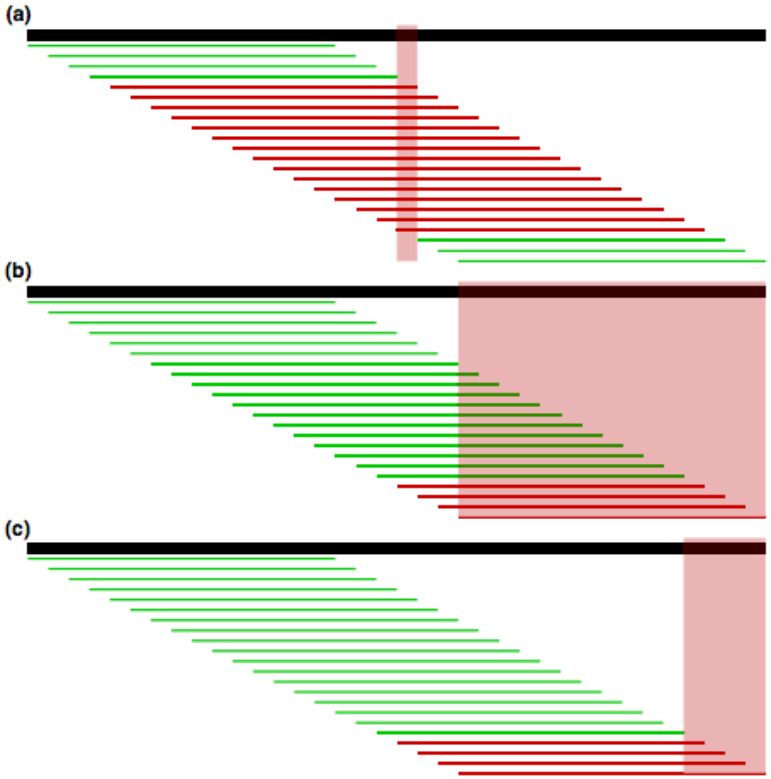
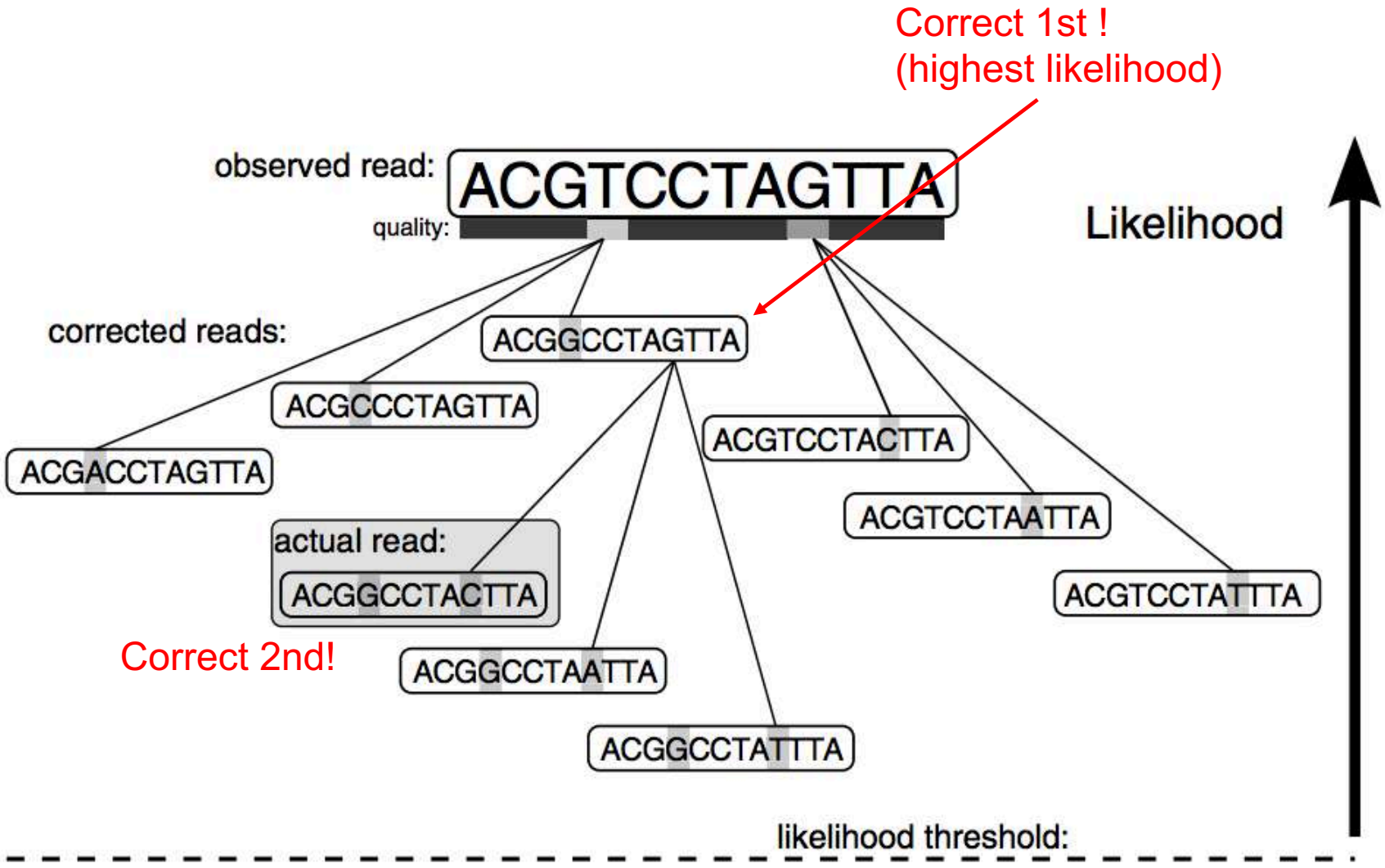


Figure 4 Localize errors. Trusted (green) and untrusted (red) 15-mers are drawn against a 36 bp read. In (a), the intersection of the untrusted *k*-mers localizes the sequencing error to the highlighted column. In (b), the untrusted *k*-mers reach the edge of the read, so we must consider the bases at the edge in addition to the intersection of the untrusted *k*-mers. However, in most cases, we can further localize the error by considering all bases covered by the right-most trusted *k*-mer to be correct and removing them from the error region as shown in (c).



Kelley *et al.*, (2010)

Velvet: first de Bruijn graph assembler

- Cited 8812 times
- Still being used in some metagenomics dataset

Resource

Velvet: Algorithms for de novo short read assembly using de Bruijn graphs

Daniel R. Zerbino and Ewan Birney[1]

EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

We have developed a new set of algorithms, collectively called "Velvet," to manipulate de Bruijn graphs for genomic sequence assembly. A de Bruijn graph is a compact representation based on short words (*k*-mers) that is ideal for high coverage, very short read (25–50 bp) data sets. Applying Velvet to very short reads and paired-ends information only, one can produce contigs of significant length, up to 50-kb N50 length in simulations of prokaryotic data and 3-kb N50 on simulated mammalian BACs. When applied to real Solexa data sets without read pairs, Velvet generated contigs of ~8 kb in a prokaryote and 2 kb in a mammalian BAC, in close agreement with our simulated results without read-pair information. Velvet represents a new approach to assembly that can leverage very short reads in combination with read pairs to produce useful assemblies.

[Supplemental material is available online at www.genome.org. The code for Velvet is freely available, under the GNU Public License, at http://www.ebi.ac.uk/~zerbino/velvet.]

Summary

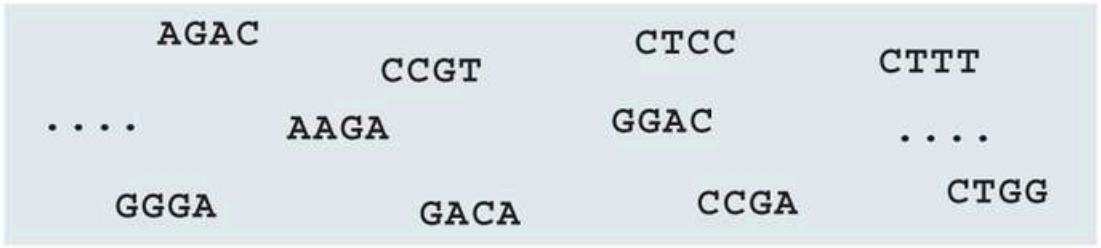
Error correction will **definitely** improve assembly

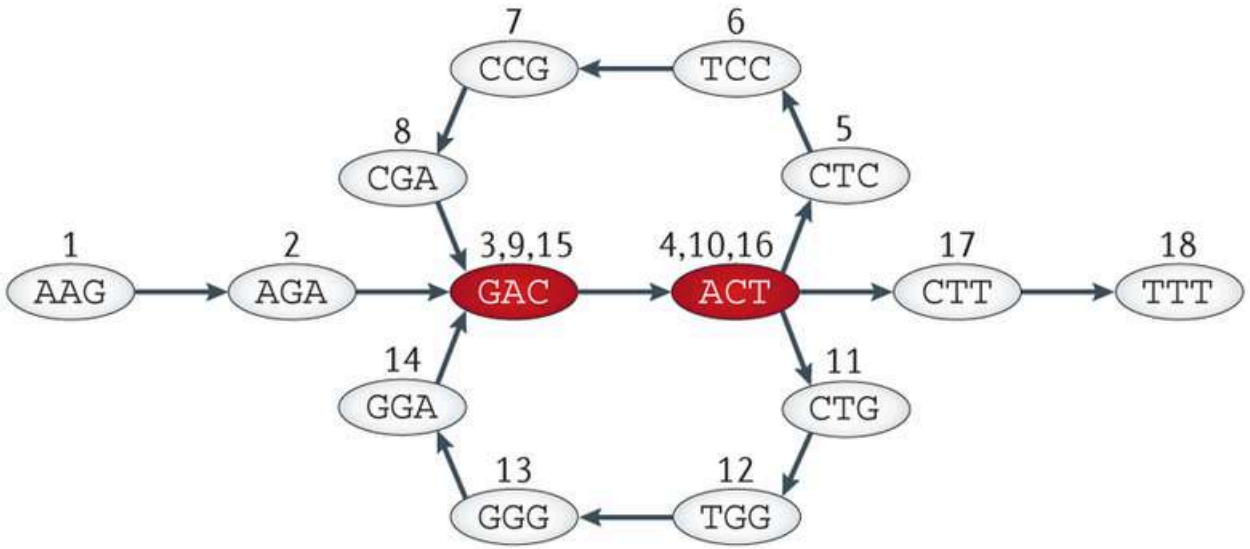
But, for it to work well:

- Sequenced coverage should be high enough
- Choose kmer wisely otherwise we can't distinguish erroneous kmer from frequent kmers

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/assembly_dbg.pdf

Summary I





de Bruijn graph

AAGACTCCGACTGGGACTTT

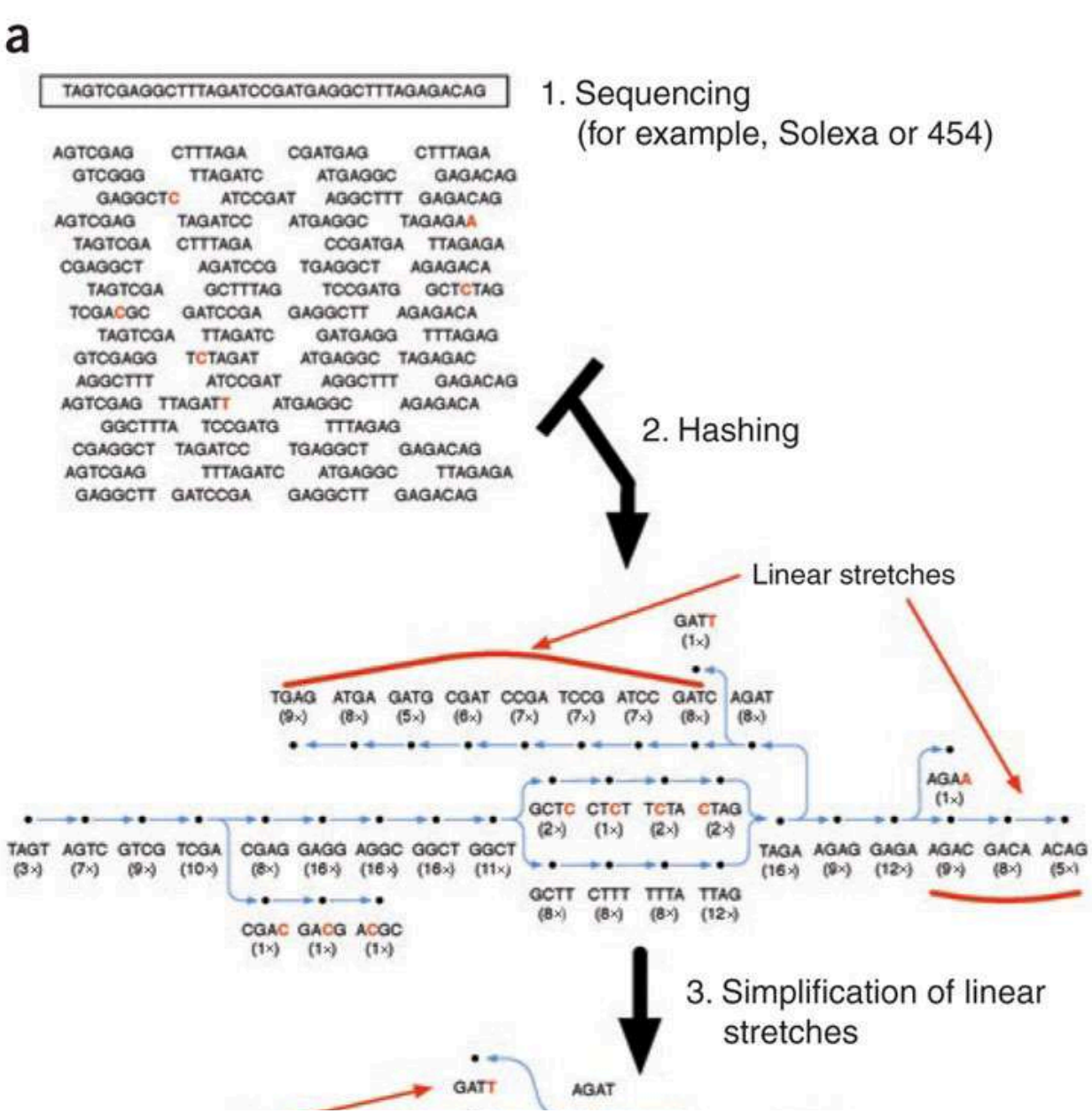
Reads

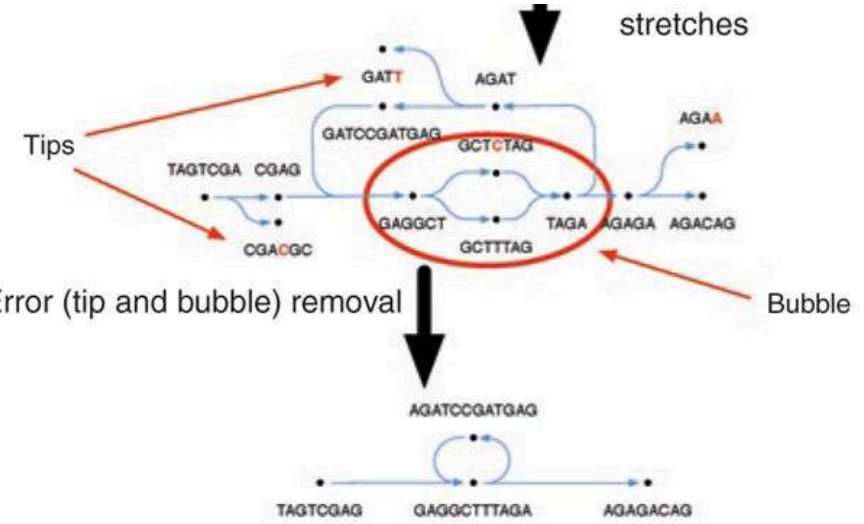
Eulerian path

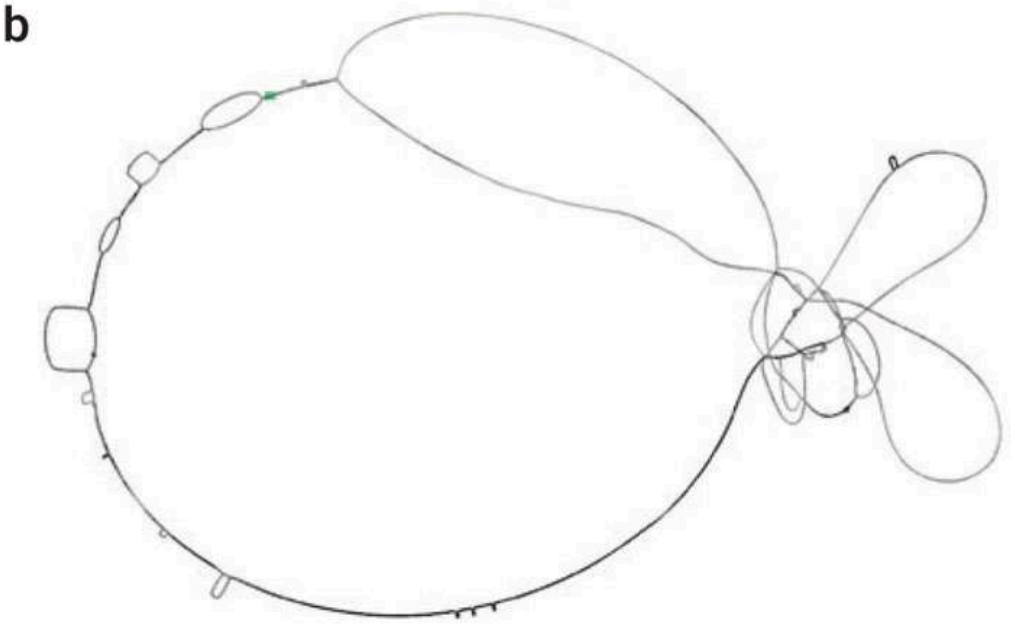
Assembly

http://www.nature.com/nrg/journal/v14/n5/full/nrg3433.html

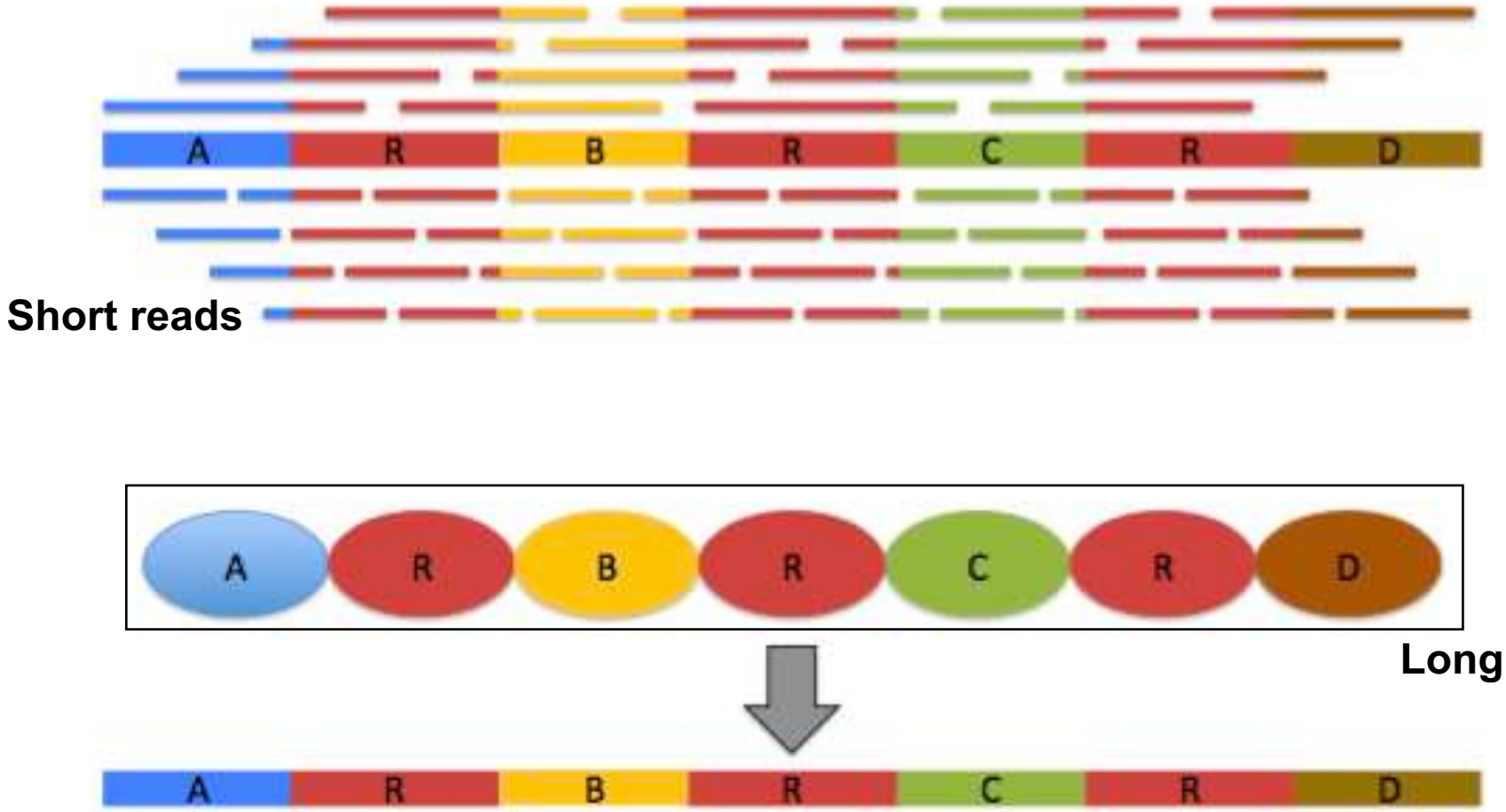
Summary II


a
TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG
1. Sequencing
(for example, Solexa or 454)
2. Hashing
Linear stretches
3. Simplification of linear stretches


stretches
Tips
4. Error (tip and bubble) removal
Bubble
AGATCCGATGAG
TAGTCGAG
GAGGCTTTAGA
AGAGACAG

