De novo metagenomics assembly

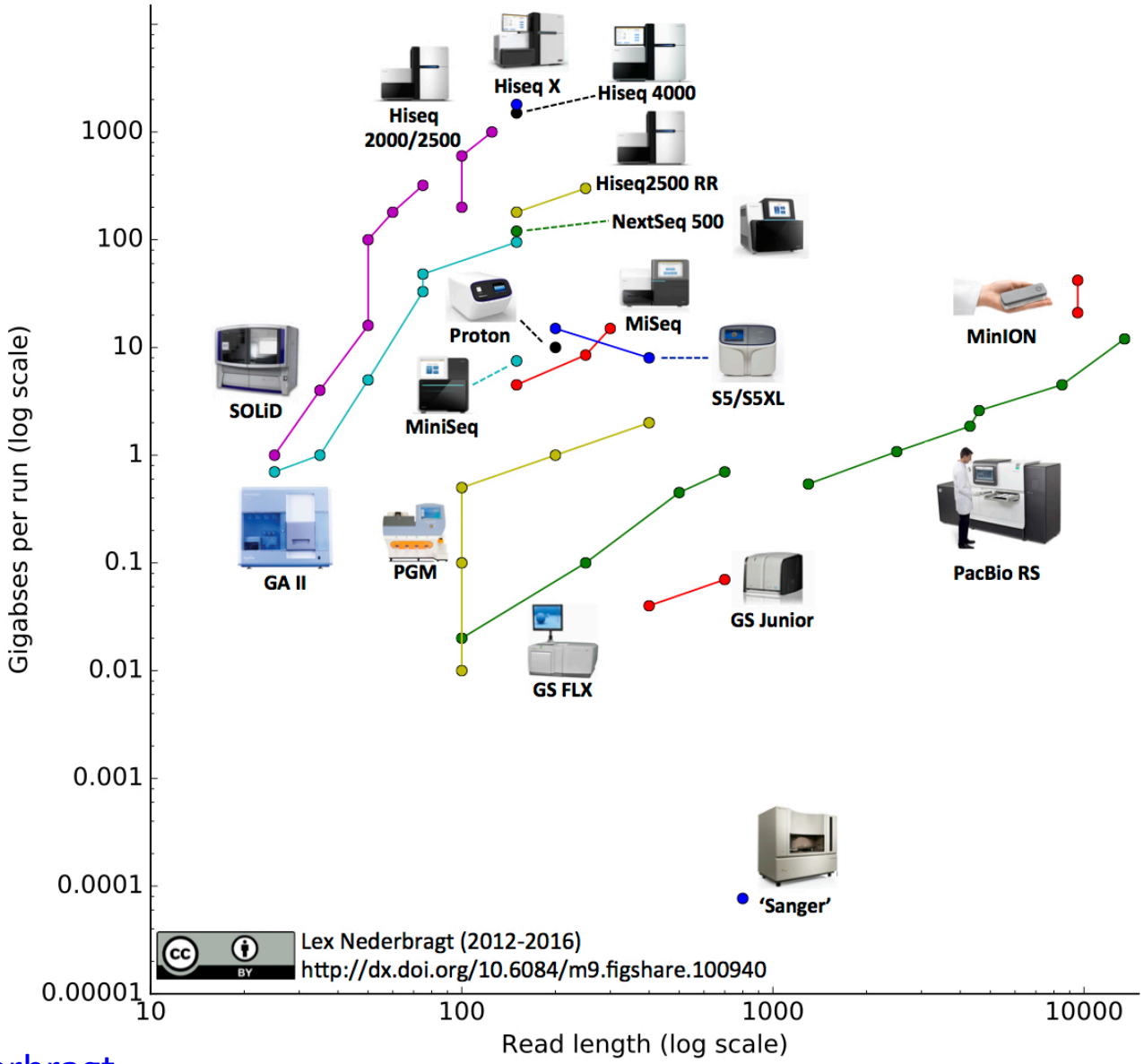
Christopher Quince
Warwick Medical School

Introduction

- What is de novo assembly?
- Why assemble?
- Reference based assembly simple but limited:
  - Map reads onto known genome
  - Requires closely related reference
  - Cannot find novel genes, plasmids, rearrangements