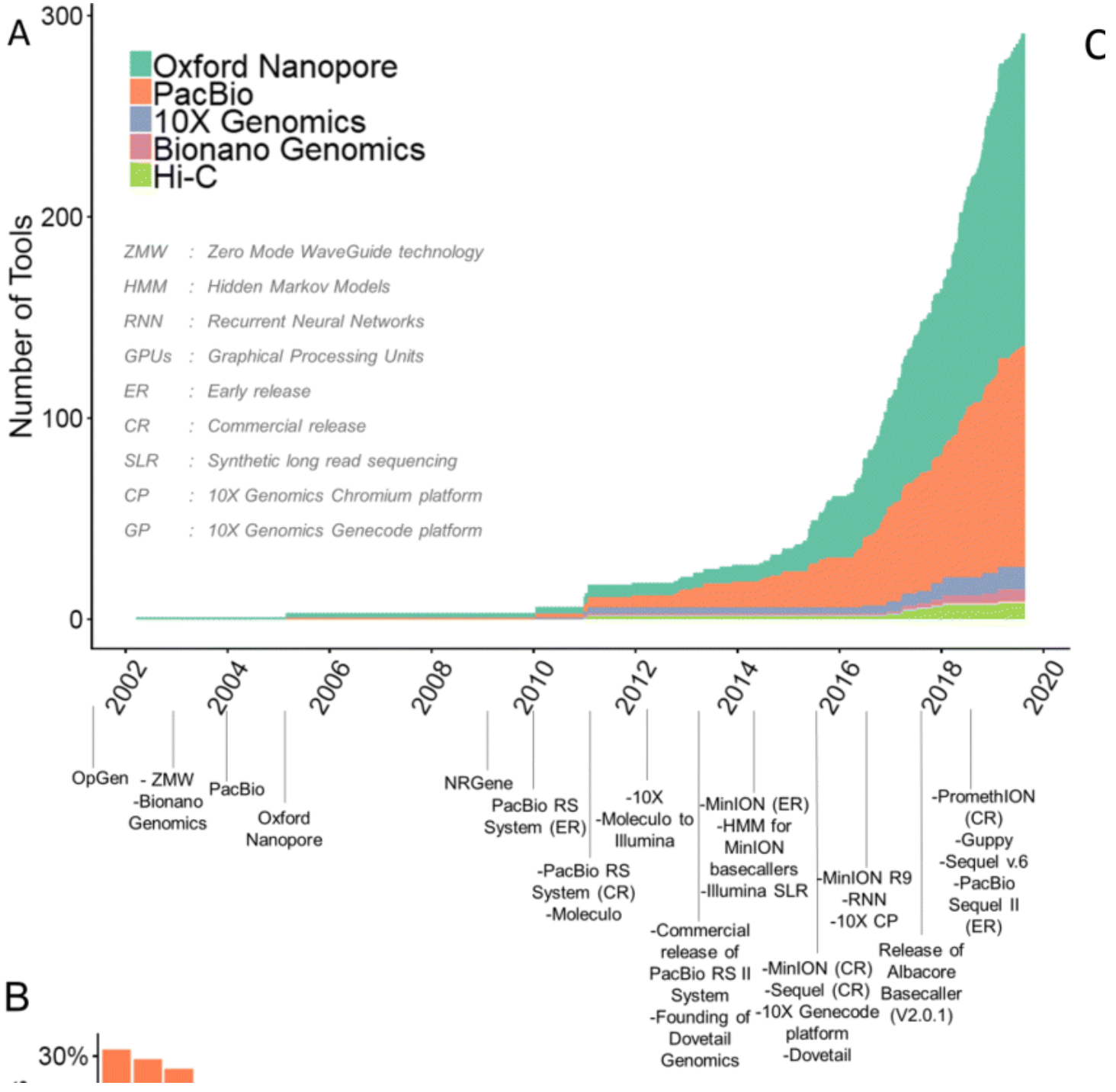

b

Long read technologies



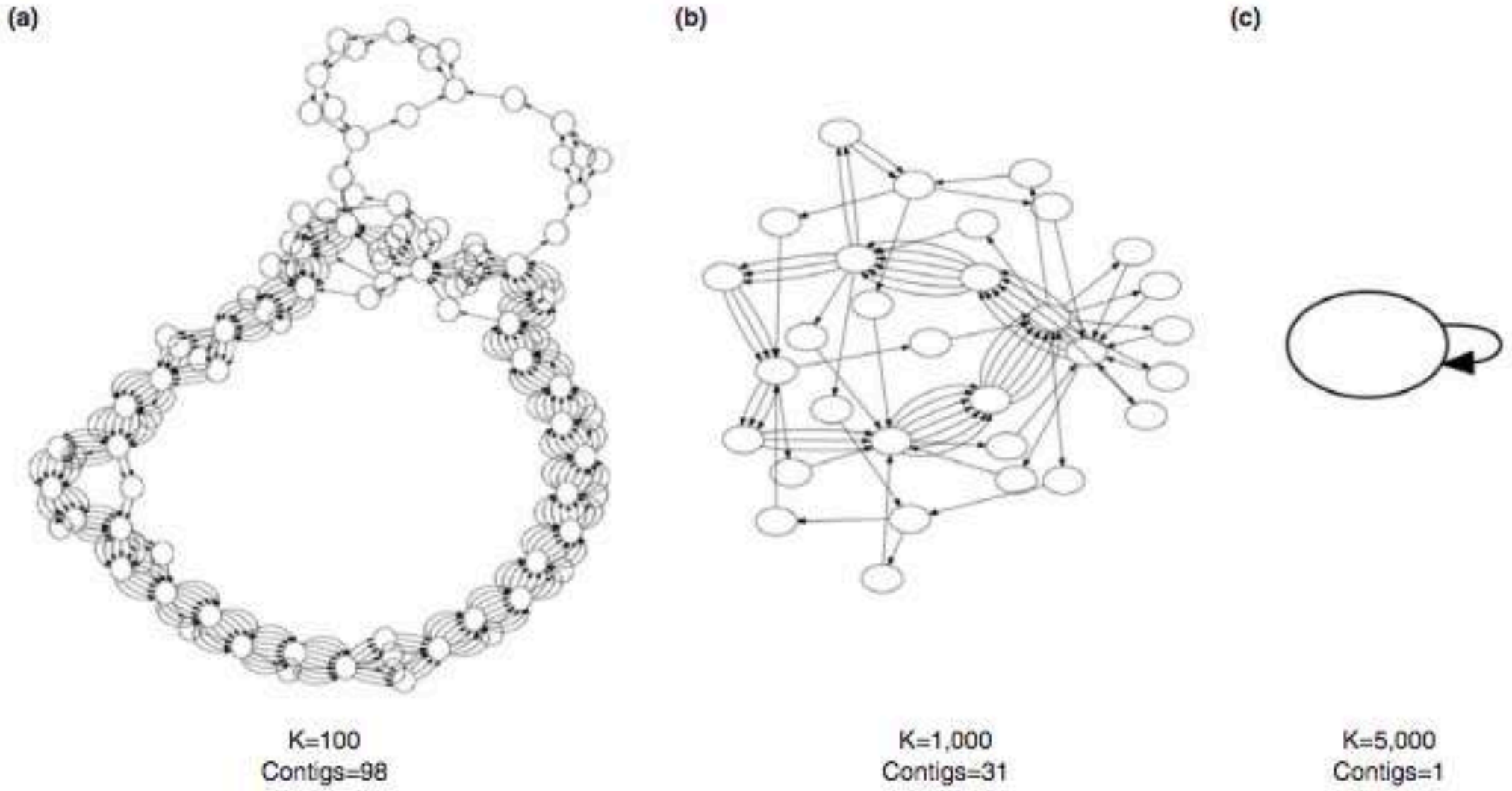
Credit: Michael Schatz

New tools continue to be developed



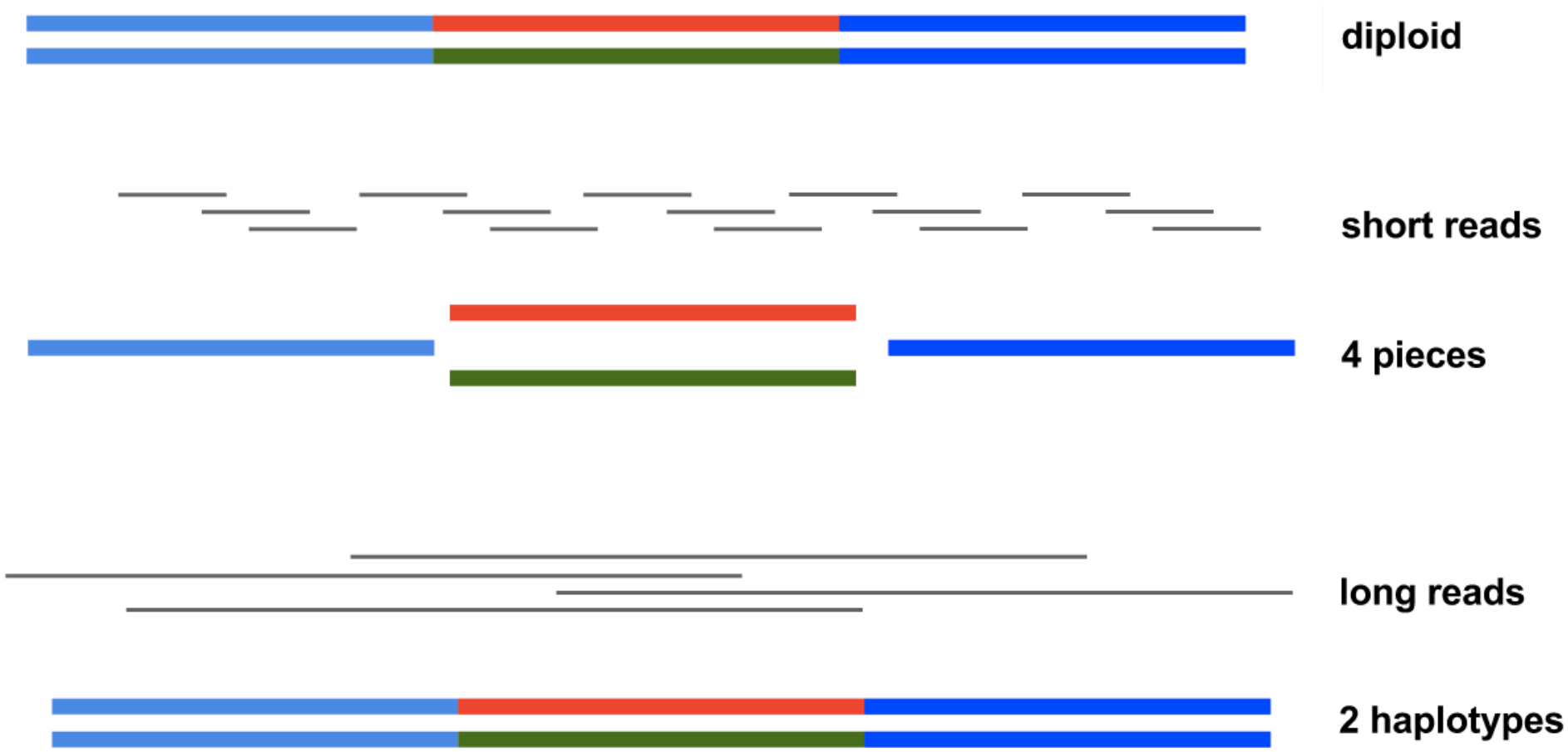
Amarasinghe *et al* (2020) Genome Biology

Simpler graphs

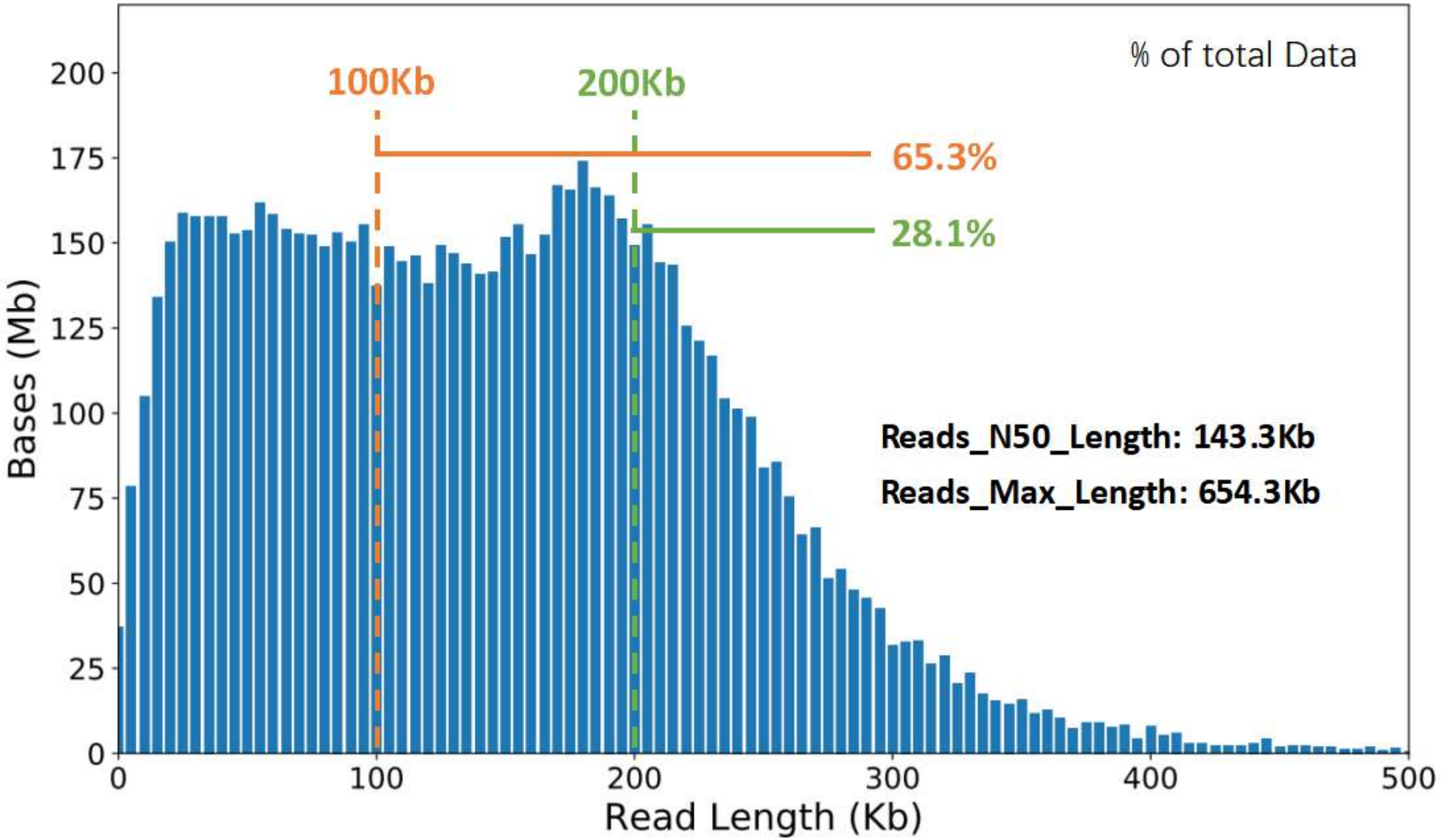


Roberts *et al.,* 2013

Long reads can also resolve haplotypes (with sufficient coverage)



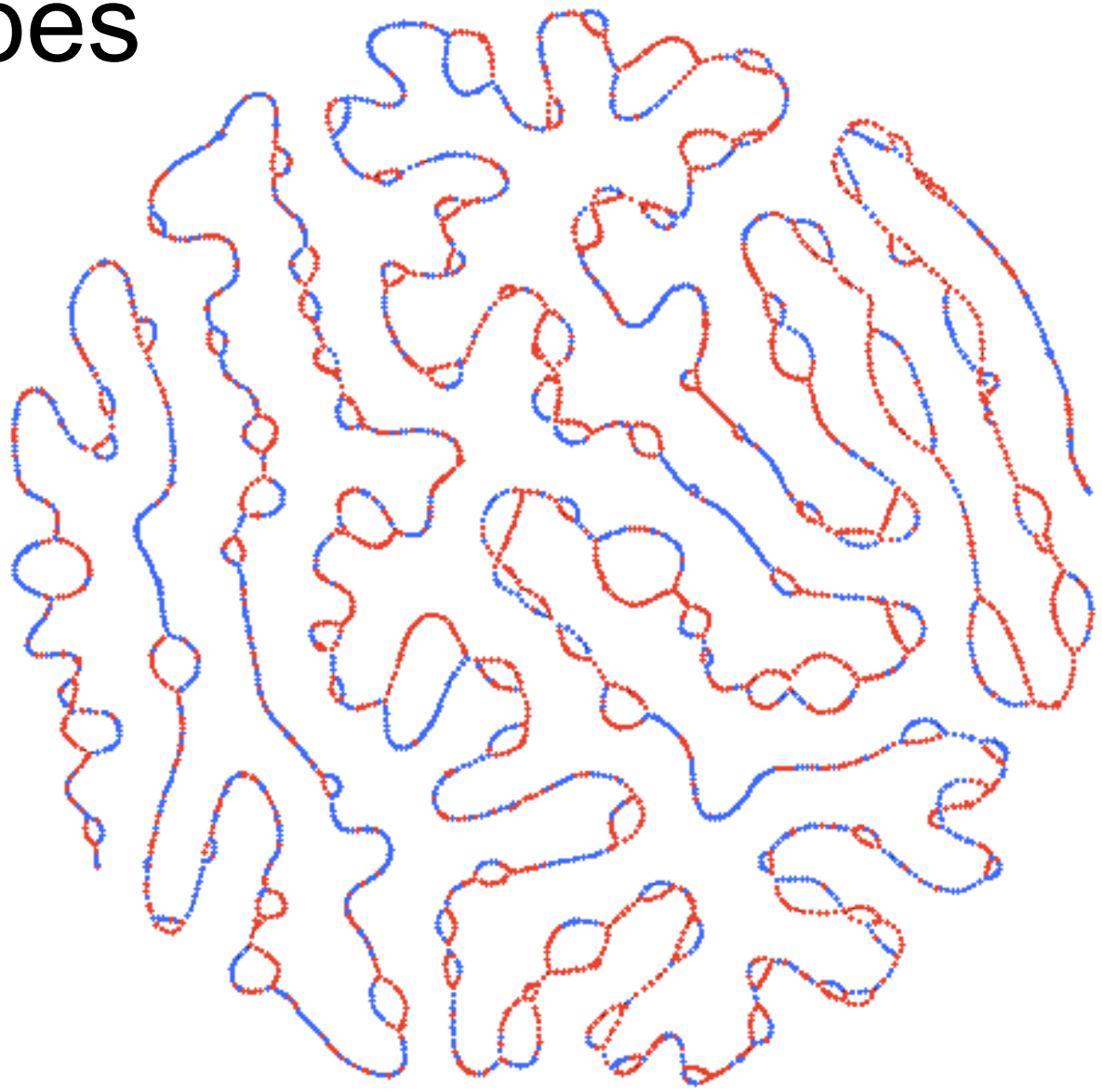
Read length and capacity go beyond



https://twitter.com/GrandOmics_Intl/status/1188724248480608257/photo/1

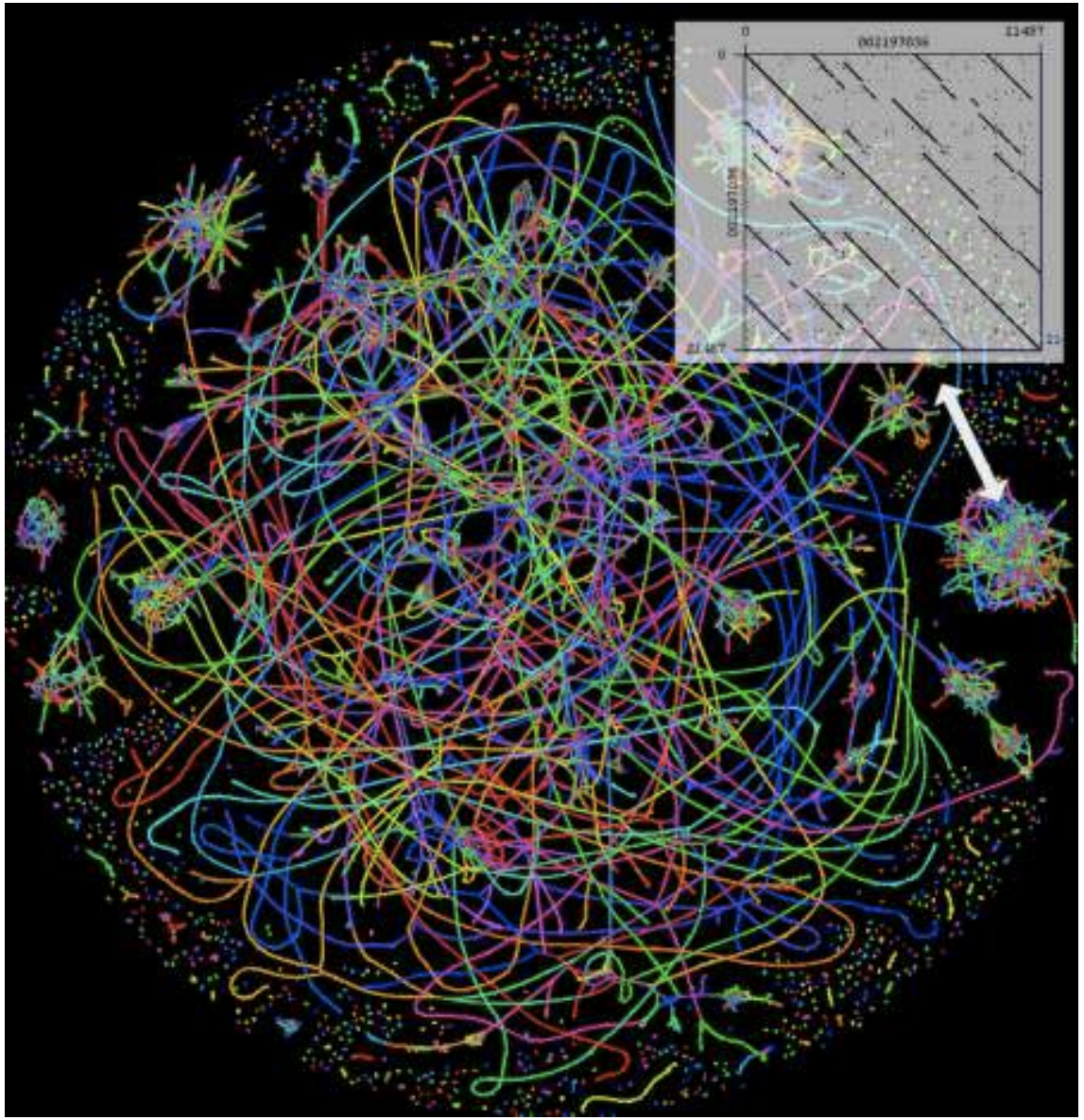
What you can do with long reads: Resolving haplotypes

Credit: Jason Chin
Two genomes. One assembly. Two colors. Many bubbles. Game against entropy.
http://t.co/uCPmxCRiZ6



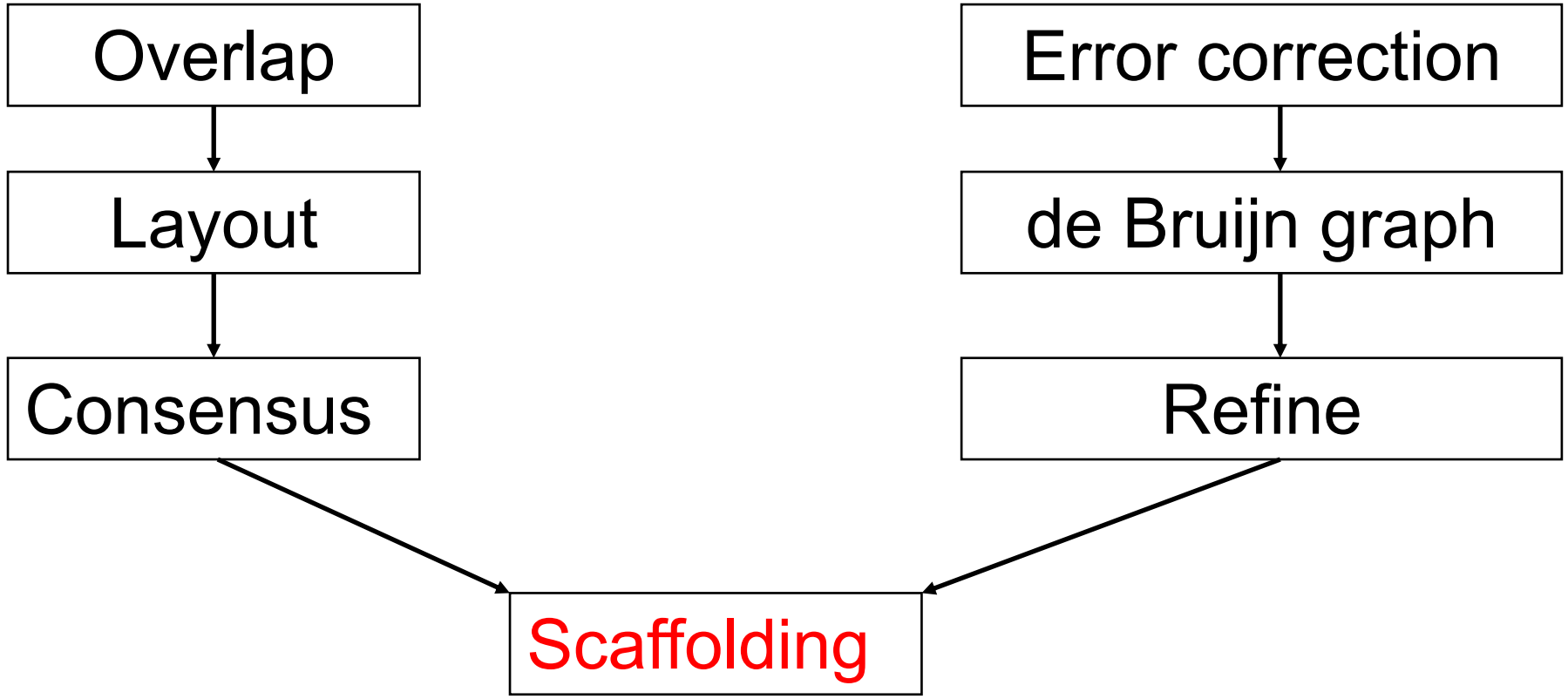
Current limitation of long reads..

Credit: Jason Chin
What are those blobs in the genome assembly graph? Intriguing repeats that have no NCBI blast hit.
http://t.co/2y7stBGs4W



Scaffolding

OLC and DBG assemblers



Scaffolding

OLC and DBG attempt to construct longest and most accurate **contigs** (**contig**uous stretch of assembled bases)

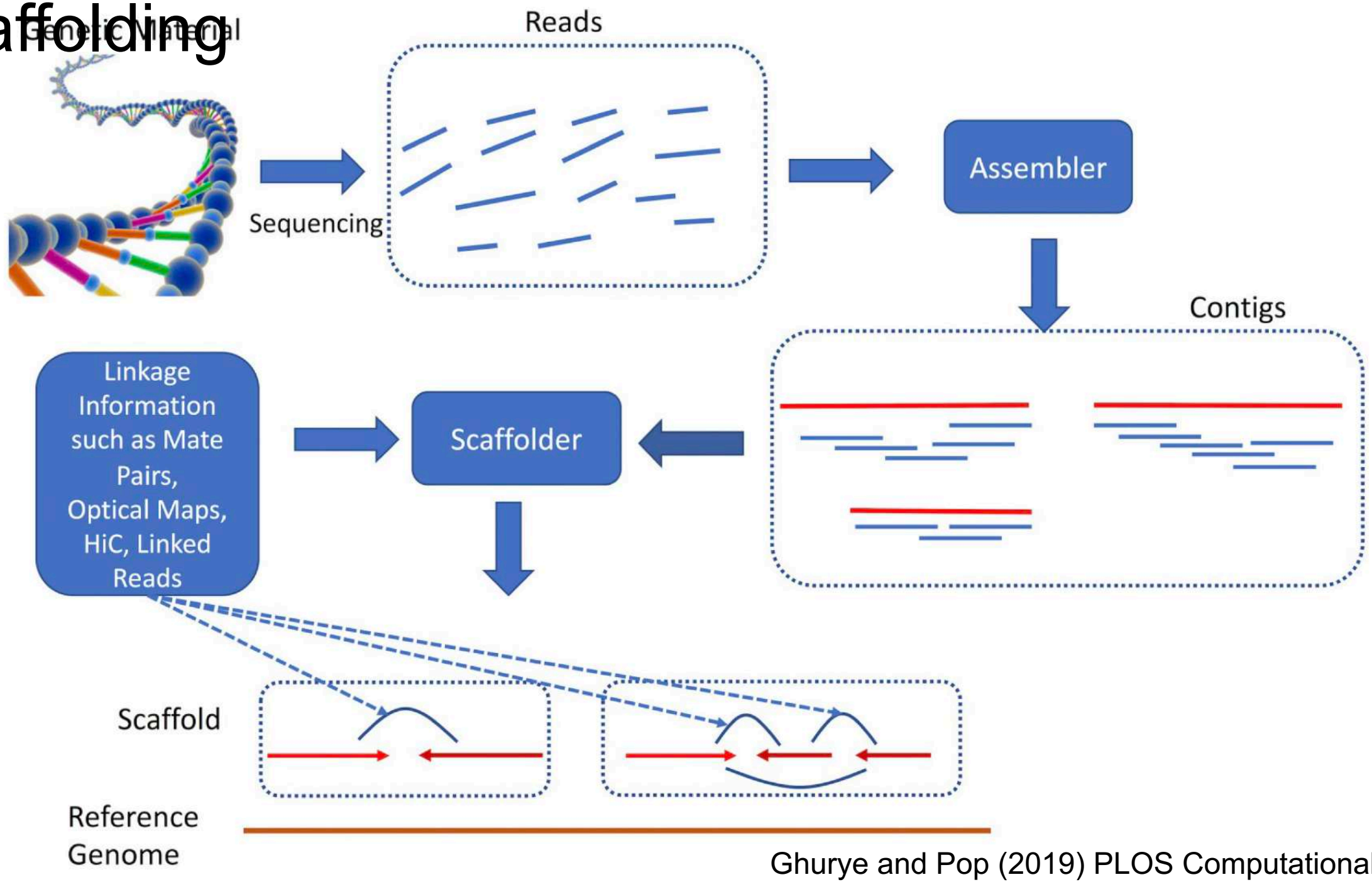
Scaffolding is to order and orient contigs with respect to each other

Various data types:

- Paired ends / Mate pairs
- Genetic map
- Additional long range information

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/scaffolding.pdf

Scaffolding



Ghurye and Pop (2019) PLOS Computational Biology

Scaffolding: Paired end sequencing

Because of technology limitation (usually ~150bp at each end), whole fragment is not sequenced. But the distance between two mates equals to length of fragment **(insert size)**

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/scaffolding.pdf

Scaffolding: Paired end sequencing

- DNA fragment (200-800 bp)

  ==========

- Single end

  ---------->==========

- Paired end (up to 800 bp span)

  ----->==========<-----

- Mate pair (up to 40 kbp span)

  ---->=========/+/==========<----

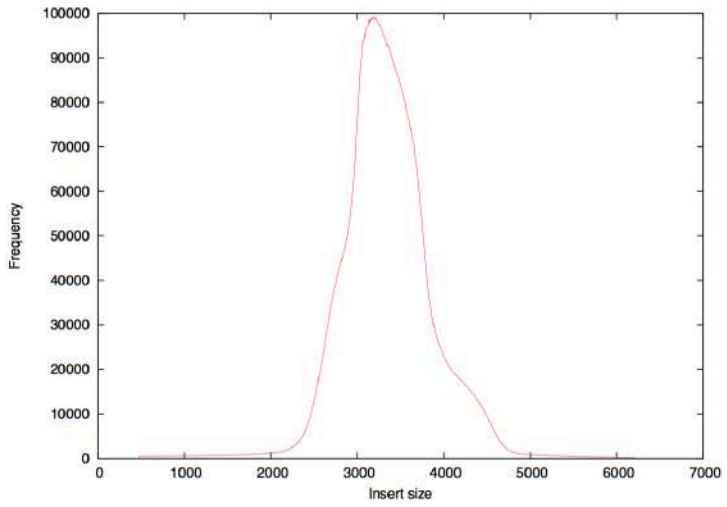
Examples

How to check insert size?

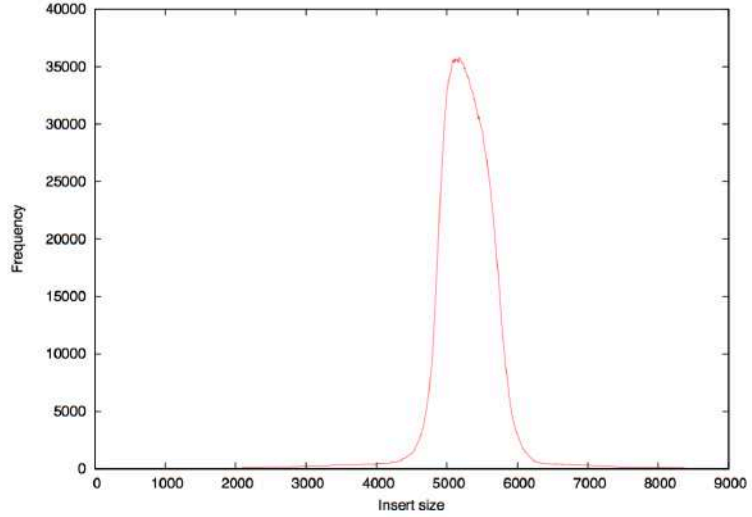
Remap the data back to the assembly

Problem can arise in larger insert sizes

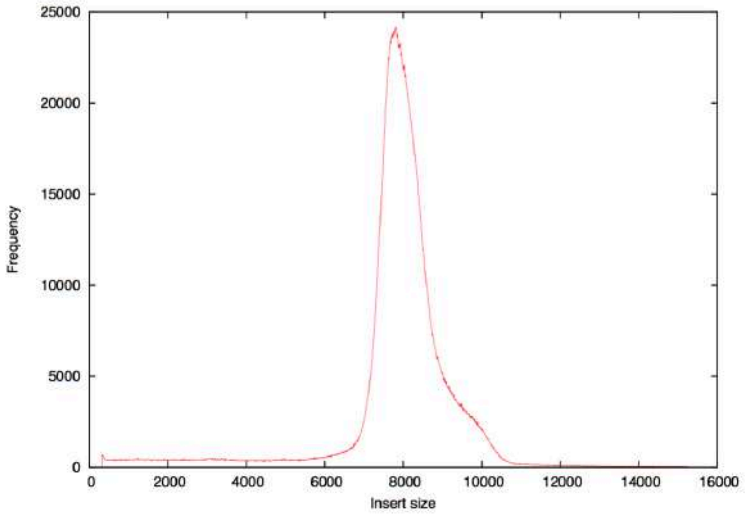
3kb



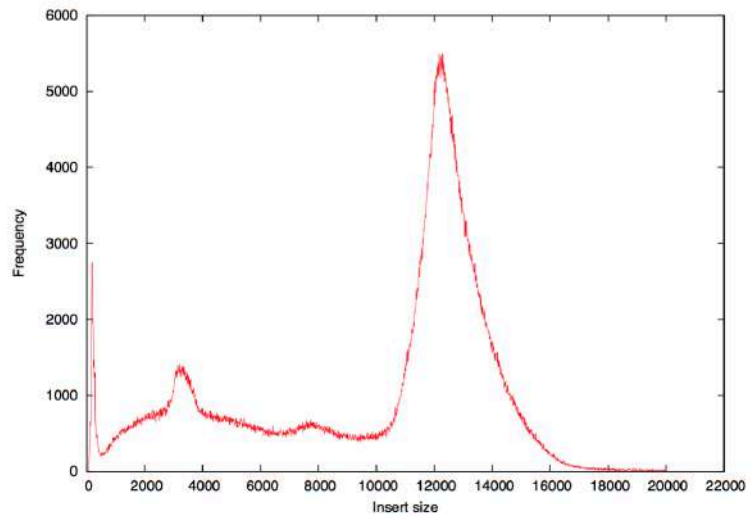
5kb



8kb



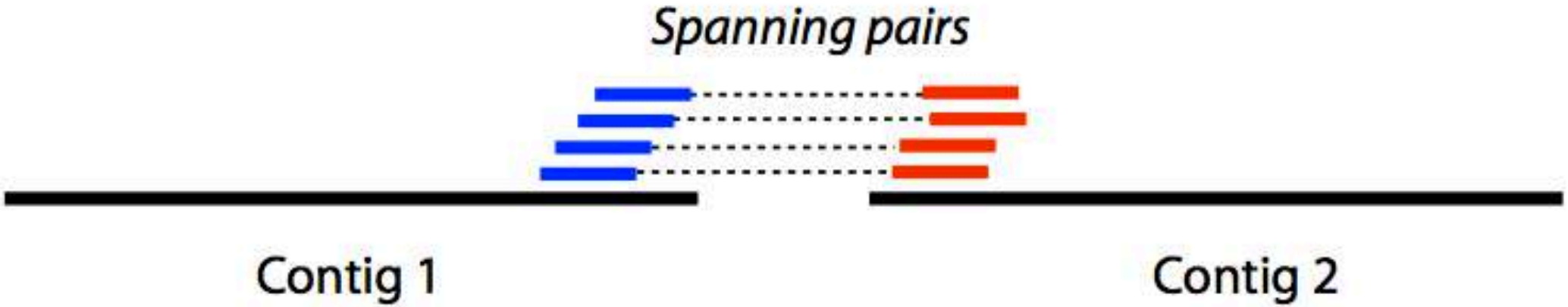
12kb



Scaffolding (Illumina; obsoleted)

Say we have a collection of pairs and we assemble them as usual

Assembly yields two contigs:



...and we discover that some of the mates at one edge of contig 1 are paired with mates in contig 2

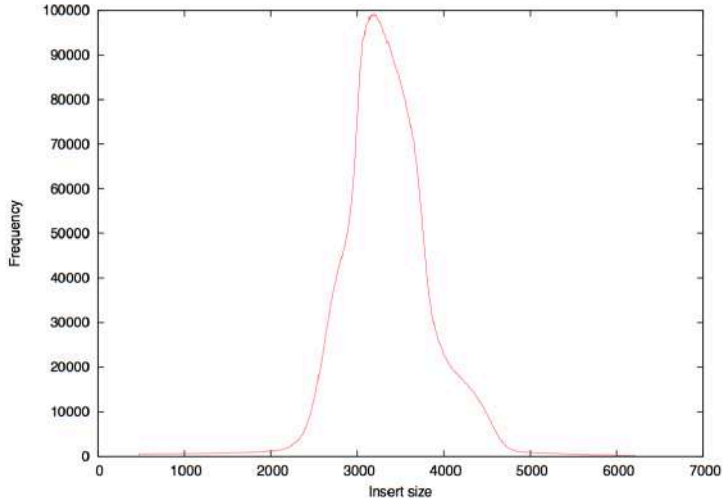
Call these *spanning pairs*

http://www.cs.jhu.edu/~langmea/resources/lecture_notes/scaffolding.pdf

Scaffolding (Illumina; obsoleted)

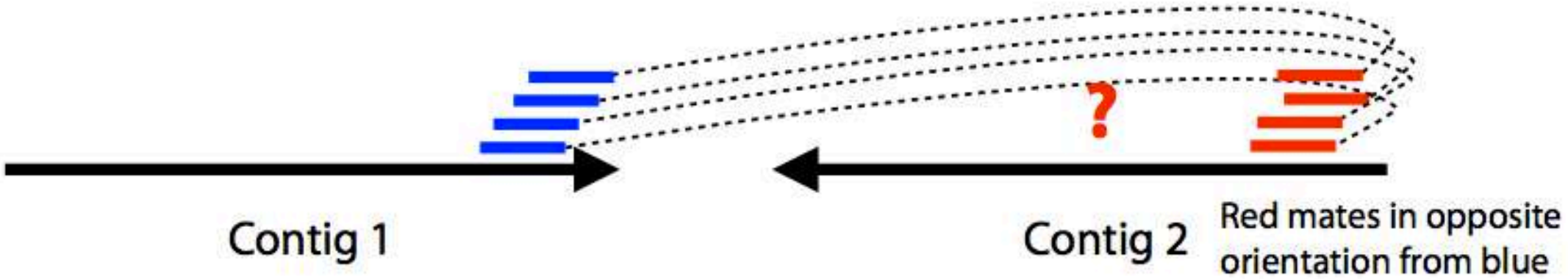
Does the distance equal to the insert size distribution mapped to the rest the genome?

Yes = Accept and Contig1NNNNContig2
No = Reject

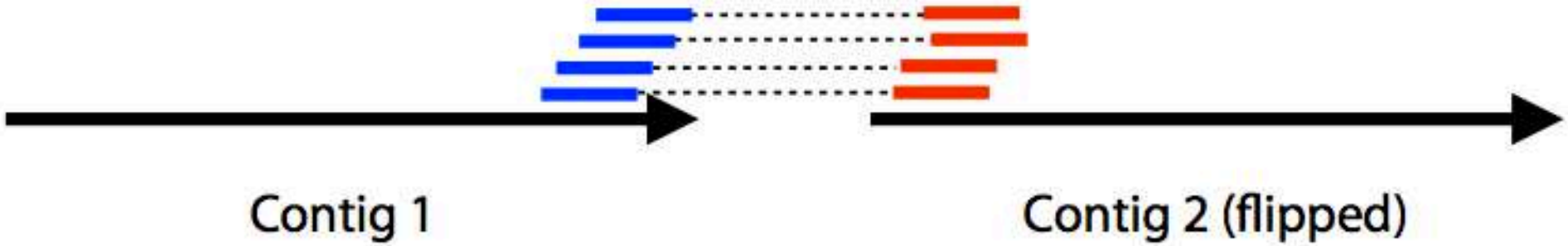


http://www.cs.jhu.edu/~langmea/resources/lecture_notes/scaffolding.pdf

Scaffolding (Illumina; obsoleted)



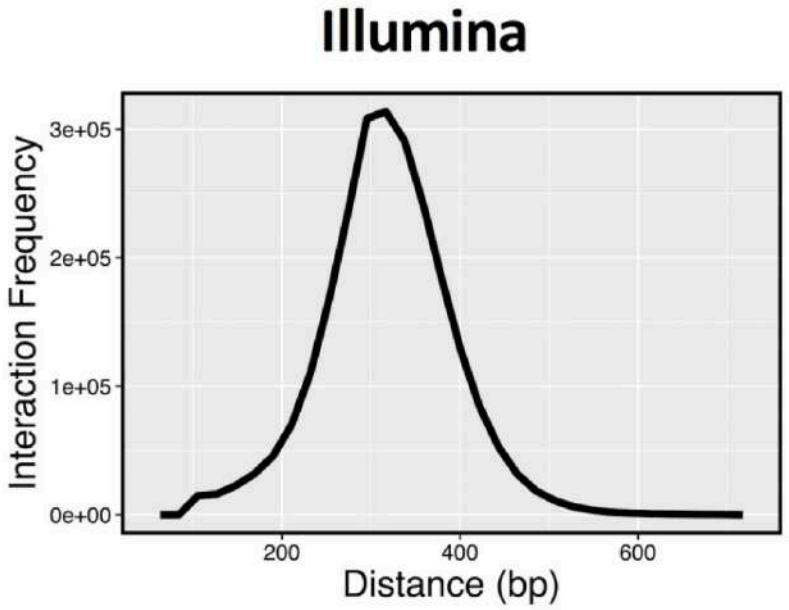
What does the picture look like if contigs 1 and 2 are close, but we assembled contig 2 "backwards" (i.e. reverse complemented)

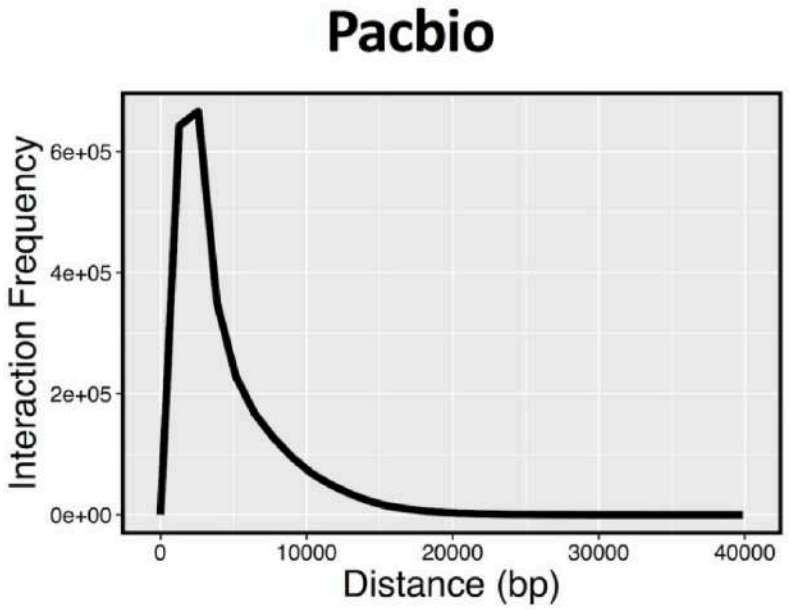


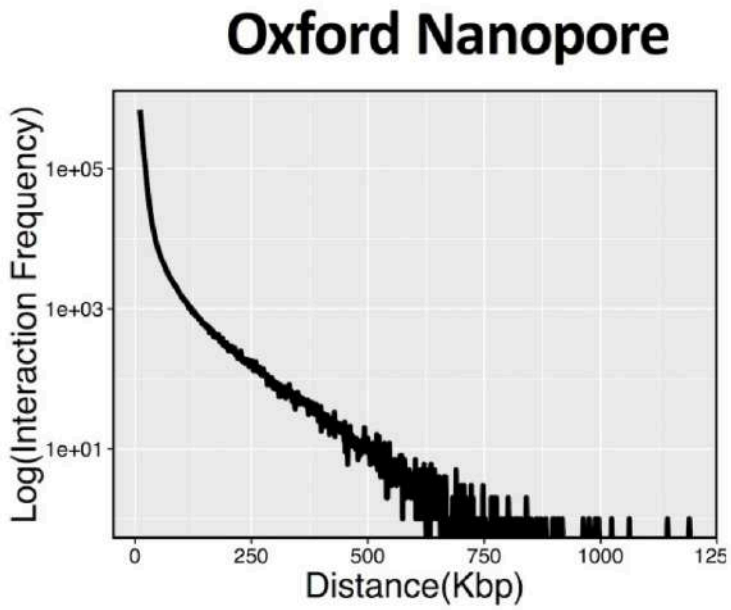
Pairs also tell us about contigs' relative *orientation*

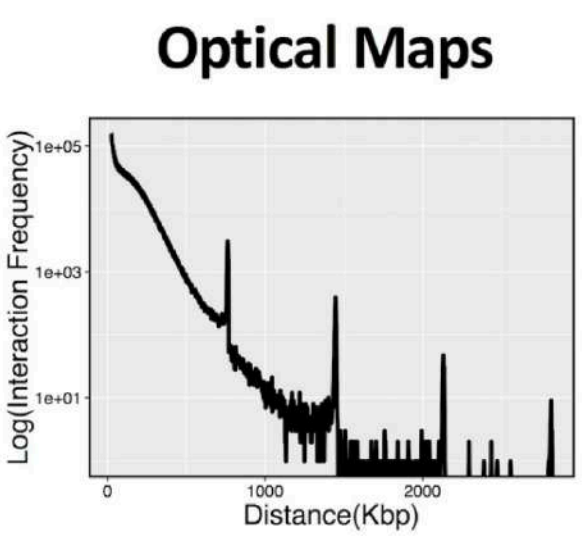
http://www.cs.jhu.edu/~langmea/resources/lecture_notes/scaffolding.pdf

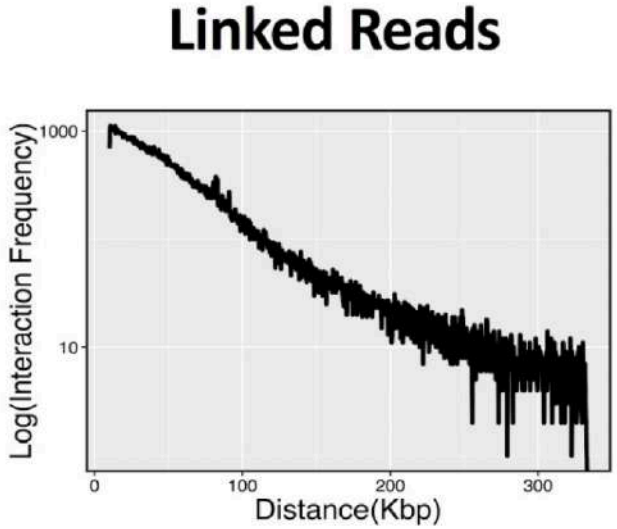
The genomic span covered by different technologies

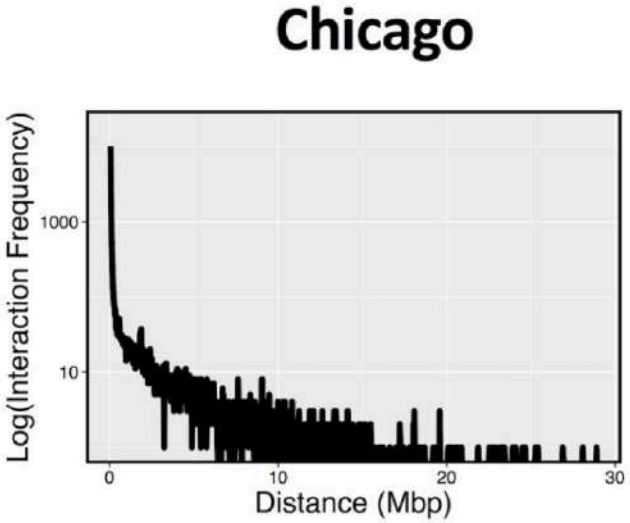


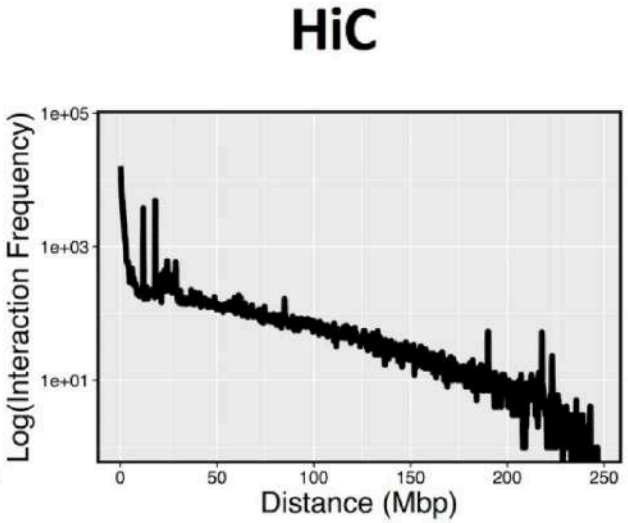






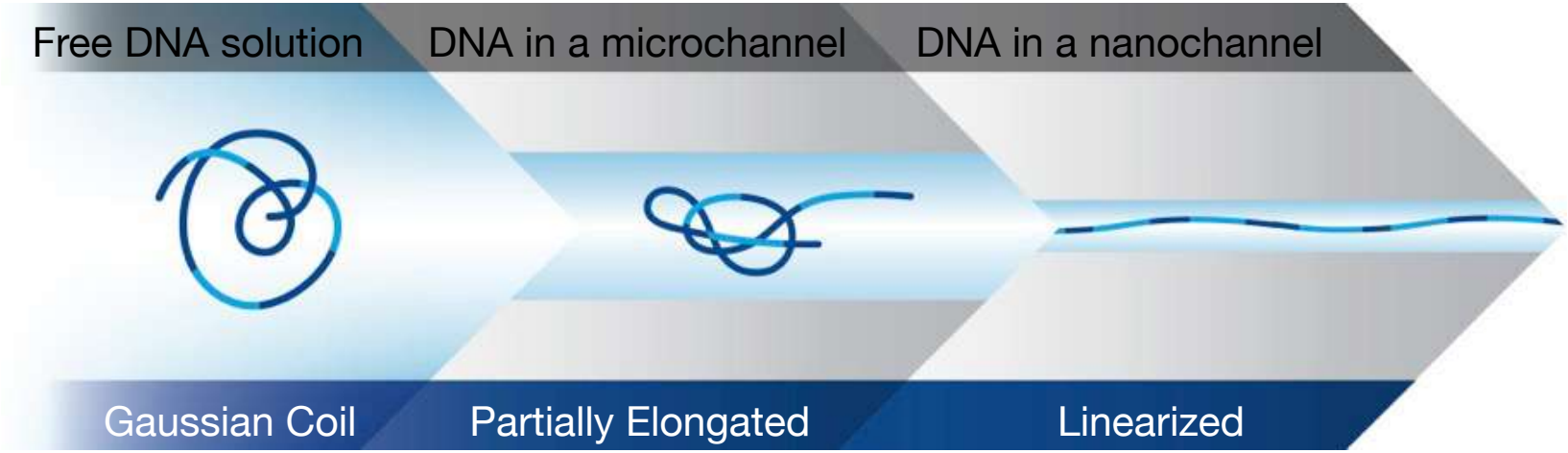




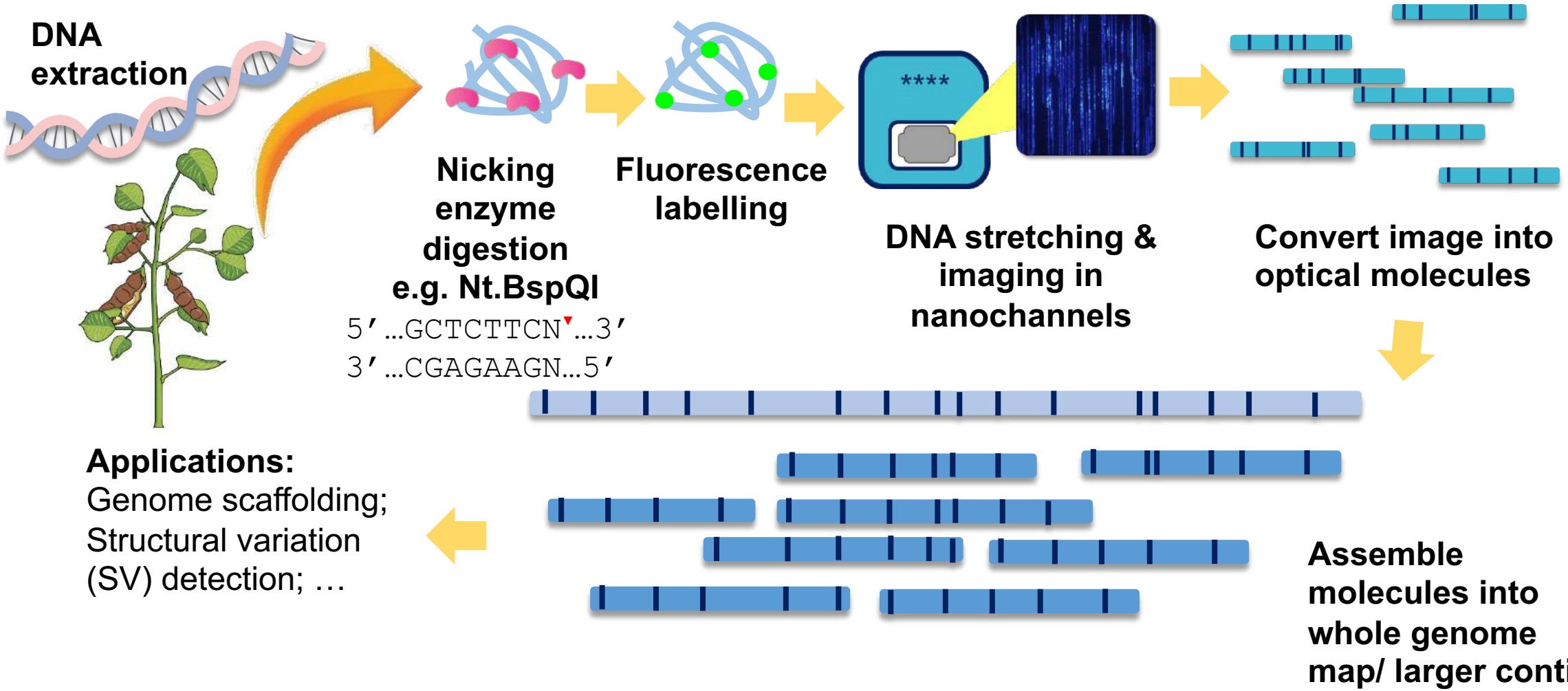


Ghurye and Pop (2019) PLOS Computational Biology

Optical map: DNA linearized in nanochannel array



Workflow of an optical mapping procedure



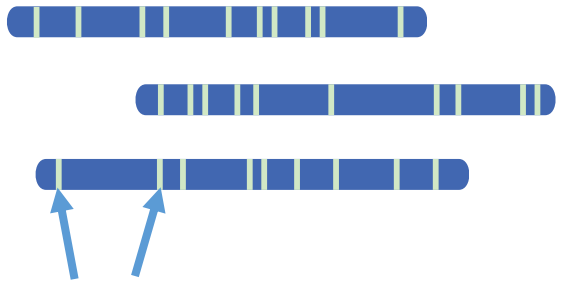
Optical mapping-assisted scaffolding principles

Sequence-specific nicking & Fluorescence labelling

Nanochannel-array fluorescent imaging

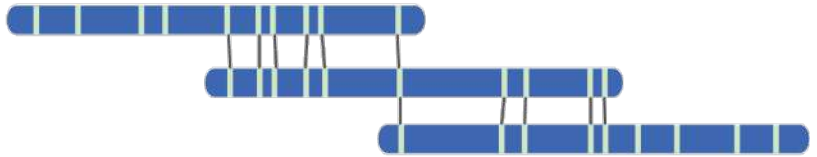
High molecular weight DNA

Optical molecules (up to 2 Mbps)

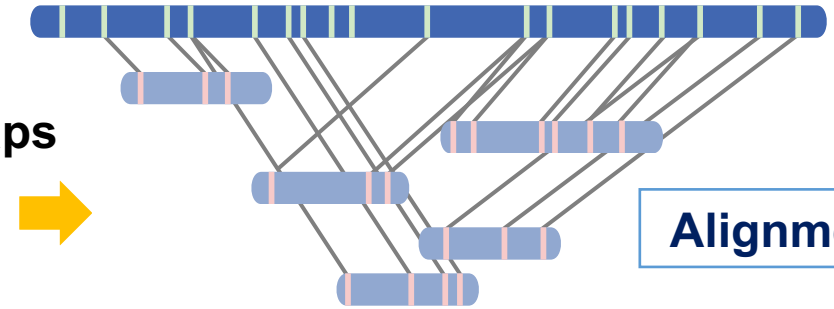


Signals (nicking sites)

De novo optical map assembly



Assembled reference optical map



Alignment

Sequence contigs

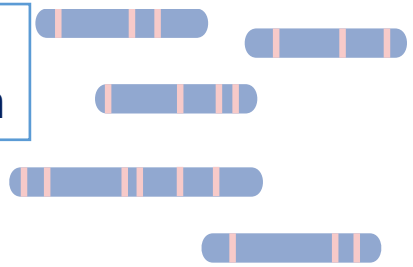
…TAAATAAATTTATATAGCT
AAAAACACGAGGGAATGA
ATTAATCAT…

…CAATAAAGGCGCGACGTGGTG
CTCAACATCATCGCCCATTAAGCA
TGAGA…

…CTGCTCAGCATTCATCTCGCC
CAGACCTTTACACGAGTGGATC
ATCAT…

In silico digestion

Predicted optical maps



Long scaffold

BIONANO GENOMICS

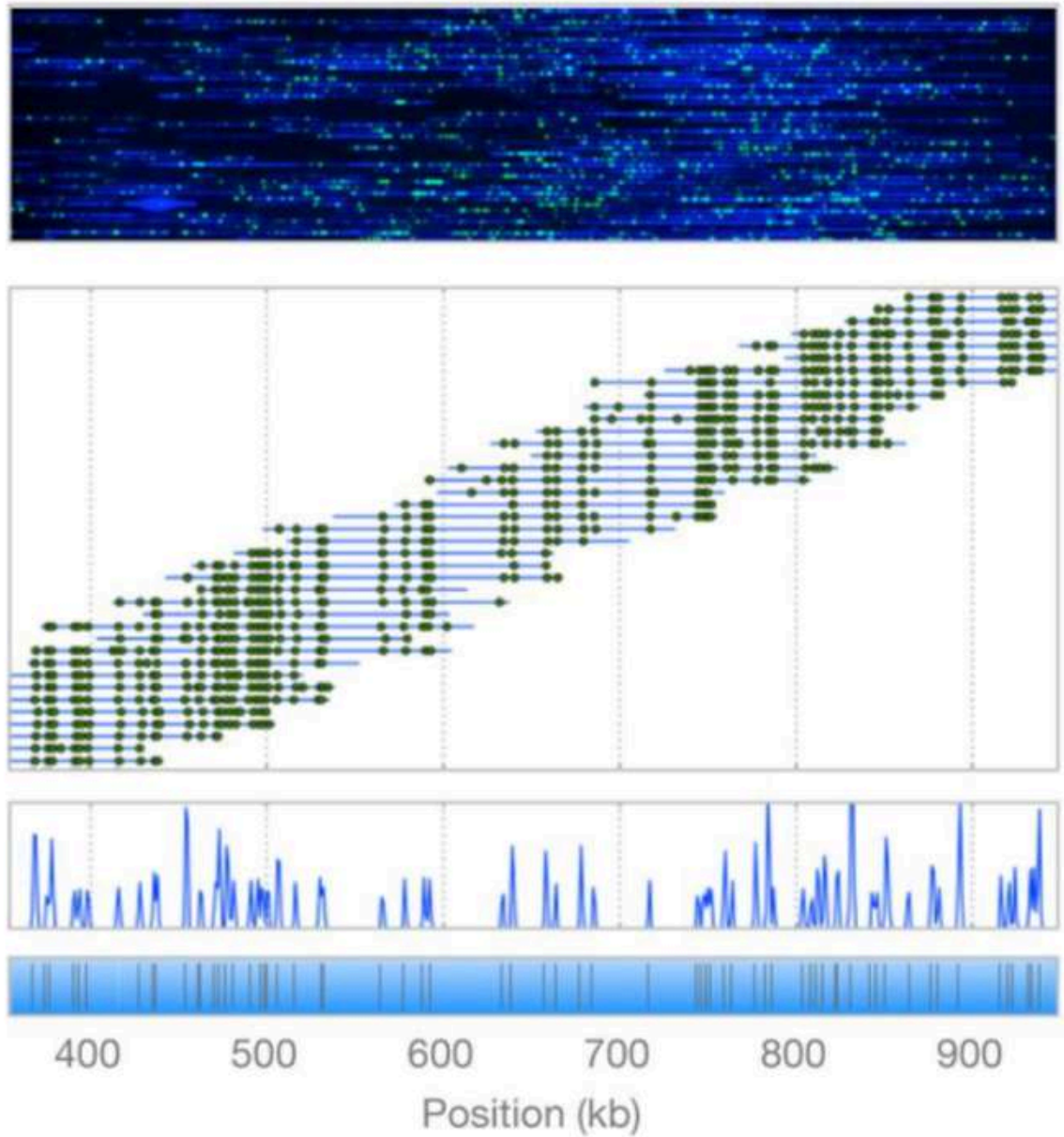
Raw Image Data

Conversion to Molecules

Population of Molecules

Aggregation & Assembly

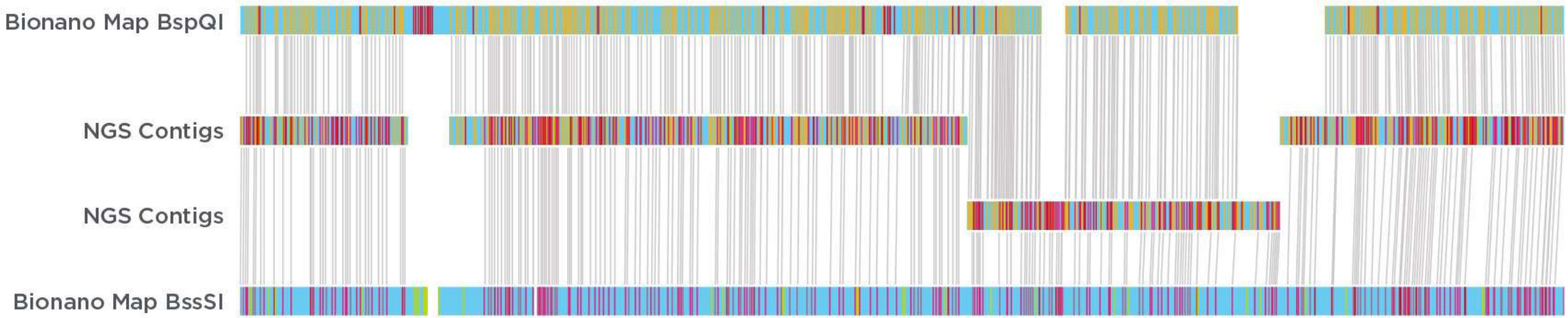
Consensus Genome Map



https://bionanogenomics.com/technology/platform-technology/

BIONANO GENOMICS

TWO ENZYME HYBRID SCAFFOLDING



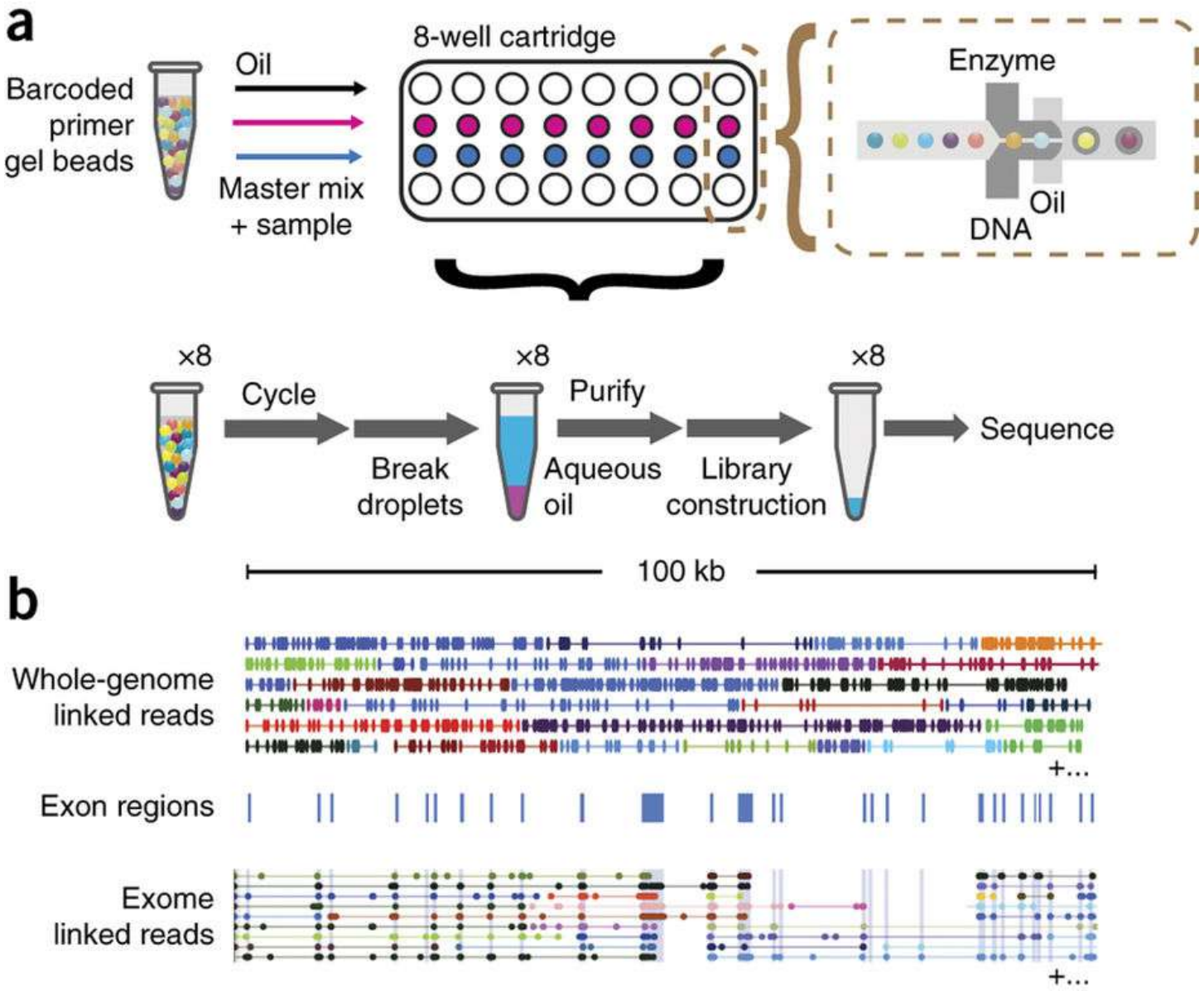
https://bionanogenomics.com/technology/platform-technology/

10X GENOMICS

- Long range information from short reads using 14bp barcodes
- Very low input DNA (ng) and 20 mins preparation time
- 1ng of DNA is split across 100,000 Gel coated beads
- Single-cell available

https://www.nature.com/articles/nbt.3432

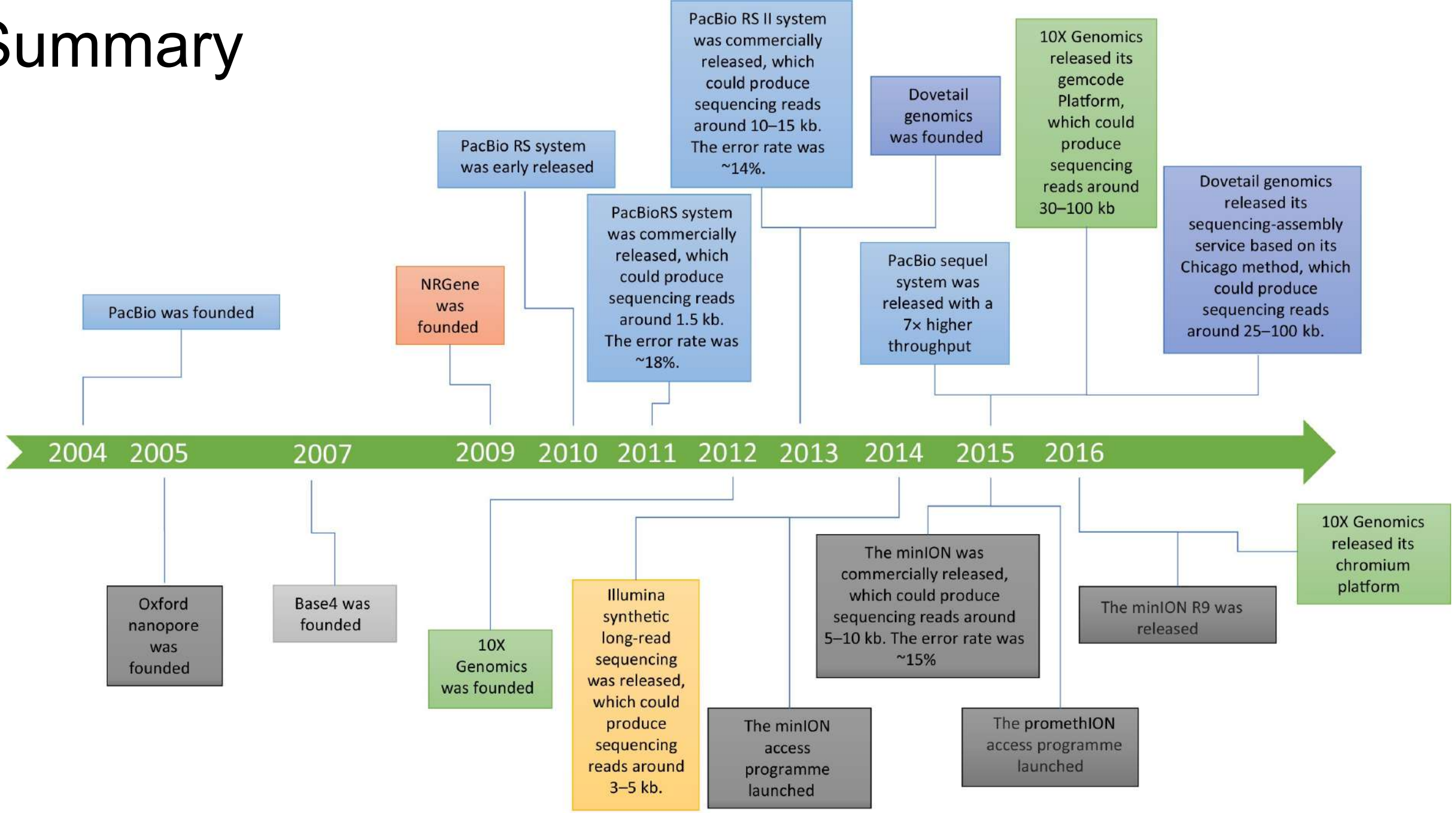
Haplotyping germline and cancer genomes with high-throughput linked-read sequencing



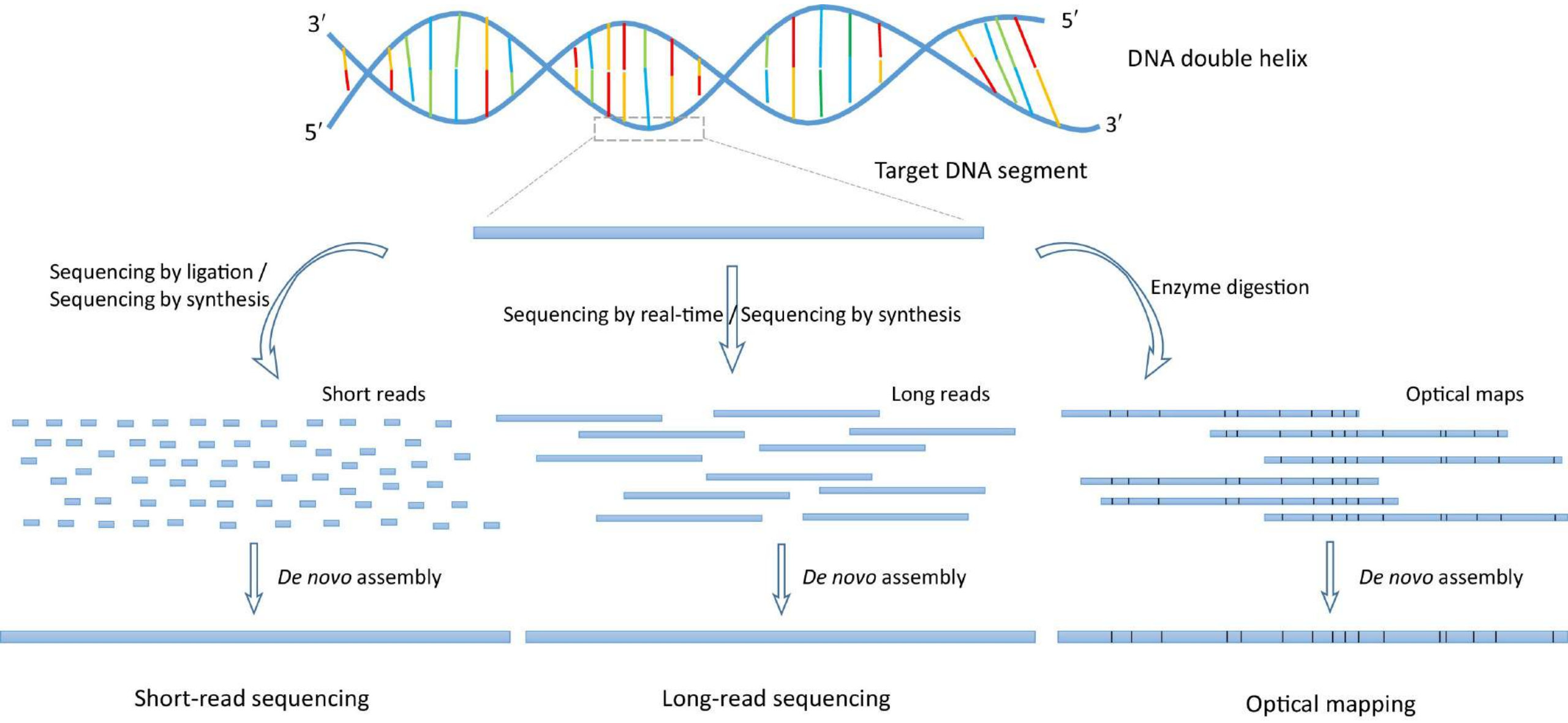
10X Genomics

https://www.youtube.com/watch?v=aUyFzwRFWJQ

Summary

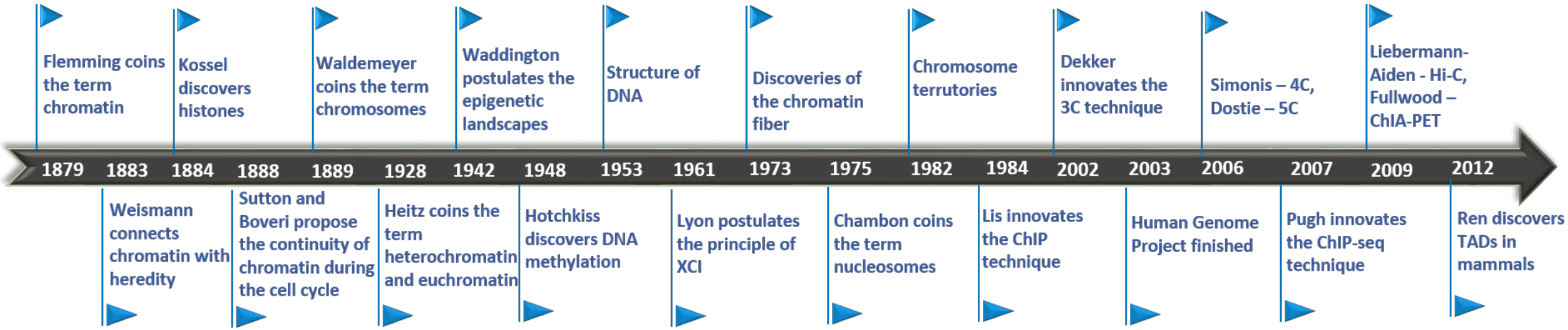

PacBio was founded
Oxford nanopore was founded
NRGene was founded
Base4 was founded
PacBio RS system was early released
10X Genomics was founded
PacBioRS system was commercially released, which could produce sequencing reads around 1.5 kb. The error rate was ~18%.
Illumina synthetic long-read sequencing was released, which could produce sequencing reads around 3–5 kb.
PacBio RS II system was commercially released, which could produce sequencing reads around 10–15 kb. The error rate was ~14%.
The minION access programme launched
Dovetail genomics was founded
PacBio sequel system was released with a 7× higher throughput
The minION was commercially released, which could produce sequencing reads around 5–10 kb. The error rate was ~15%
The promethION access programme launched
10X Genomics released its gemcode Platform, which could produce sequencing reads around 30–100 kb
Dovetail genomics released its sequencing-assembly service based on its Chicago method, which could produce sequencing reads around 25–100 kb.
The minION R9 was released
10X Genomics released its chromium platform
2004
2005
2007
2009
2010
2011
2012
2013
2014
2015
2016

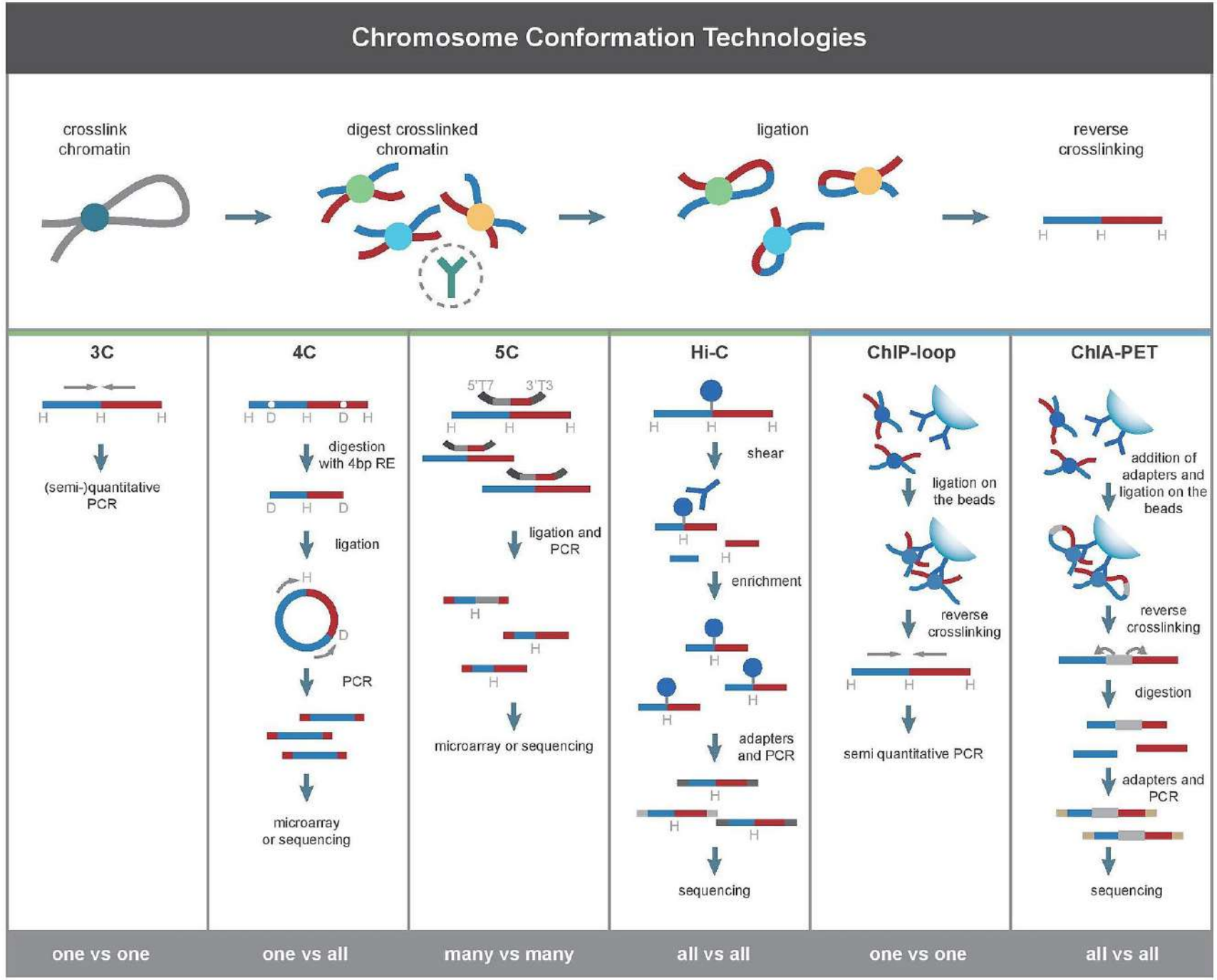
Summary


3′
5′
5′
3′
DNA double helix
Target DNA segment
Sequencing by ligation /
Sequencing by synthesis
Sequencing by real-time / Sequencing by synthesis
Enzyme digestion
Short reads
Long reads
Optical maps
De novo assembly
De novo assembly
De novo assembly
Short-read sequencing
Long-read sequencing
Optical mapping

Scaffolding using Chromosome conformation capture

Chromosome conformation capture techniques (often abbreviated to 3C technologies or 3C-based methods) are a set of molecular biology methods used to analyze the spatial organization of chromatin in a cell. These methods quantify the **number of interactions between genomic loci that are nearby in 3-D space**, but may be **separated by many nucleotides** in the linear genome




Chromosome Conformation Technologies
crosslink chromatin
digest crosslinked chromatin
ligation
reverse crosslinking
H H H
3C
H H H
(semi-)quantitative PCR
4C
H D H D H
digestion with 4bp RE
D H D
ligation
H
D
PCR
microarray or sequencing
5C
5'T7 3'T3
H H H
ligation and PCR
H
H
H
microarray or sequencing
Hi-C
H H H
shear
H
H
enrichment
H
H
H
H
adapters and PCR
H
H H
sequencing
ChIP-loop
ligation on the beads
reverse crosslinking
H H H
semi quantitative PCR
ChIA-PET
addition of adapters and ligation on the beads
reverse crosslinking
digestion
adapters and PCR
sequencing
one vs one
one vs all
many vs many
all vs all
one vs one
all vs all

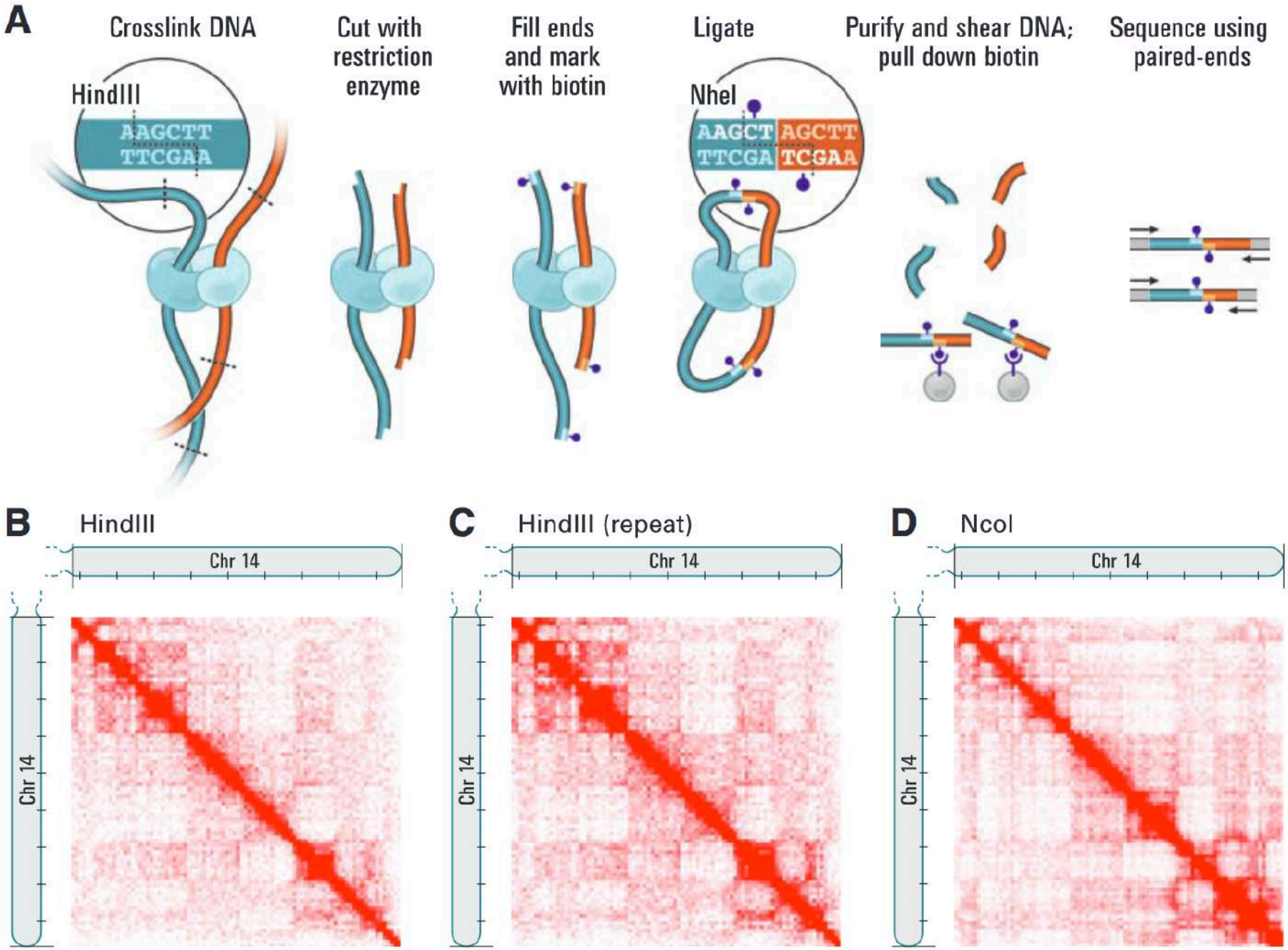
Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

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]†

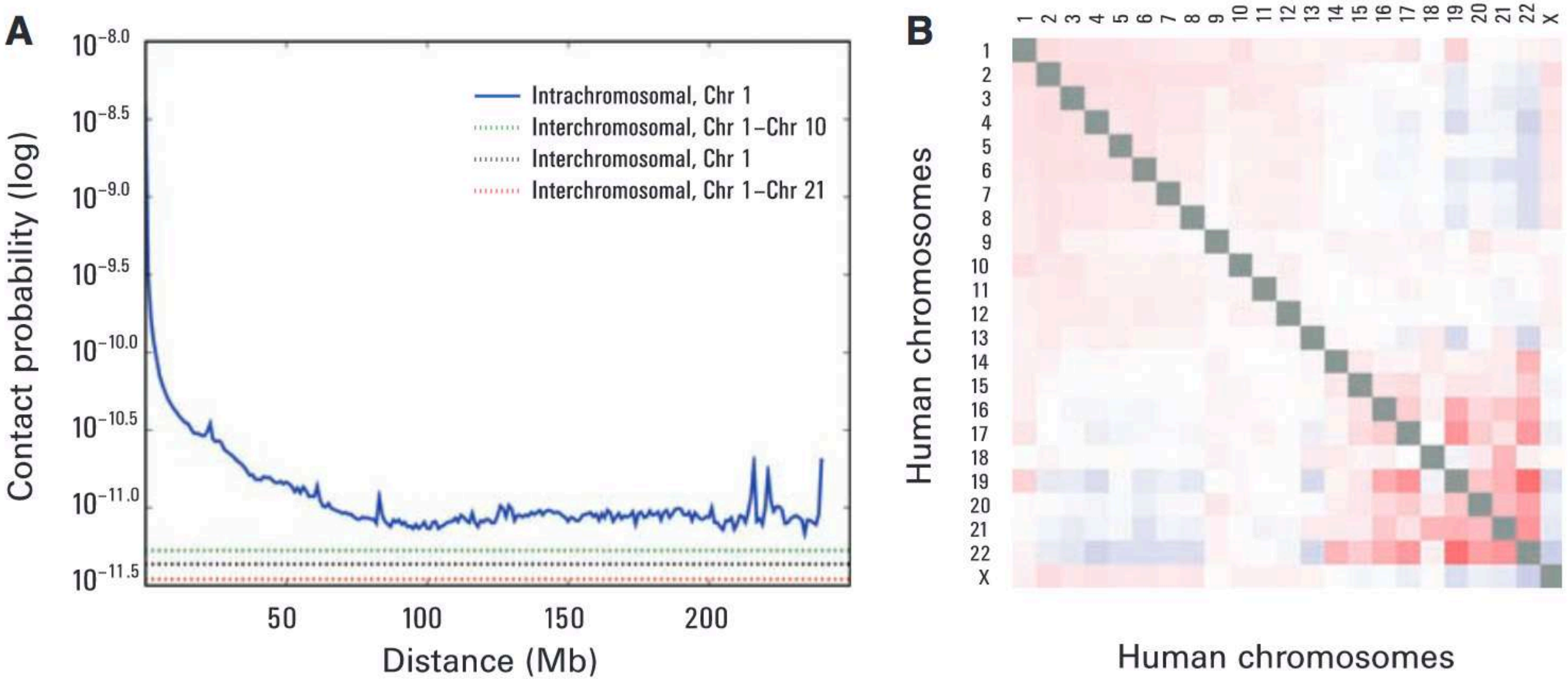
We describe Hi-C, a method that probes the three-dimensional architecture of whole genomes by coupling proximity-based ligation with massively parallel sequencing. We constructed spatial proximity maps of the human genome with Hi-C at a resolution of 1 megabase. These maps confirm the presence of chromosome territories and the spatial proximity of small, gene-rich chromosomes. We identified an additional level of genome organization that is characterized by the spatial segregation of open and closed chromatin to form two genome-wide compartments.

Lieberman-Aiden (2009) Science 10.1126/science.1181369

~Cited 3000 times



Lieberman-Aiden (2009) Science 10.1126/science.1181369



Lieberman-Aiden (2009) Science 10.1126/science.1181369

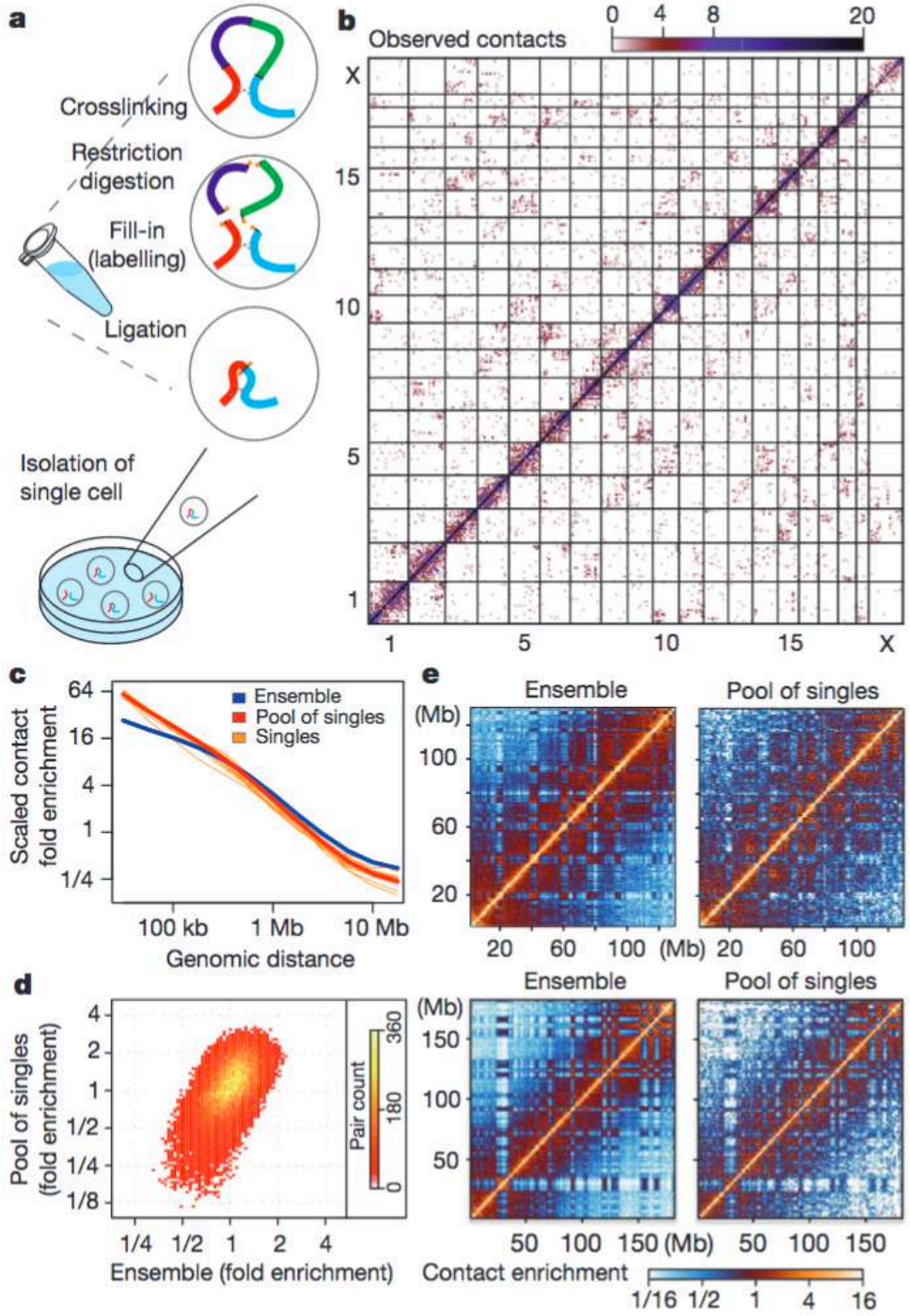
Hi-C at single cell level

ARTICLE

doi:10.1038/nature12593

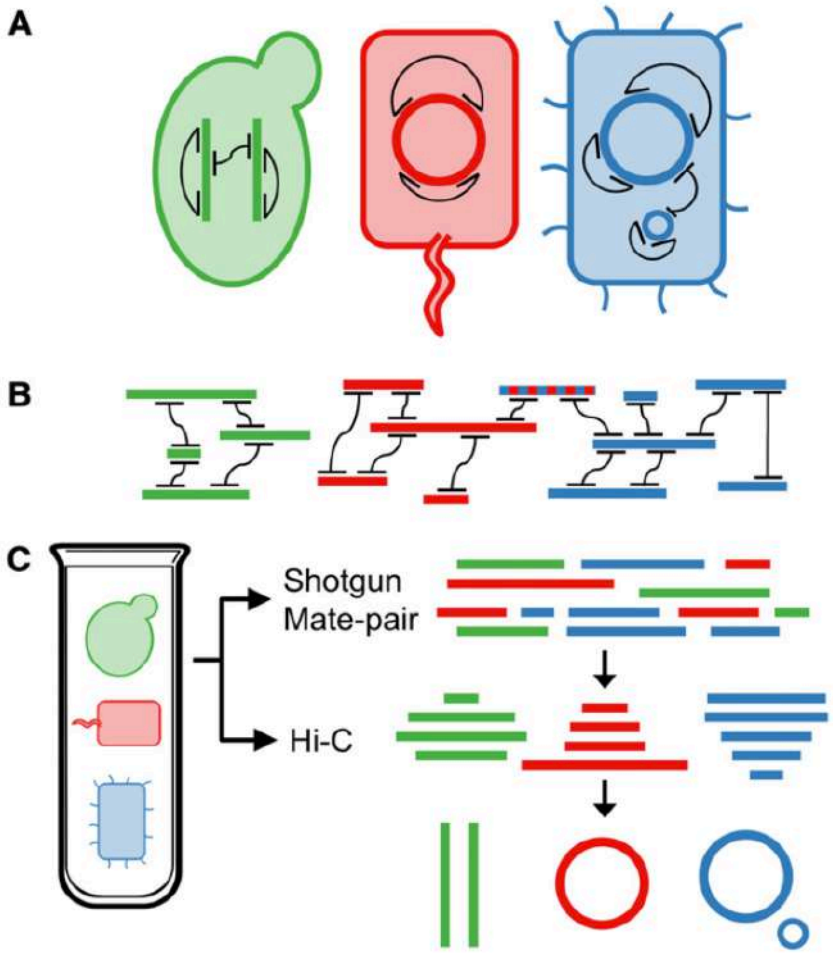
Single-cell Hi-C reveals cell-to-cell variability in chromosome structure

Takashi Nagano[1]*, Yaniv Lubling[2]*, Tim J. Stevens[3]*, Stefan Schoenfelder[1], Eitan Yaffe[2], Wendy Dean[4], Ernest D. Laue[3], Amos Tanay[2] & Peter Fraser[1]



Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps

Joshua N. Burton,[1] Ivan Liachko,[1] Maitreya J. Dunham,[2] and Jay Shendure[2]
Department of Genome Sciences, University of Washington, Seattle, Washington 98195-5065



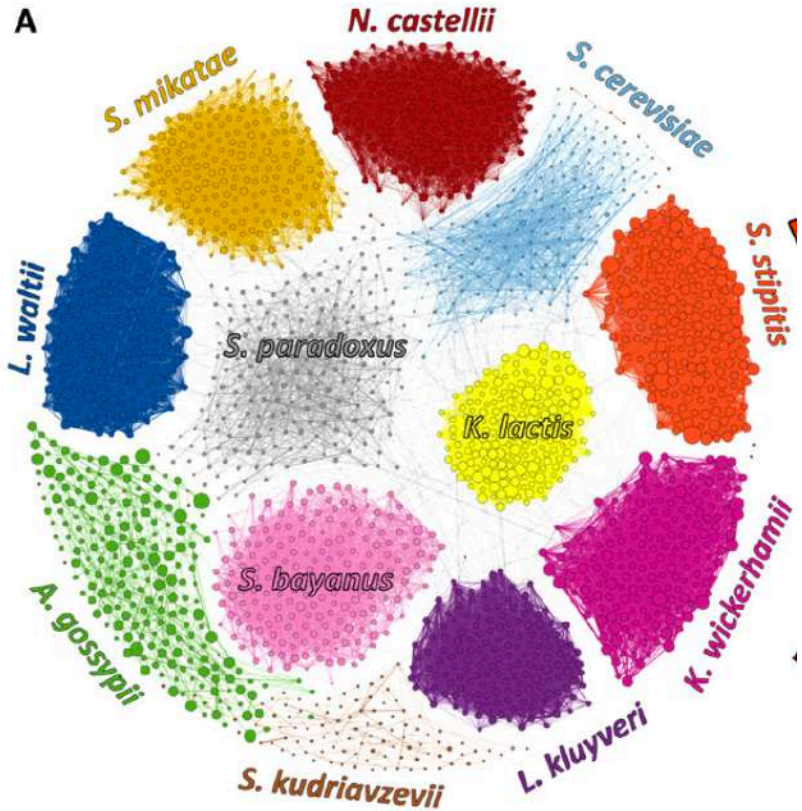
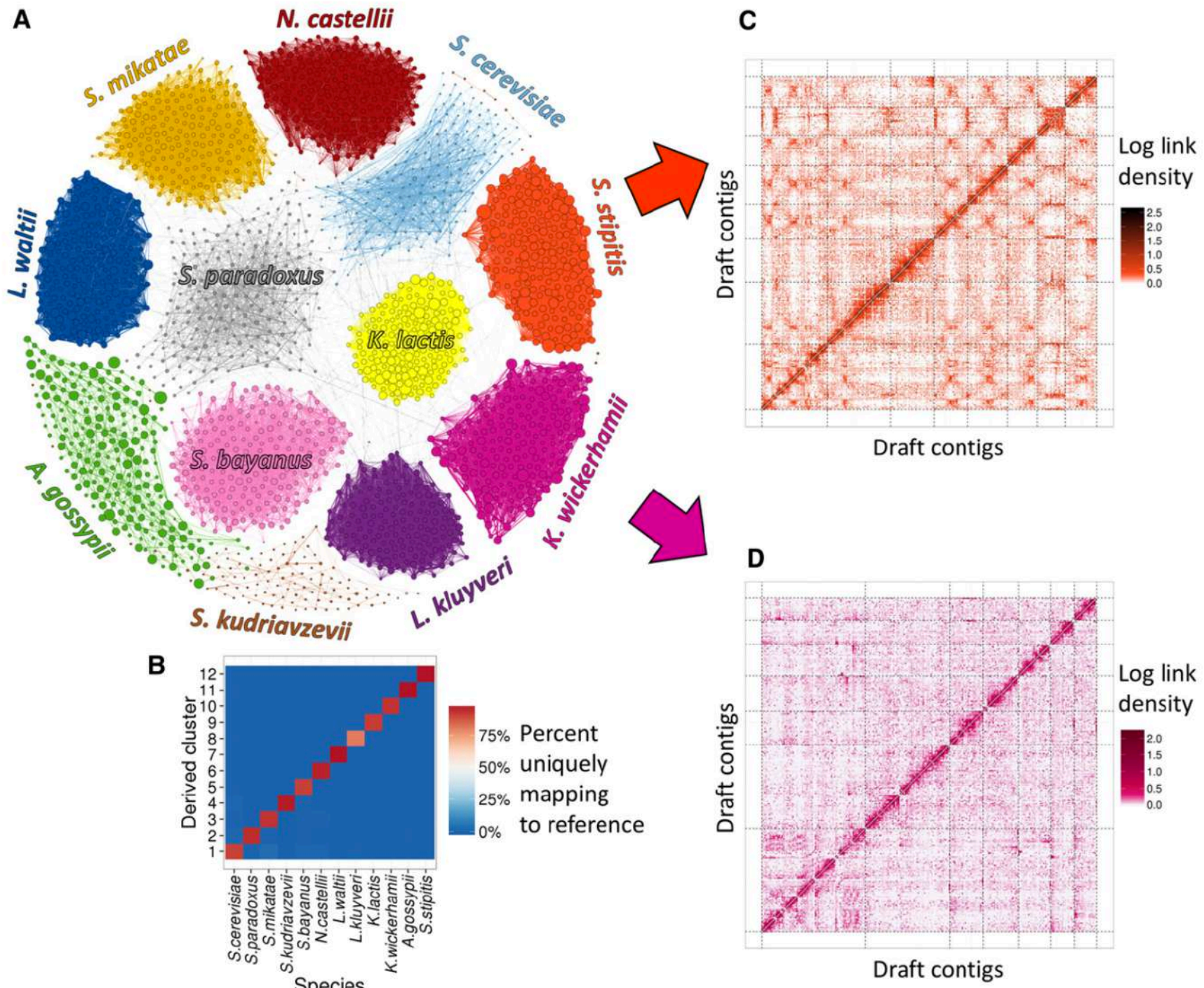


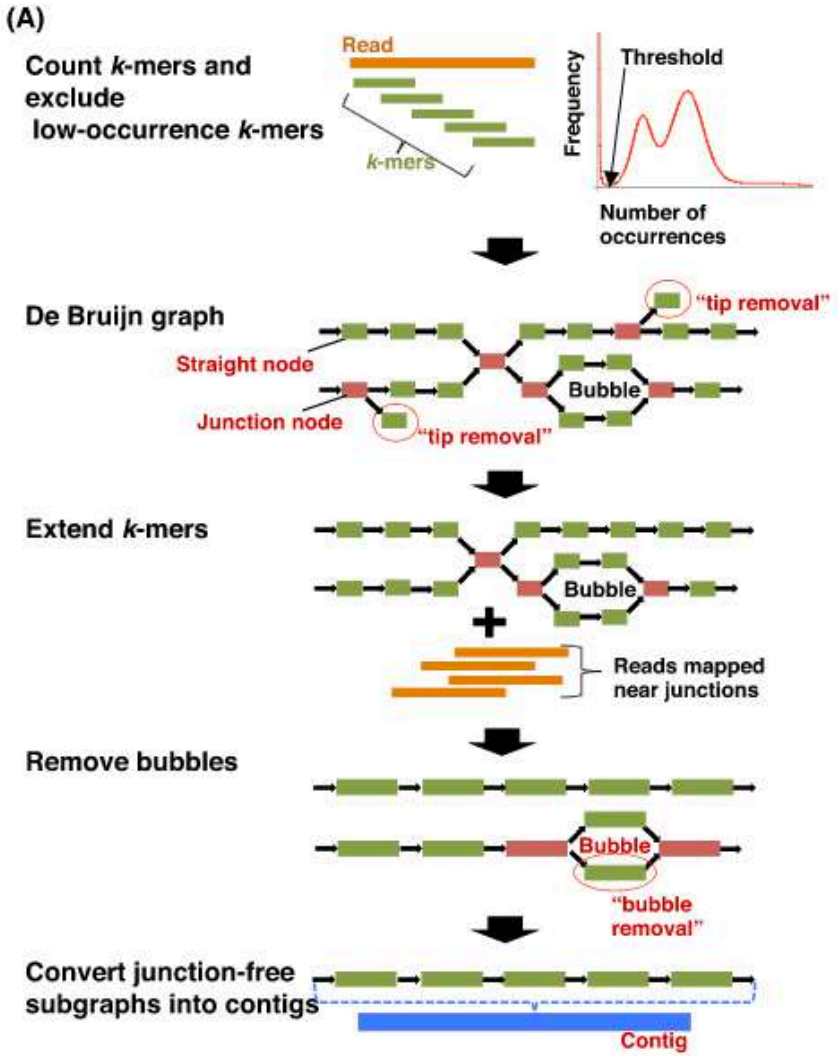
Table 2 Sequencing libraries used in MetaPhase analyses

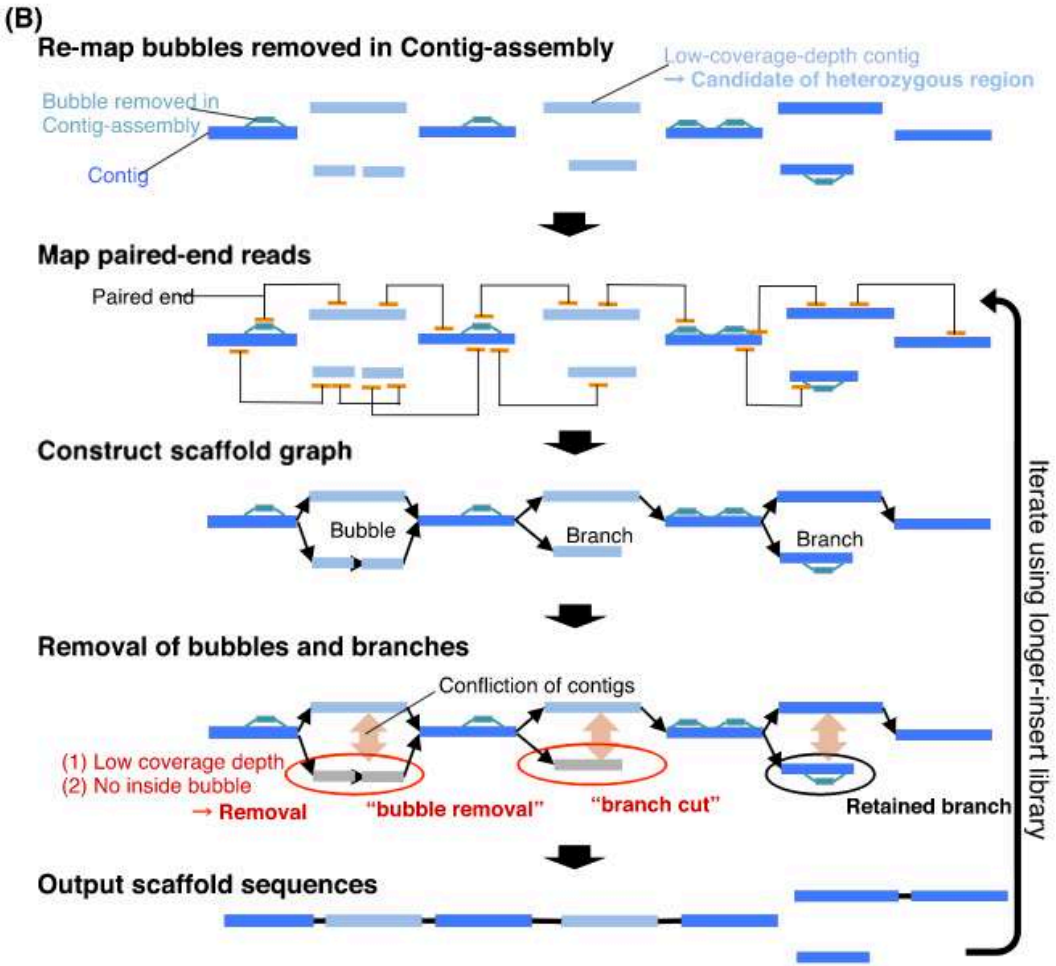
Sample	Library Type	Read Length, bp	Read Pairs, millions
M-Y	Shotgun	101	85.7
	Mate-pair	100	9.2
	Hi-C	100	81.0

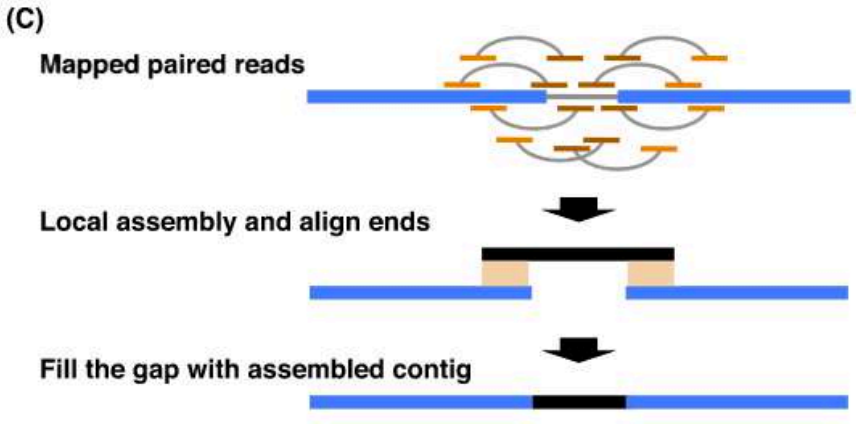

A
N. castellii
S. mikatae
S. cerevisiae
L. waltii
S. paradoxus
S. stipitis
K. lactis
A. gossypii
S. bayanus
K. wickerhamii
L. kluyveri
S. kudriavzevii
B
Derived cluster
12
11
10
9
8
7
6
5
4
3
2
1
S.cerevisiae
S.paradoxus
S.mikatae
S.kudriavzevii
S.bayanus
N.castellii
L.waltii
L.kluyveri
K.lactis
K.wickerhamii
A.gossypii
S.stipitis
Species
75%
50%
25%
0%
Percent uniquely mapping to reference
C
Draft contigs
Draft contigs
Log link density
2.5
2.0
1.5
1.0
0.5
0.0
D
Draft contigs
Draft contigs
Log link density
2.0
1.5
1.0
0.5
0.0

Case studies

Would you understand everything in this paper?







Resource

Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads

Why sequence a genome? (2020 version)

- **Genomics advance our understanding of organisms across tree of life**
  - All previous published genomes that were fragmented are being redone again
    - To reveal greater insights and ease of use for community
  - New genomes are expected to be of good quality

- **Reveal more variations within species**
  - Population genomics is not just **remapping** anymore
  - More accurate inference of structure variation
    - Gene level

- Better analysis power in a genomics world

Lower quality genomes are getting improved

LETTER

OPEN

doi:10.1038/nature22971

Improved maize reference genome with single-molecule technologies

Yinping Jiao[1], Paul Peluso[2], Jinghua Shi[3], Tiffany Liang[3], Michelle C. Stitzer[4], Bo Wang[1], Michael S. Campbell[1], Joshua C. Stein[1], Xuehong Wei[1], Chen-Shan Chin[2], Katherine Guill[5], Michael Regulski[1], Sunita Kumari[1], Andrew Olson[1], Jonathan Gent[6], Kevin L. Schneider[7], Thomas K. Wolfgruber[7], Michael R. May[8], Nathan M. Springer[9], Eric Antoniou[1], W. Richard McCombie[1], Gernot G. Presting[7], Michael McMullen[5], Jeffrey Ross-Ibarra[10], R. Kelly Dawe[6], Alex Hastie[3], David R. Rank[2] & Doreen Ware[1,11]

RESEARCH ARTICLE Open Access

An improved genome assembly uncovers prolific tandem repeats in Atlantic cod

Ole K. Tørresen[1*], Bastiaan Star[1], Sissel Jentoft[1,2], William B. Reinar[1], Harald Grove[3], Jason R. Miller[4], Brian P. Walenz[5], James Knight[6], Jenny M. Ekholm[7], Paul Peluso[7], Rolf B. Edvardsen[8], Ave Tooming-Klunderud[1], Morten Skage[1], Sigbjørn Lien[3], Kjetill S. Jakobsen[1] and Alexander J. Nederbragt[1,9*]

An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing

Aleksey V. Zimin[1,2], Kristian A. Stevens[3], Marc W. Crepeau[3], Daniela Puiu[2], Jill L. Wegrzyn[4], James A. Yorke[1], Charles H. Langley[3], David B. Neale[5] and Steven L. Salzberg[2,6,*]

Research

Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling

Edward S. Rice,[1] Satomi Kohno,[2] John St. John,[3] Son Pham,[4] Jonathan Howard,[5] Liana F. Lareau,[6] Brendan L. O'Connell,[1,7] Glenn Hickey,[1] Joel Armstrong,[1] Alden Deran,[1] Ian Fiddes,[1] Roy N. Platt II,[8] Cathy Gresham,[9] Fiona McCarthy,[10] Colin Kern,[11] David Haan,[1] Tan Phan,[12] Carl Schmidt,[13] Jeremy R. Sanford,[14] David A. Ray,[8] Benedict Paten,[15] Louis J. Guillette Jr.,[16,†] and Richard E. Green[1,6,7]

Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity

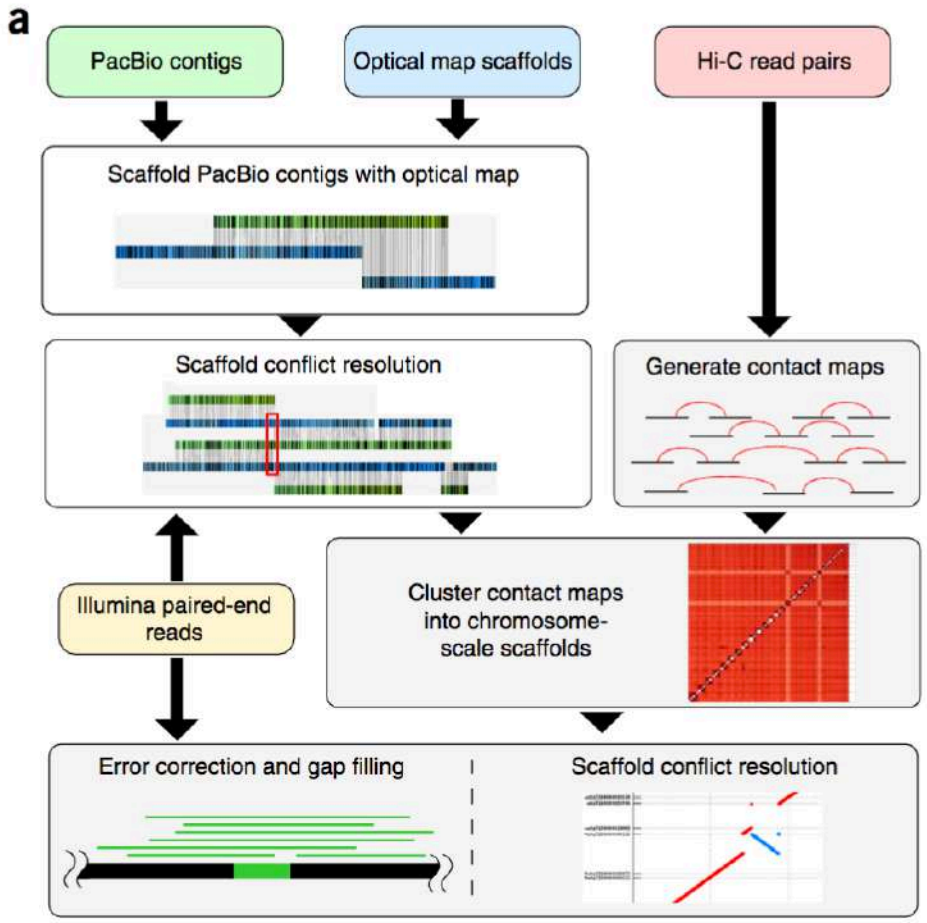
Patrick P. Edger[1,2,*,†], Robert VanBuren[1,†], Marivi Colle[1], Thomas J. Poorten[3], Ching Man Wai[1], Chad E. Niederhuth[4], Elizabeth I. Alger[1], Shujun Ou[1,2], Charlotte B. Acharya[3], Jie Wang[5], Pete Callow[1], Michael R. McKain[6], Jinghua Shi[7], Chad Collier[7], Zhiyong Xiong[8], Jeffrey P. Mower[9], Janet P. Slovin[10], Timo Hytönen[11], Ning Jiang[1,2], Kevin L. Childs[5,12] and Steven J. Knapp[3,*]

An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations

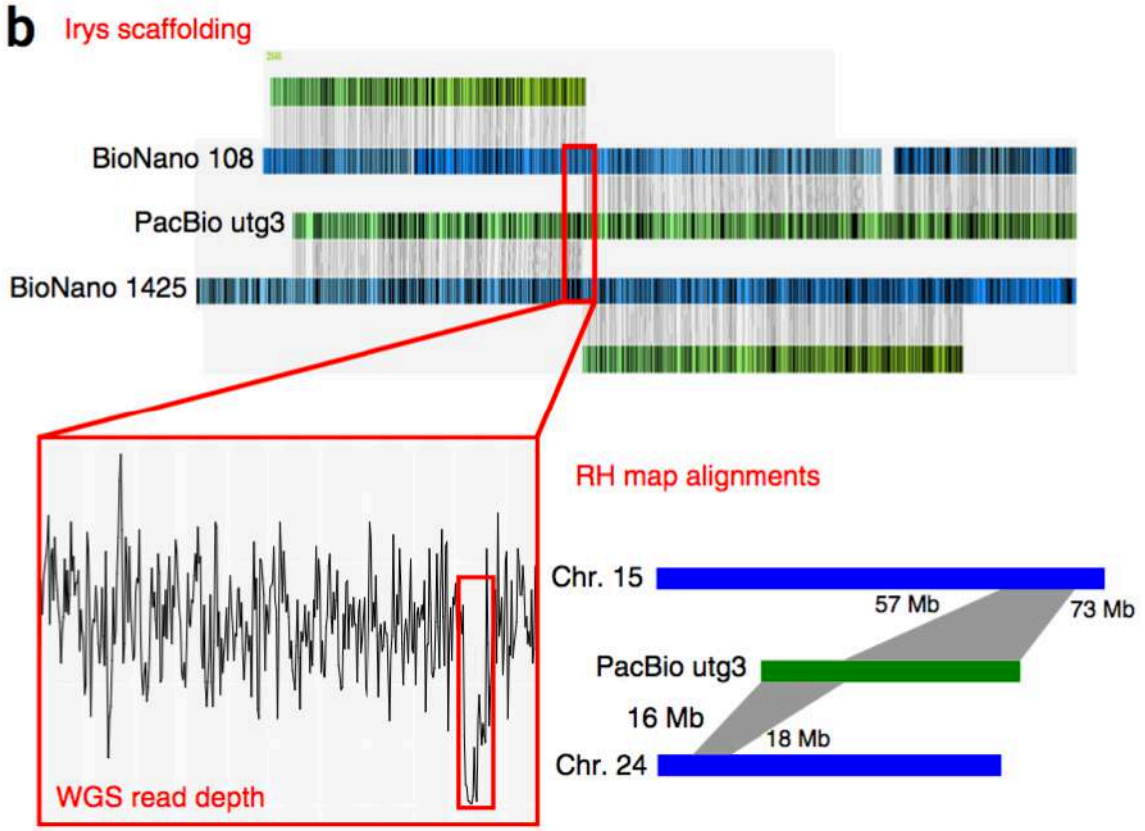
Bernardo J. Clavijo,[1,9] Luca Venturini,[1,9] Christian Schudoma,[1] Gonzalo Garcia Accinelli,[1] Gemy Kaithakottil,[1] Jonathan Wright,[1] Philippa Borrill,[2] George Kettleborough,[1] Darren Heavens,[1] Helen Chapman,[1] James Lipscombe,[1] Tom Barker,[1] Fu-Hao Lu,[2] Neil McKenzie,[2] Dina Raats,[1] Ricardo H. Ramirez-Gonzalez,[1,2] Aurore Coince,[1] Ned Peel,[1] Lawrence Percival-Alwyn,[1] Owen Duncan,[3] Josua Trösch,[3] Guotai Yu,[2] Dan M. Bolser,[4] Guy Namaati,[4] Arnaud Kerhornou,[4] Manuel Spannagl,[5] Heidrun Gundlach,[5] Georg Haberer,[5] Robert P. Davey,[1,6] Christine Fosker,[1] Federica Di Palma,[1,6] Andrew L. Phillips,[7] A. Harvey Millar,[3] Paul J. Kersey,[4] Cristobal Uauy,[2] Ksenia V. Krasileva,[1,6,8] David Swarbreck,[1,6] Michael W. Bevan,[2] and Matthew D. Clark[1,6]

Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome

Derek M Bickhart[1,18], Benjamin D Rosen[2,18], Sergey Koren[3,18], Brian L Sayre[4], Alex R Hastie[5], Saki Chan[5], Joyce Lee[5], Ernest T Lam[5], Ivan Liachko[6], Shawn T Sullivan[7], Joshua N Burton[6], Heather J Huson[8], John C Nystrom[8], Christy M Kelley[9], Jana L Hutchison[2], Yang Zhou[2,10], Jiajie Sun[11], Alessandra Crisà[12], F Abel Ponce de León[13], John C Schwartz[14], John A Hammond[14], Geoffrey C Waldbieser[15], Steven G Schroeder[2], George E Liu[2], Maitreya J Dunham[6], Jay Shendure[6,16], Tad S Sonstegard[17], Adam M Phillippy[3], Curtis P Van Tassell[2] & Timothy P L Smith[9]



New genomes are expected to have high quality



Correcting misassemblies

… ARS1 comprises just 31 scaffolds and 649 gaps covering 30 of the 31 haploid, acrocentric goat chromosomes (excluding only the Y chromosome),

Table 1 Assembly statistics

Assembly[a]	Contigs[b]	Scaffolds	Unplaced contigs[c]	Degenerate contigs[d]	Contig NG50 (Mb)[e]	Scaffold NG50 (Mb)[e,f]	Assembly size (Gb)	Assembly in scaffolds (%)
PacBio	3,074	–	–	30,693	3.795	–	2.914	N/A
Optical Map	–	2,944	–	–	–	1.487	2.748	N/A
PacBio + Optical Map	1,109	333	1,242	30,693	10.197	20.623	2.910	90.89
PacBio + Hi-C	2,115	31	959	30,693	3.795	88.799	2.910	87.97
PacBio + Optical Map + Hi-C	1,780	31	571	30,693	10.197	87.347	2.910	89.05
ARS1	680	31	654	29,315	18.702	87.277	2.924	88.32

[a]Assemblies are listed in order of inclusion of scaffolding technologies toward the final assembly (ARS1), with the original contigs (PacBio) scaffolded using different technologies (Optical Map and Hi-C). Because the optical map program (Irys Scaffold) generates an assembly from the consensus of labeled DNA molecules, we have included scaffold statistics from these data (optical map) for comparison. [b]The number of continuous stretches of sequence within the scaffold without gaps >3 bases in length of at least 10 bases. [c]Unplaced contigs are defined as input contigs or scaffolds that were not placed by the optical map or Hi-C in a scaffold were excluded from the scaffold counts. [d]Degenerate contigs were assembled unitigs that had less than 50 PacBio reads supporting their assembly (**Supplementary Note**). Differences in degenerate contig counts in the final ARS1 assembly are due to PBJelly merging of degenerate contigs (538 contigs) or removal due to no supporting PacBio read alignments (840). [e]All NG50 values are based on the ARS1 assembly size (2.924 Gb). [f]No scaffolds were generated for the PacBio entry.

A golden goat genome

Kim C Worley

The newly described *de novo* goat genome sequence is the most contiguous diploid vertebrate assembly generated thus far using whole-genome assembly and scaffolding methods. The contiguity of this assembly is approaching that of the finished human and mouse genomes and suggests an affordable roadmap to high-quality references for thousands of species.

…. This report generates sighs of relief from researchers frustrated with the highly fragmented genome sequences available for most species….

… The lower costs and greater accessibility of these methods bring potential for wider impact….

a

Most genomes in this N50 range

Contig N50

Genomes with highest N50

Number of vertebrate genomes		Less than 10 kb	10 kb to 100 kb	100 kb to 1 Mb	1 Mb to 10 Mb	Greater than 10 Mb
Diploid human	21	–	8	6	3	4
Non-human mammal	196	34	34	14	1	3
Non-mammal	193	43	133	14	3	0
Total	410	77	285	34	7	7

b

	Goat CHIR_1.0	Goat ARS1	Human GRCh38
Total sequence length	2.6 Gb	2.9 Gb	3.2 Gb
Total assembly gap length	140 Mb	38 Mb	160 Mb
Gaps between scaffolds	411	0	349
Number of scaffolds	77,431	29,907	735
Scaffold N50	14 Mb	87 Mb	67 Mb
Number of contigs	337,494	30,399	1,385
Contig N50	18.9 kb	26.2 Mb	56.4 Mb
Number of chromosomes and plasmids	30	31	25

c

Difficult regions are being targeted

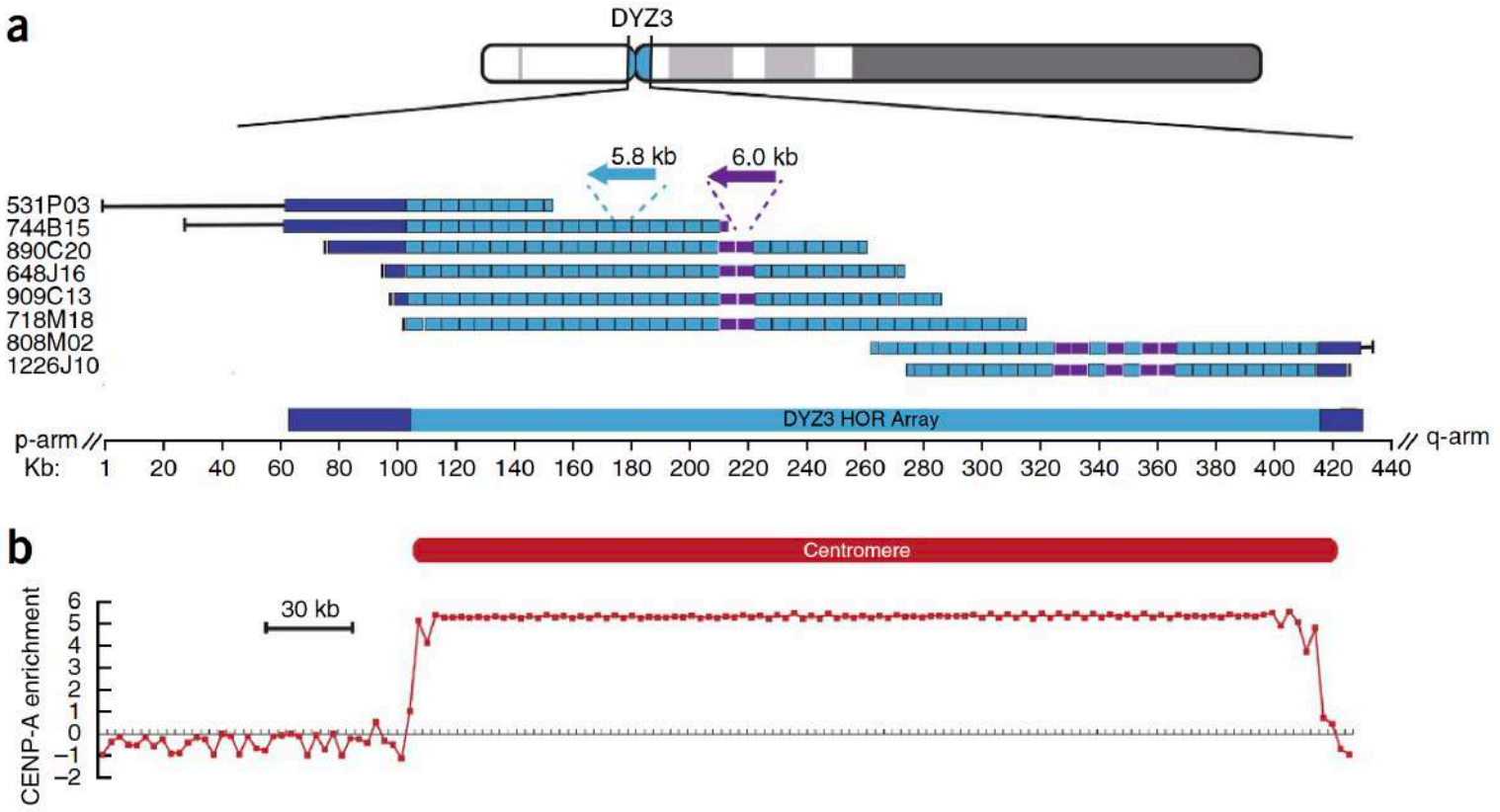
nature biotechnology

OPEN

Linear assembly of a human centromere on the Y chromosome

Miten Jain[1,5], Hugh E Olsen[1,5], Daniel J Turner[2], David Stoddart[2], Kira V Bulazel[3], Benedict Paten[1], David Haussler[1], Huntington F Willard[3,4], Mark Akeson[1] & Karen H Miga[1,3]

The human genome reference sequence remains incomplete owing to the challenge of assembling long tracts of near-identical tandem repeats in centromeres. We implemented a nanopore sequencing strategy to generate high-quality reads that span hundreds of kilobases of highly repetitive DNA in a human Y chromosome centromere. Combining these data with short-read variant validation, we assembled and characterized the centromeric region of a human Y chromosome.



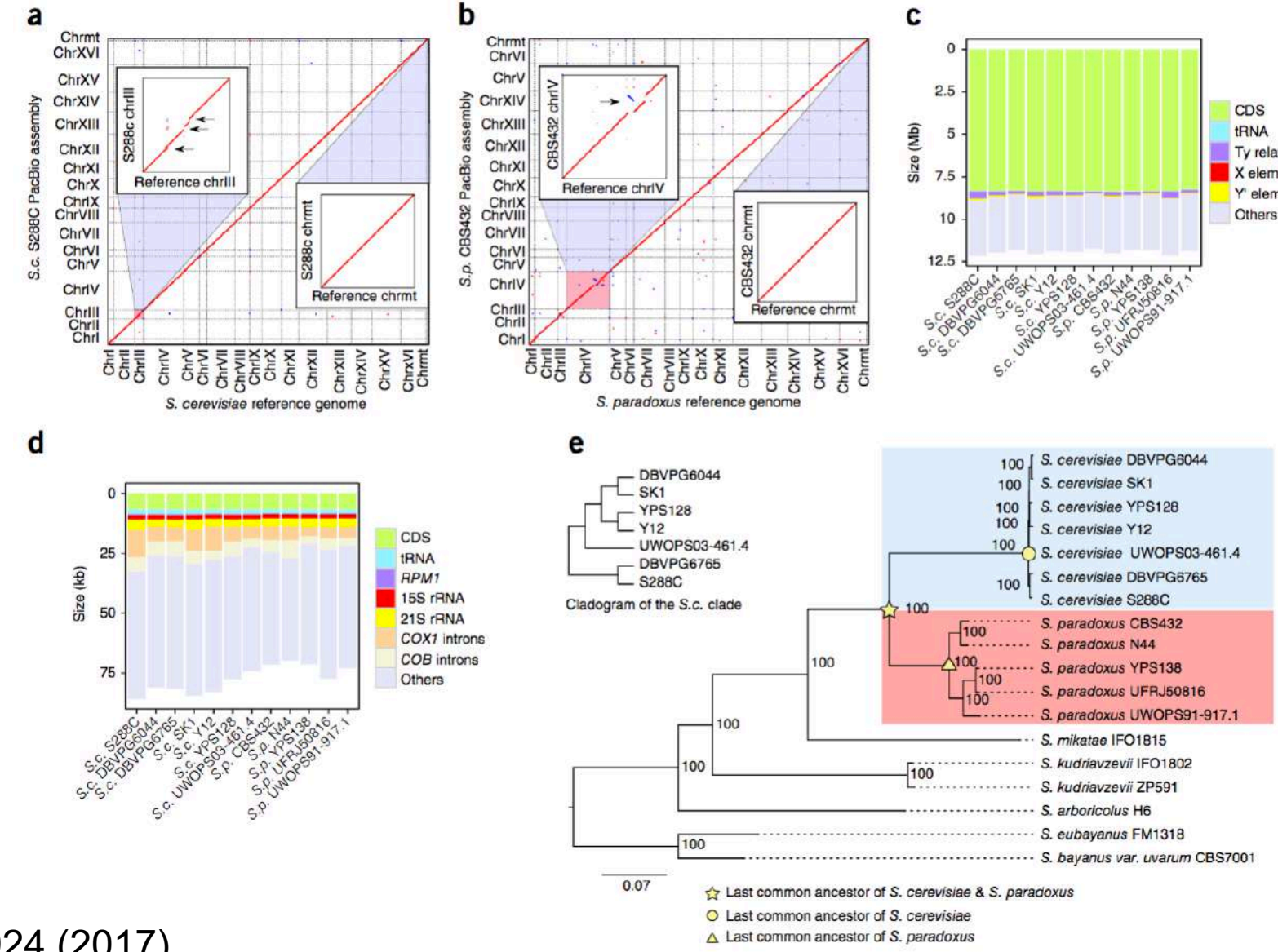
Jain *et al* (2018) Nature Biotechnology

Contrasting evolutionary genome dynamics between domesticated and wild yeasts

Jia-Xing Yue[1], Jing Li[1], Louise Aigrain[2], Johan Hallin[1], Karl Persson[3], Karen Oliver[2], Anders Bergström[2], Paul Coupland[2,5], Jonas Warringer[3], Marco Cosentino Lagomarsino[4], Gilles Fischer[4], Richard Durbin[2] & Gianni Liti[1]

Population genomics

- long-read sequencing to generate **end-to-end genome assemblies** for **12 strains** representing major subpopulations of the partially domesticated yeast *Saccharomyces cerevisiae* and its wild relative *S. paradoxus*.



Nature Genetics **volume 49**, pages 913–924 (2017)

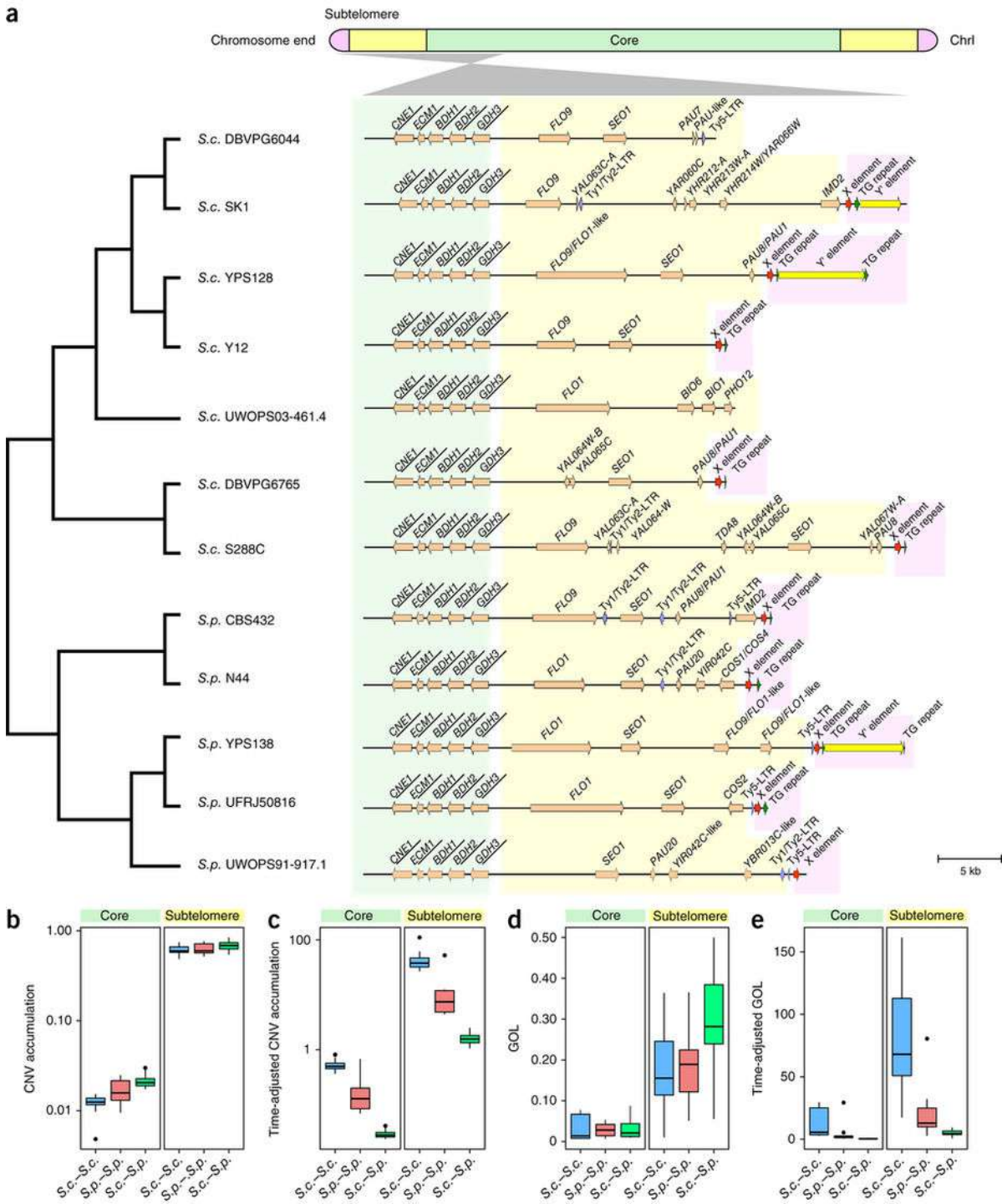
Contrasting evolutionary genome dynamics between domesticated and wild yeasts

Jia-Xing Yue[1], Jing Li[1], Louise Aigrain[2], Johan Hallin[1], Karl Persson[3], Karen Oliver[2], Anders Bergström[2], Paul Coupland[2,5], Jonas Warringer[3], Marco Cosentino Lagomarsino[4], Gilles Fischer[4], Richard Durbin[2] & Gianni Liti[1]

- enable precise definition of chromosomal boundaries between cores and subtelomeres
- *S. paradoxus* shows faster accumulation of balanced rearrangements (inversions, reciprocal translocations and transpositions), *S. cerevisiae* accumulates unbalanced rearrangements (novel insertions, deletions and duplications) more rapidly.

- Such striking contrasts between wild and domesticated yeasts are likely to reflect the influence of human activities on structural genome evolution.

Nature Genetics **volume 49**, pages 913–924 (2017)



LETTER

OPEN

doi:10.1038/nature23264

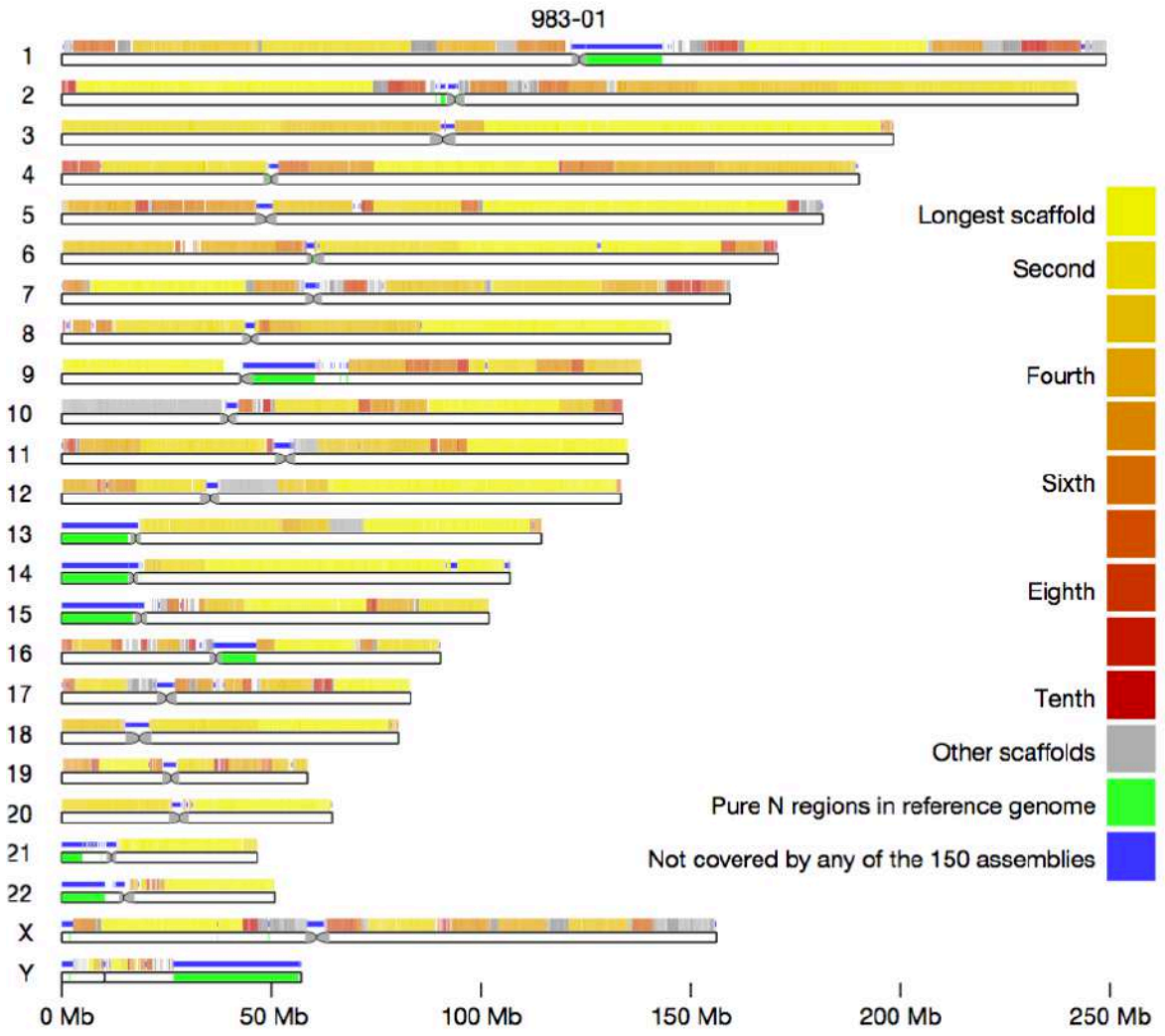
Sequencing and *de novo* assembly of 150 genomes from Denmark as a population reference

Lasse Maretty[1]*, Jacob Malte Jensen[2,3]*, Bent Petersen[4]*, Jonas Andreas Sibbesen[1]*, Siyang Liu[1,5]*, Palle Villesen[2,3,6]*, Laurits Skov[2,3]*, Kirstine Belling[4]*, Christian Theil Have[7], Jose M. G. Izarzugaza[4], Marie Grosjean[4], Jette Bork-Jensen[7], Jakob Grove[3,8,9], Thomas D. Als[3,8,9], Shujia Huang[10,11], Yuqi Chang[10], Ruiqi Xu[5], Weijian Ye[5], Junhua Rao[5], Xiaosen Guo[10,12], Jihua Sun[5,7], Hongzhi Cao[10], Chen Ye[10], Johan van Beusekom[4], Thomas Espeseth[13,14], Esben Flindt[12], Rune M. Friborg[2,3], Anders E. Halager[2,3], Stephanie Le Hellard[14,15], Christina M. Hultman[16], Francesco Lescai[3,8,9], Shengting Li[3,8,9], Ole Lund[4], Peter Løngren[4], Thomas Mailund[2,3], Maria Luisa Matey-Hernandez[4], Ole Mors[3,6,9], Christian N. S. Pedersen[2,3], Thomas Sicheritz-Pontén[4], Patrick Sullivan[16,17], Ali Syed[4], David Westergaard[4], Rachita Yadav[4], Ning Li[5], Xun Xu[10], Torben Hansen[7], Anders Krogh[1], Lars Bolund[8,10], Thorkild I. A. Sørensen[7,18,19], Oluf Pedersen[7], Ramneek Gupta[4], Simon Rasmussen[4]§, Søren Besenbacher[2,6]§, Anders D. Børglum[3,8,9]§, Jun Wang[3,10,12]§, Hans Eiberg[20]§, Karsten Kristiansen[10,12]§, Søren Brunak[4,21]§ & Mikkel Heide Schierup[2,3,22]§

- Hundreds of thousands of human genomes are now being sequenced to characterize genetic variation and use this information to augment association mapping studies of complex disorders and other phenotypic traits.

- Genetic variation is identified mainly by mapping short reads to the reference genome. However, these approaches are biased against discovery of structural variants and variation in the more complex parts of the genome.

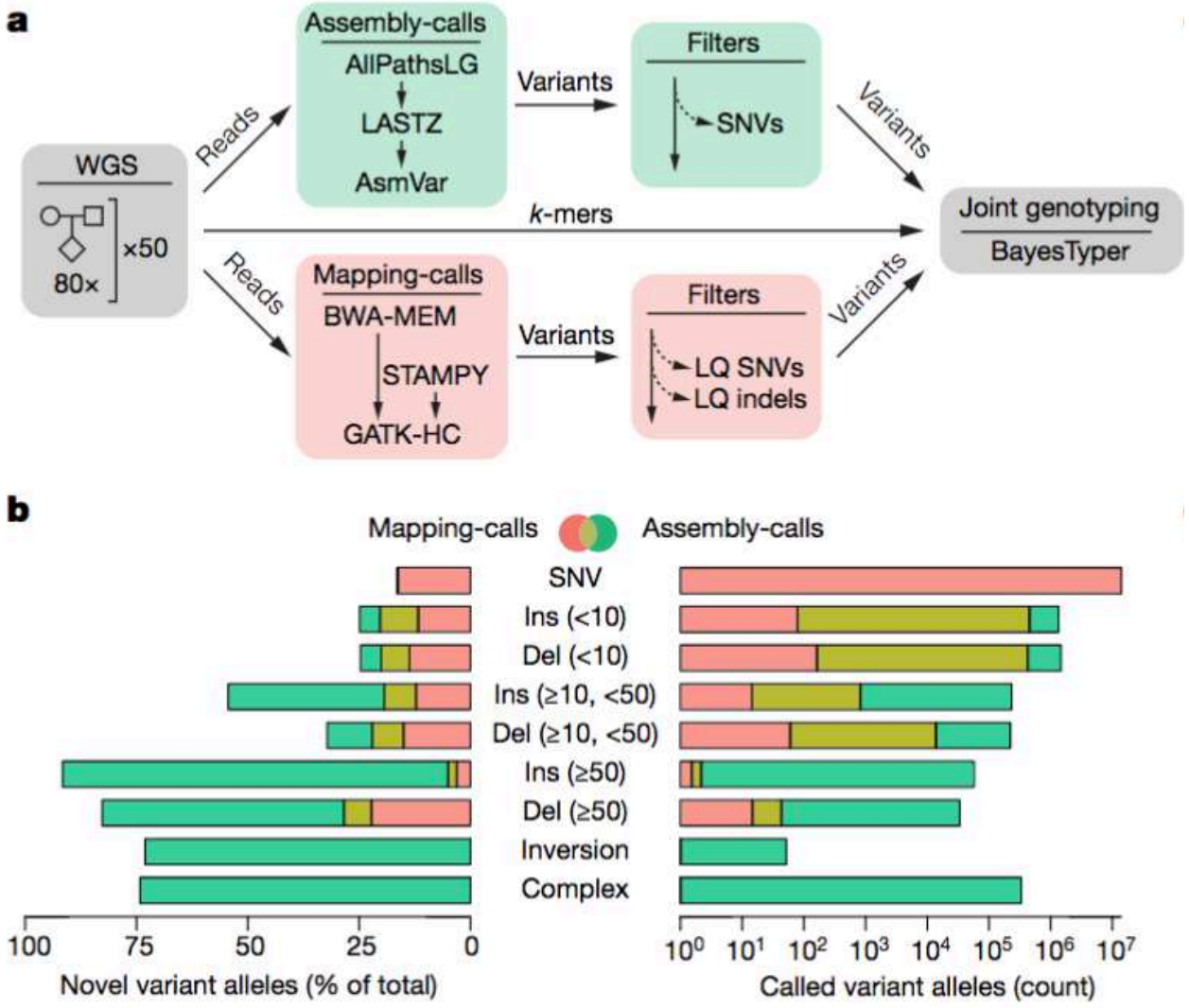
- report *de novo* assemblies of 150 individuals (50 trios) from the GenomeDenmark project.



We found that 16.4% of the called SNVs were novel (not in the Single Nucleotide Polymorphism database 142 (dbSNP142) or 1000 Genomes Project phase 3 structural variant call-set), **whereas as many as 91.6% of insertions ≥50 bp were novel** (Fig. 2b).

The fraction of novel variants increased rapidly with variant length, especially for insertions (Fig. 2d), with **most longer variants contributed by the assembly-based approach…**

For instance, …we **called 33,653 deletions ≥50 bp**, whereas the **1000 Genomes Project identified 42,279 such variants in 25 times more individuals who were more diverse than our study population**.



ARTICLE

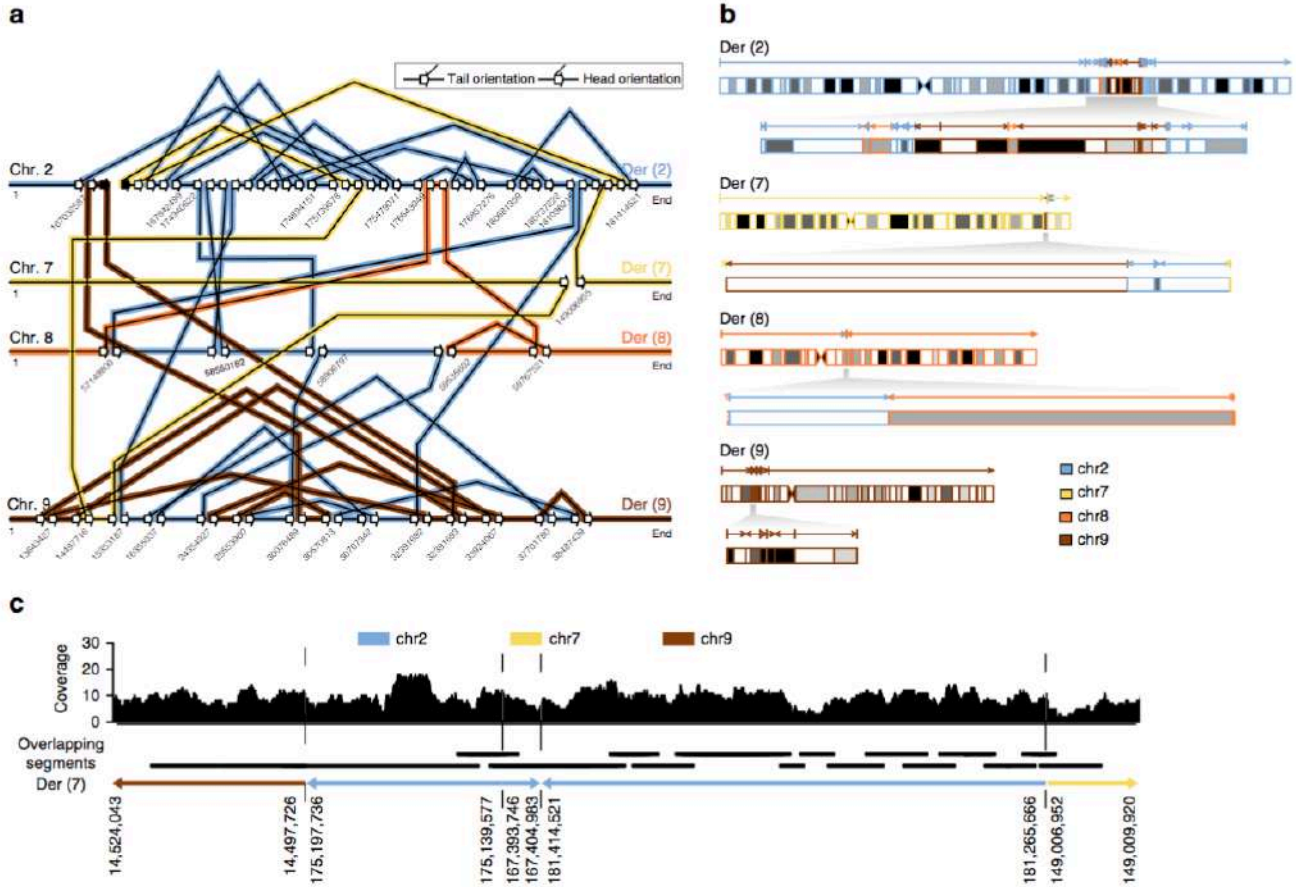
DOI: 10.1038/s41467-017-01343-4 OPEN

Mapping and phasing of structural variation in patient genomes using nanopore sequencing

Mircea Cretu Stancu[1], Markus J. van Roosmalen[1], Ivo Renkens[1], Marleen M. Nieboer[1], Sjors Middelkamp[1], Joep de Ligt[1], Giulia Pregno[2], Daniela Giachino[2], Giorgia Mandrile[2], Jose Espejo Valle-Inclan[1], Jerome Korzelius[1], Ewart de Bruijn[1], Edwin Cuppen[3], Michael E. Talkowski[4,5,6], Tobias Marschall[7,8], Jeroen de Ridder[1] & Wigard P. Kloosterman[1]

Long reads lead to better SV

- long reads are superior to short reads with regard to detection of de novo chromothripsis rearrangements.
- long reads also enable efficient phasing of genetic variations, which we leveraged to determine the parental origin of all de novo chromothripsis breakpoints and to resolve the structure of these complex rearrangements.



Example of metagenomics

Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen

Matthias Hess,[1,2]* Alexander Sczyrba,[1,2]* Rob Egan,[1,2] Tae-Wan Kim,[3] Harshal Chokhawala,[3] Gary Schroth,[4] Shujun Luo,[4] Douglas S. Clark,[3,5] Feng Chen,[1,2] Tao Zhang,[1,2] Roderick I. Mackie,[6] Len A. Pennacchio,[1,2] Susannah G. Tringe,[1,2] Axel Visel,[1,2] Tanja Woyke,[1,2] Zhong Wang,[1,2] Edward M. Rubin[1,2]†

- 268Gb of metagenomics data
- Identified 27,755 putative carbohydrate-active genes from a cow rumen metagenome
- Expressed 90 candidates of which 57% had enzymatic activity against cellulosic substrates
- Assembled 15 uncultured microbial genomes

Hess *et al*., 2011 **Science**

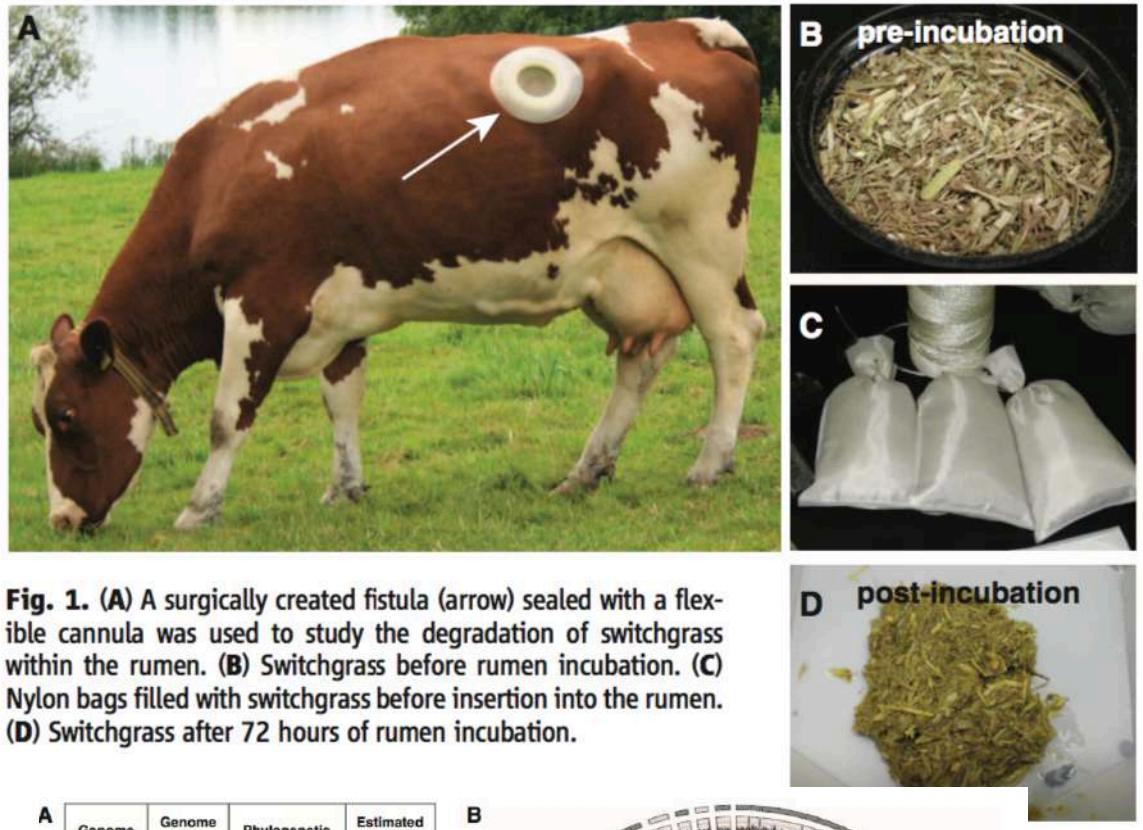
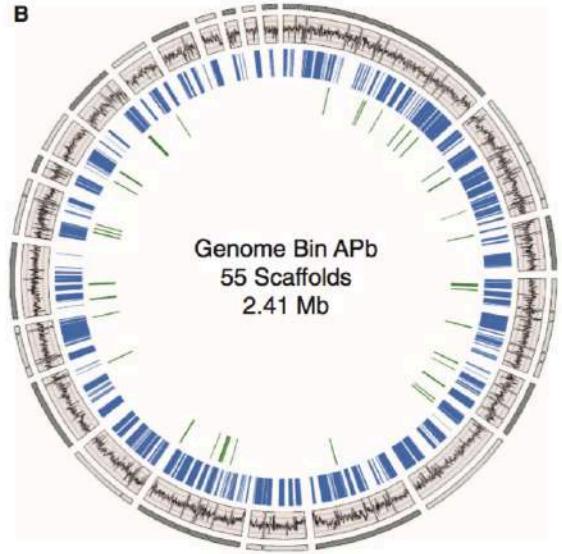


Fig. 1. (**A**) A surgically created fistula (arrow) sealed with a flexible cannula was used to study the degradation of switchgrass within the rumen. (**B**) Switchgrass before rumen incubation. (**C**) Nylon bags filled with switchgrass before insertion into the rumen. (**D**) Switchgrass after 72 hours of rumen incubation.

A

Genome Bin	Genome Size (Mb)	Phylogenetic Order	Estimated Completeness
AFa	2.87	Spirochaetales	92.98%
AMa	2.21	Spirochaetales	91.23%
Ala	2.53	Clostridiales	90.10%
AGa	3.08	Bacteroidales	89.77%
AN	2.02	Clostridiales	78.50%
AJ	2.24	Bacteroidales	75.96%
AC2a	2.07	Bacteroidales	75.96%
AWa	2.02	Clostridiales	75.77%
AH	2.52	Bacteroidales	75.45%
AQ	1.91	Bacteroidales	71.36%
AS1a	1.75	Clostridiales	70.99%
APb	2.41	Clostridiales	64.85%
BOa	1.67	Clostridiales	64.16%
ADa	2.99	Myxococcales	62.13%
ATa	1.87	Clostridiales	60.41%

B



enome; innermost circle (green tick marks), location of glycoside hydrolase genes on draft genome.

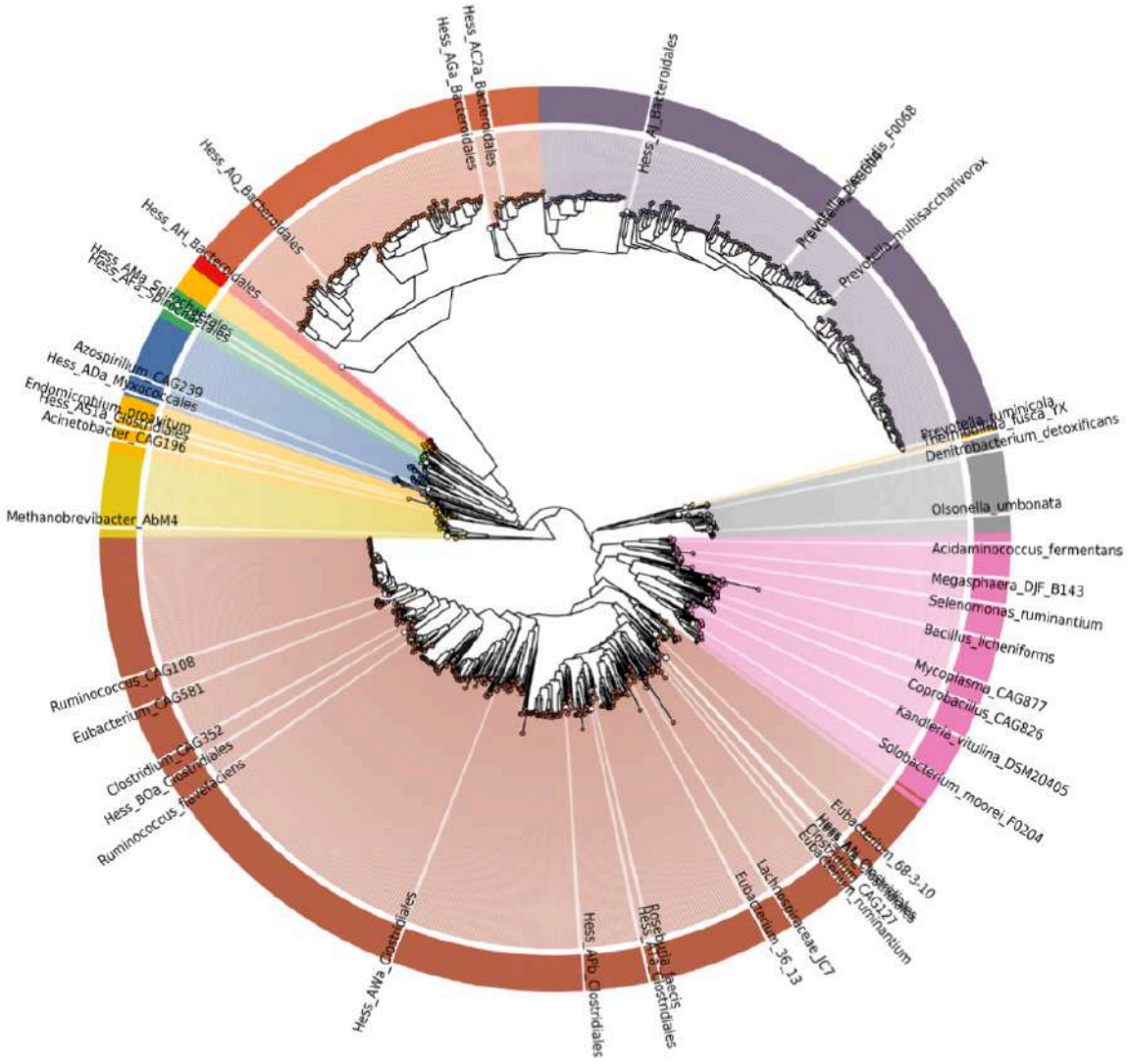
ARTICLE

DOI: 10.1038/s41467-018-03317-6 OPEN

Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen

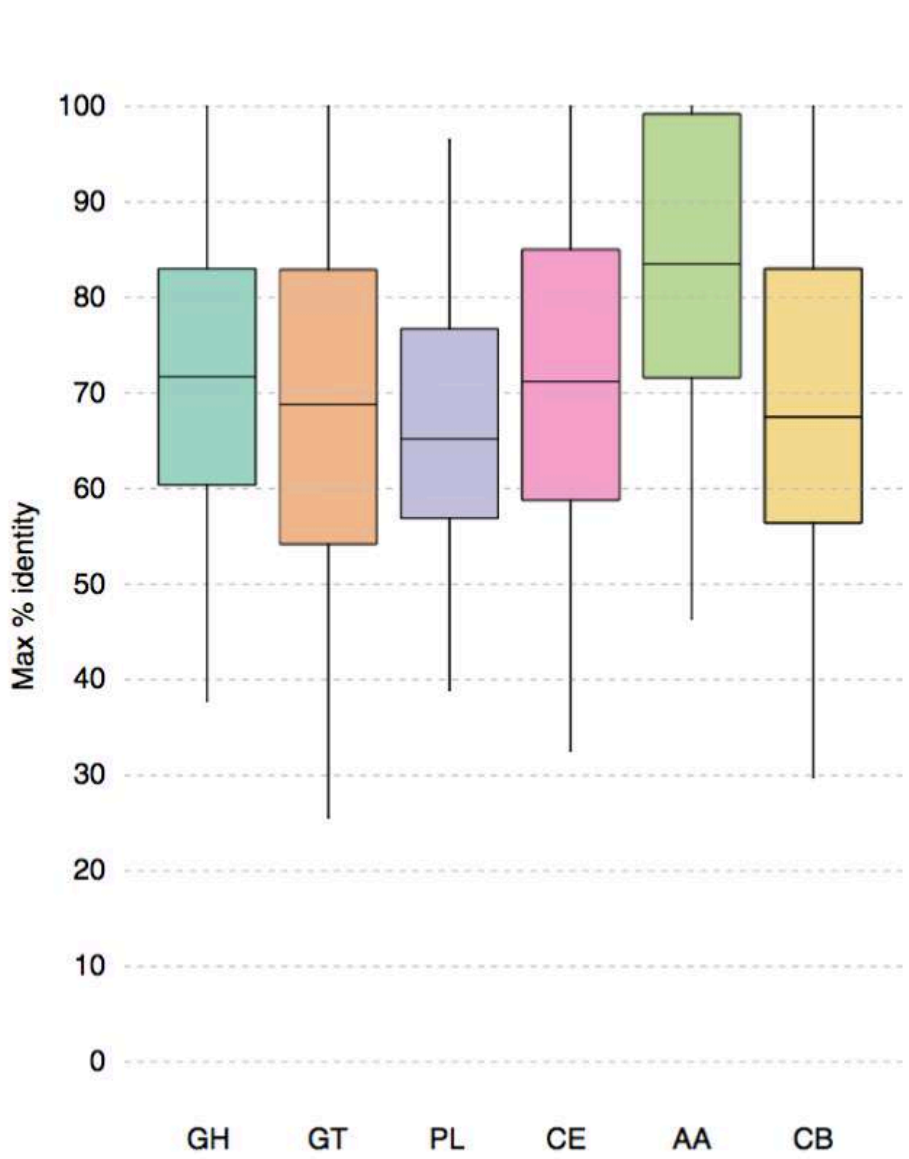
Robert D. Stewart[1], Marc D. Auffret[2], Amanda Warr[1], Andrew H. Wiser[3], Maximilian O. Press[3], Kyle W. Langford[3], Ivan Liachko[3], Timothy J. Snelling[4], Richard J. Dewhurst[2], Alan W. Walker[4], Rainer Roehe[2] & Mick Watson[1]

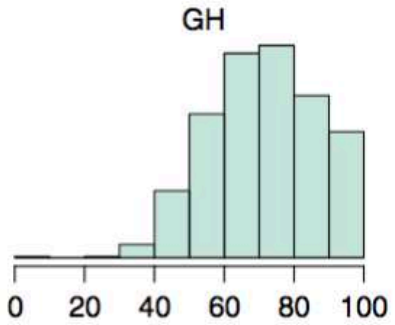
- 800 Gb of sequence data derived from 43 Scottish cattle
- Using metagenomic binning and Hi-C techniques
- Assembly of 913 draft bacterial and archaeal genomes
- Most of these genomes represent previously unsequenced strains and species.
- The draft genomes contain over 69,000 proteins predicted to be involved in carbohydrate metabolism, over 90% of which do not have a good match in public databases.

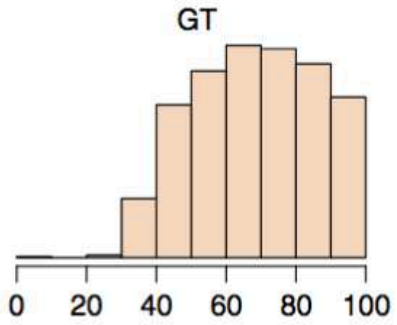


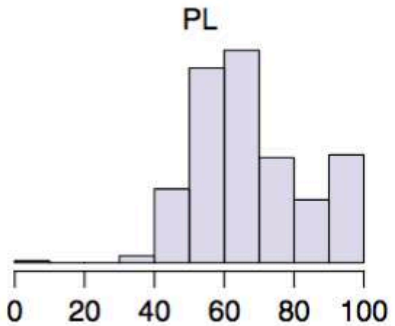
Stewart *et al* (2018) Nature Communications

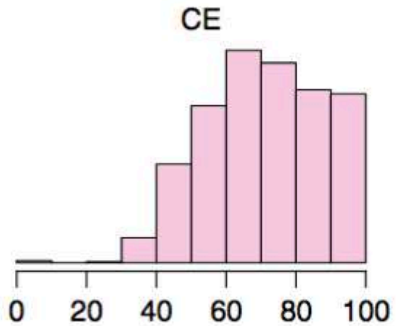
- The draft genomes contain over 69,000 proteins predicted to be involved in carbohydrate metabolism, over 90% of which do not have a good match in public databases.

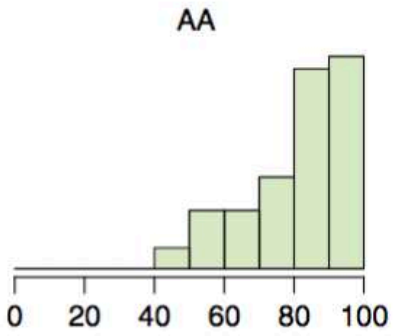


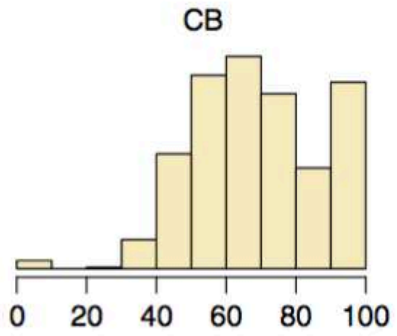












- Inclusion of the 913 genomes presented here improves metagenomic read classification by sevenfold against the study's own data, and by fivefold against other publicly available rumen datasets.

- dataset substantially improves the coverage of rumen microbial genomes in the public databases and represents a valuable resource for biomass-degrading enzyme discovery and studies of the rumen microbiome

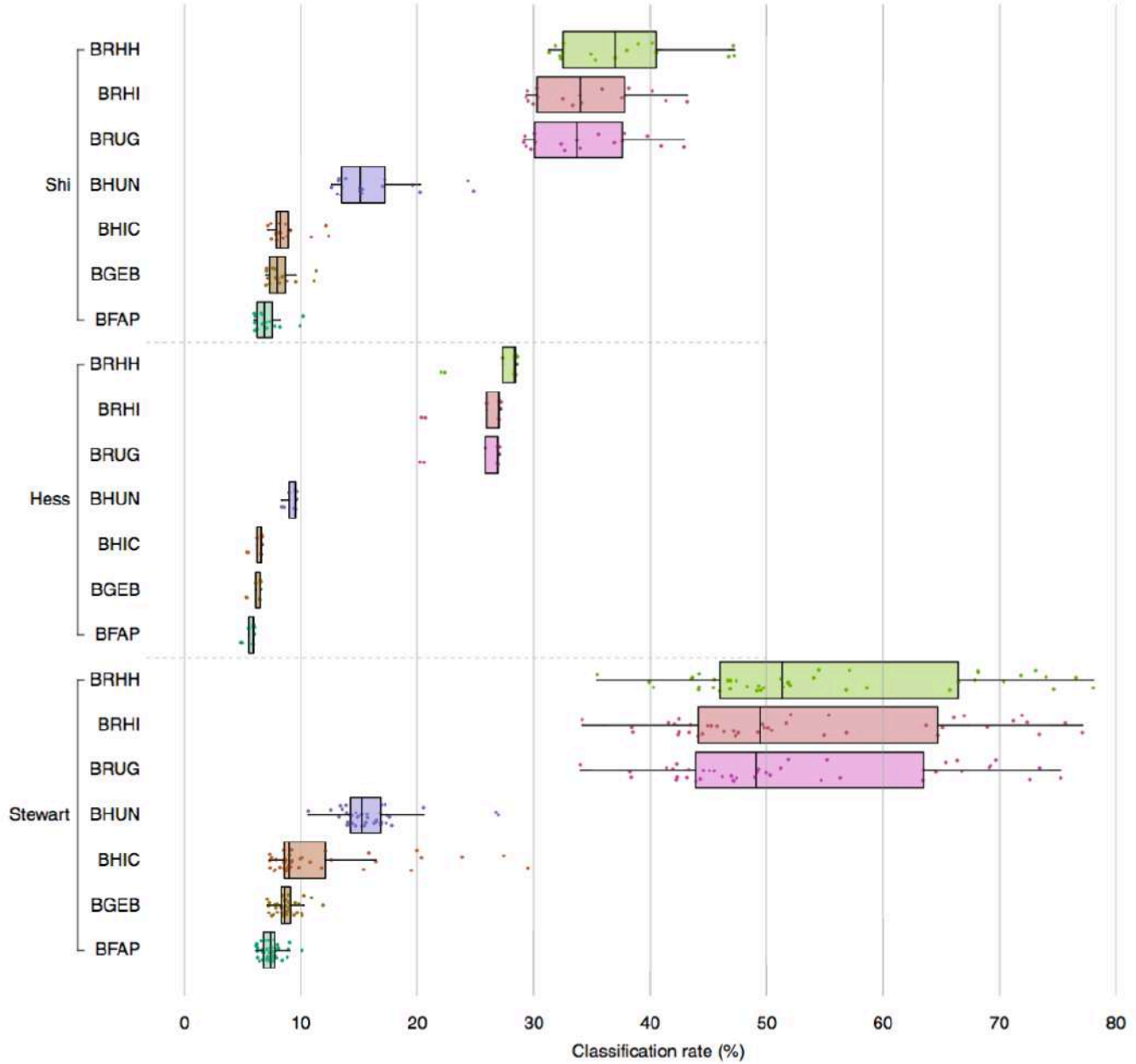


Fig. 4 Classification rate for three datasets against various Kraken databases. BFAP bacterial, archaeal, fungal and protozoan genomes from RefSeq, BGEB BFAP + 1003 GEBA genomes, BHIC BFAP + 63 hRUG genomes, BHUN BFAP + 410 genomes from the Hungate 1000 project, BRUG BFAP + 850 RUG MAGs, BRHI BFAP + all 913 genomes from this study, BRHH BFAP + 913 RUGs + 410 Hungate 1000 genomes. Addition of rumen-specific RUGs or Hungate 1000 genomes has the most dramatic effect

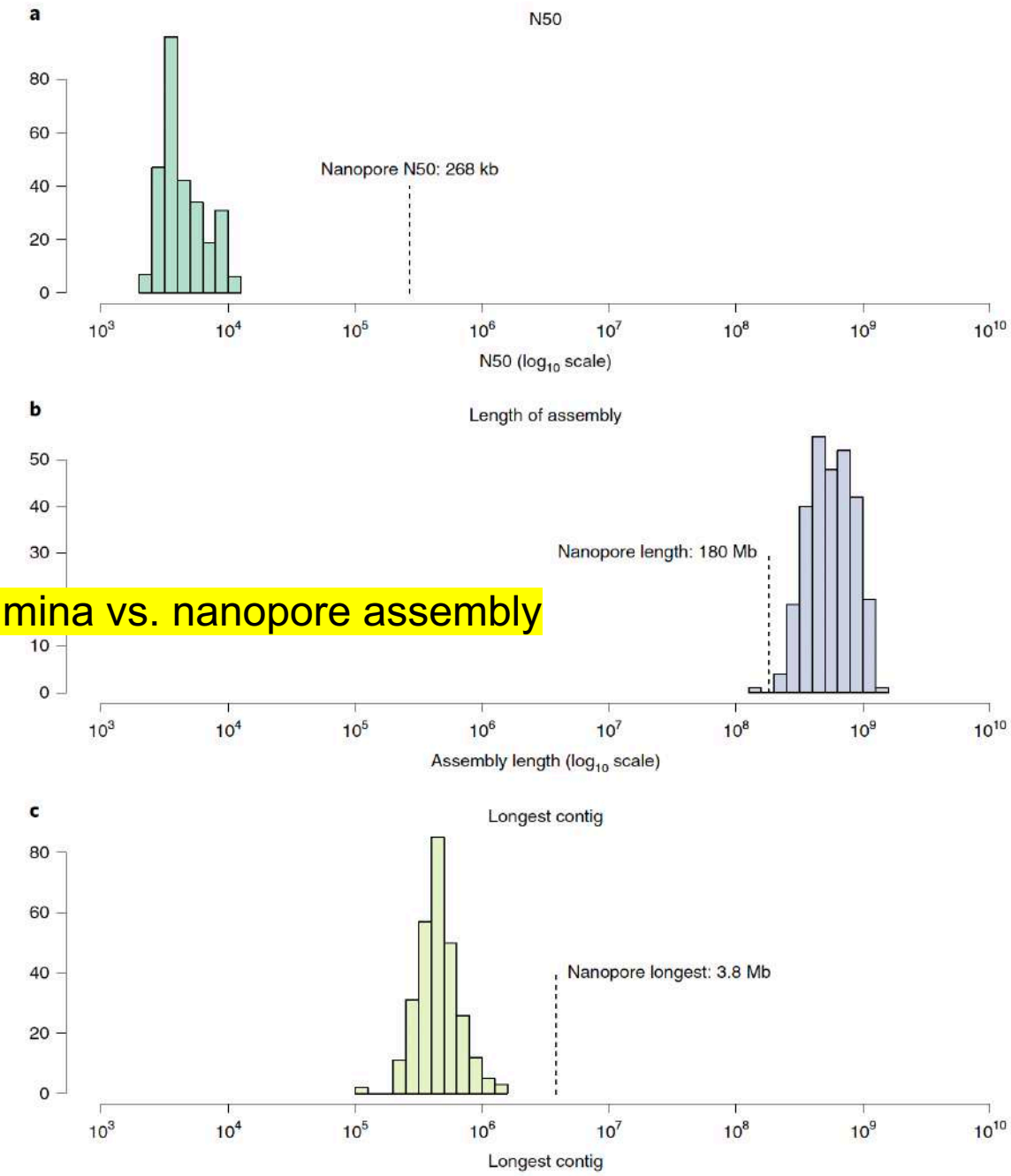
Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery

Robert D. Stewart[1], Marc D. Auffret[2], Amanda Warr[1], Alan W. Walker[3], Rainer Roehe[2] and Mick Watson[1*]

- **6.5 Tb** of sequence data derived from **283** ruminant cattles
- Using metagenomic binning and Hi-C techniques

- Assembly of **4,941** draft bacterial and archaeal genomes

- Long read is being used: “We also present a metagenomic assembly of nanopore (MinION) sequencing data (from one rumen sample) that contains at least three whole bacterial chromosomes as single contigs”



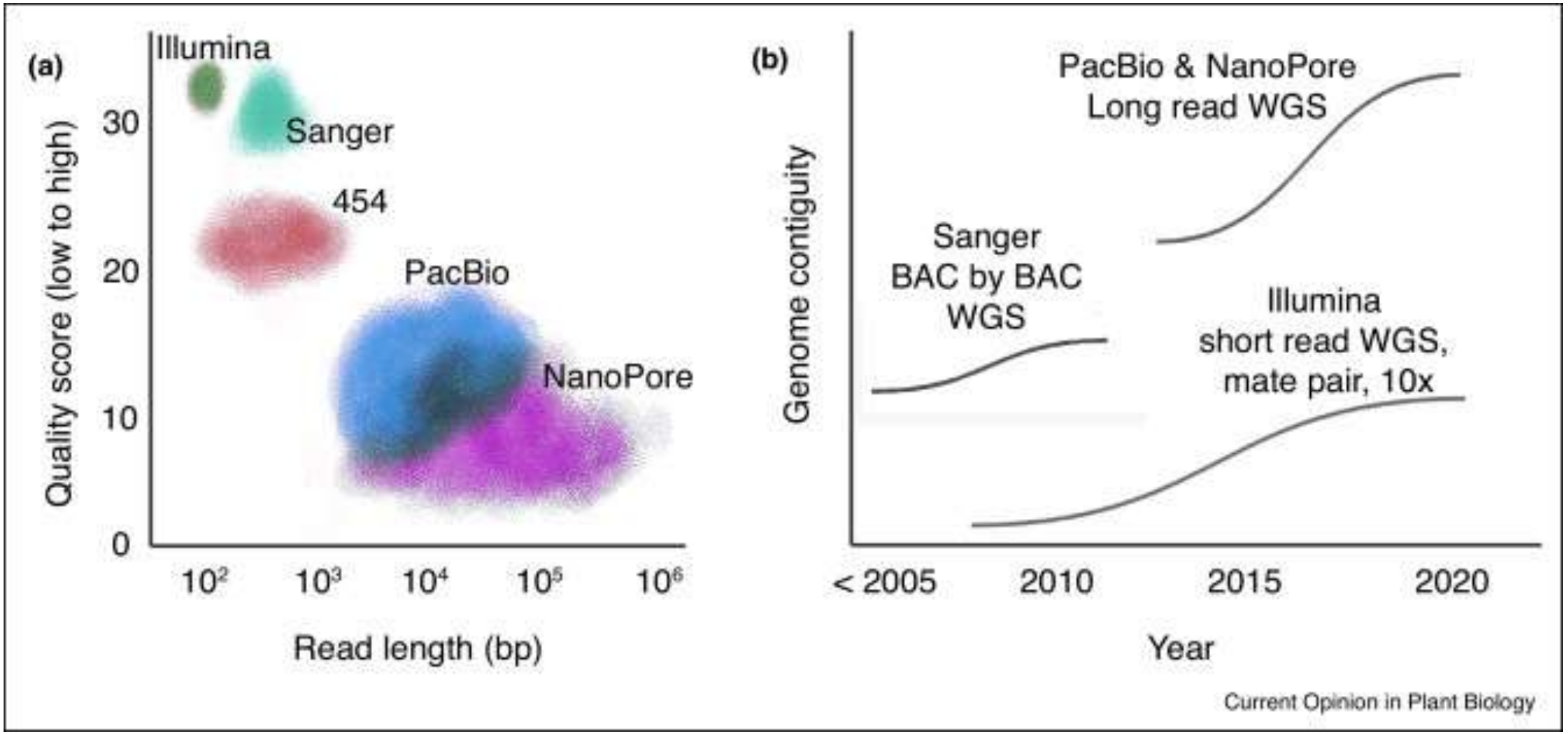
282 Illumina vs. nanopore assembly

Stewart *et al* (2019) Nature Biotechnology

Summary and opinions

Summary and opinions

Advances in sequencing technology have dramatically improved genome contiguity over the last two decades

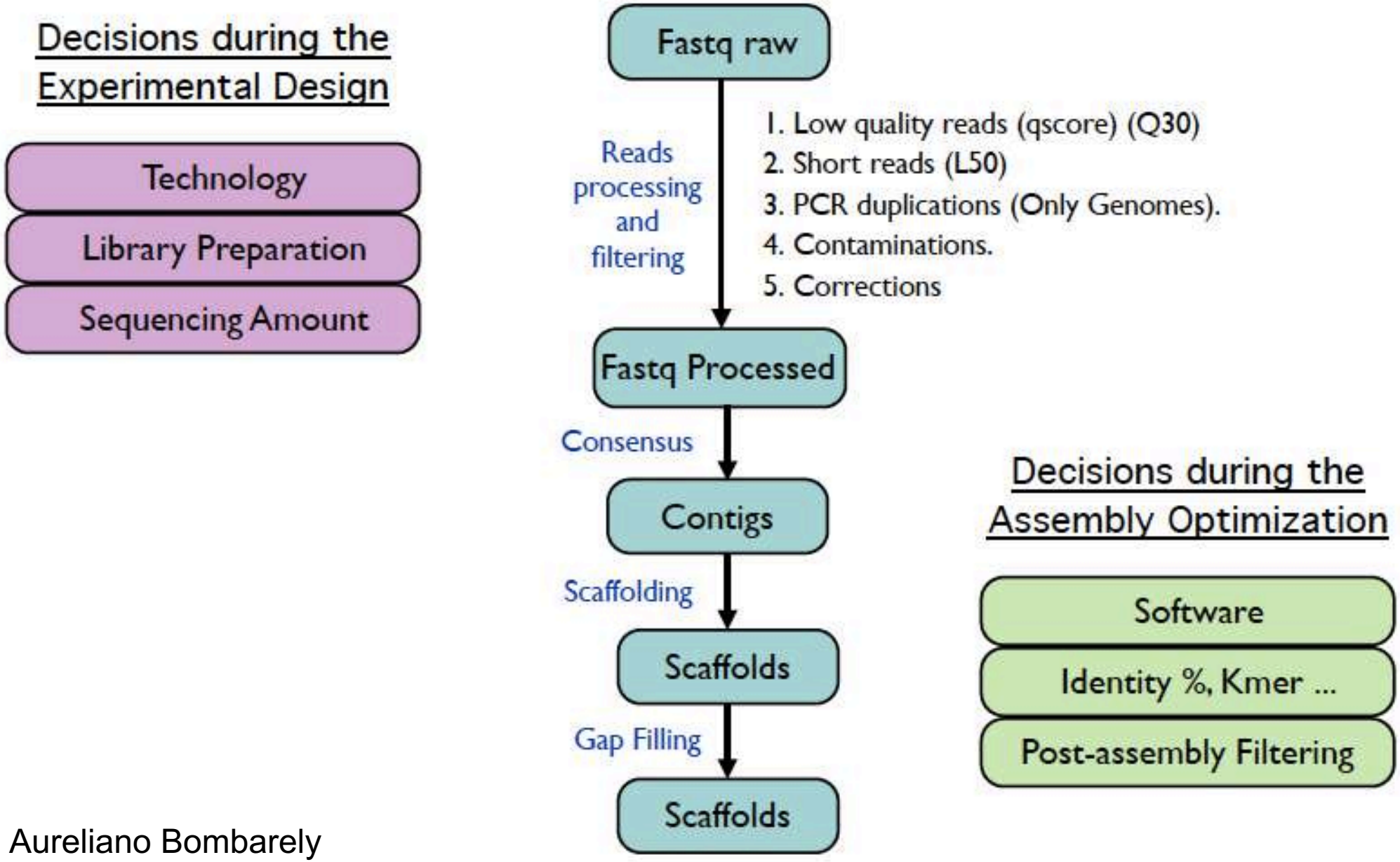


Todd and VanBuren (2020) Current Opinion in Plant Biology

Summary

- Expectation is higher in assembly of a genome
- Population genomics is moving from a mapping (resequencing) to assembly approach
- Long read technologies are improving rapidly fast so every standard lab can generate a high quality assembly
- Assembly processes need to scale up to accommodate the advancing technologies and changing biological questions

Overview of a sequencing project



Credit: Aureliano Bombarely

Things to consider

- $$$$$$$$$
- **Project type**
  virus, bacteria, eukaryote, meta-genome
- **Goals**
  Just an assembly to showcase the world?
    Sequence pandemic species = conservation? (No right or wrong answer)
  Any biological question?
  Why *de novo* sequence a species?

- **Hardware**
  - You need CPUs, but RAM is more important
    - Imagine storing all the hashes or kmers
  - This may change depending on nature of data (all long reads within n years?)

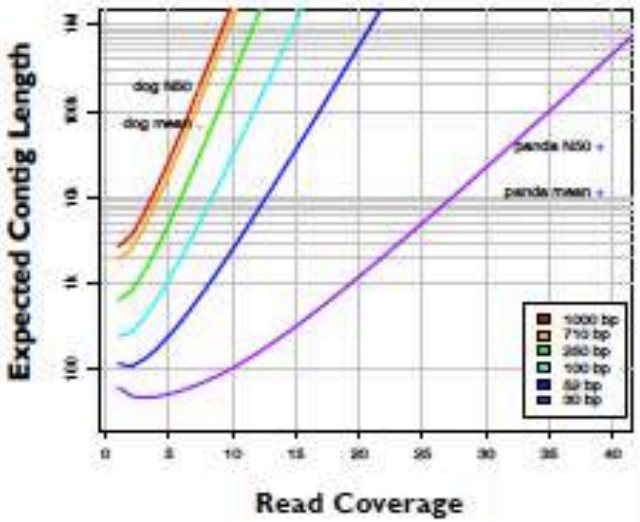
Consider technologies and experiments

- Use multiple techniques to answer your questions
  - Long read only
  - Long read + Hi-C
  - Long read + optical maps + Hi-C
  - 10X linked reads
  - 10X + Hi-C
  - Single cell?
  - Experimental advancement?

- Sometimes limitations becomes experiment preparation rather than the technological one

Ingredients for a good assembly

Coverage



High coverage is required

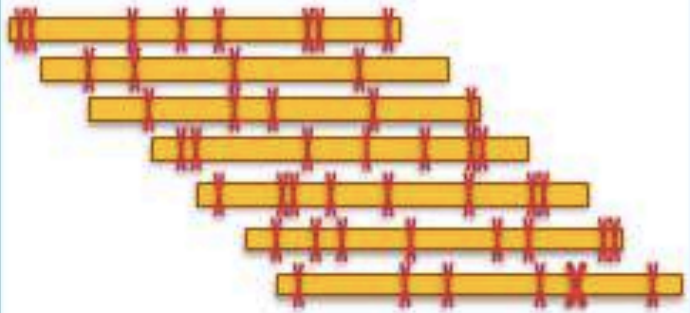
- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

Read Length

Reads & mates must be longer than the repeats

- Short reads will have *false overlaps* forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality



Errors obscure overlaps

- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Credit: Michael Schatz

FUTURE ISSUES

1. High-error reads: Third-generation DNA sequencing technologies (e.g., from Pacific Biosciences and Oxford Nanopore) generate much longer reads than previously possible (tens of thousands of base pairs), but with the cost of much higher error rates.
2. Metagenomics/mixtures of organisms: Increasingly, scientists are sequencing the genomes within mixtures of organisms, whether in the context of metagenomics or within clinical samples (e.g., mixtures of tumor cells). Most of the theoretical framework for sequence assembly was developed for isolated genomes. New methods will need to be developed that can both cope with and characterize heterogeneity within closely related genomes.
3. Dealing with big data: As DNA sequencing costs drop, scientists are increasingly able to focus on larger genomes (such as those of plants) and mixtures (such as soil metagenomes). New approaches will be needed that allow genome sequence assemblers to scale with the amount of data being generated.

Simpson and Green (2015) Annual Review of Genomics and Human Genetics

Not covered but should be

Alignment method in overlap graph
String graph (reduced overlap graph)
Shortest Superstring Problem (SSP)
Hamiltonian path

Choice of kmers in DBG
Bloom filter

References

shorturl.at/xBC06

[Announcement]

- Possible changes in lectures
- May have to cancel practical classes ;
  - Or alternative formats, for instance, I will teach R live and go through an example
- May have to move presentation forward
- Please email me for suggestions
  - possible topics that you wish to discuss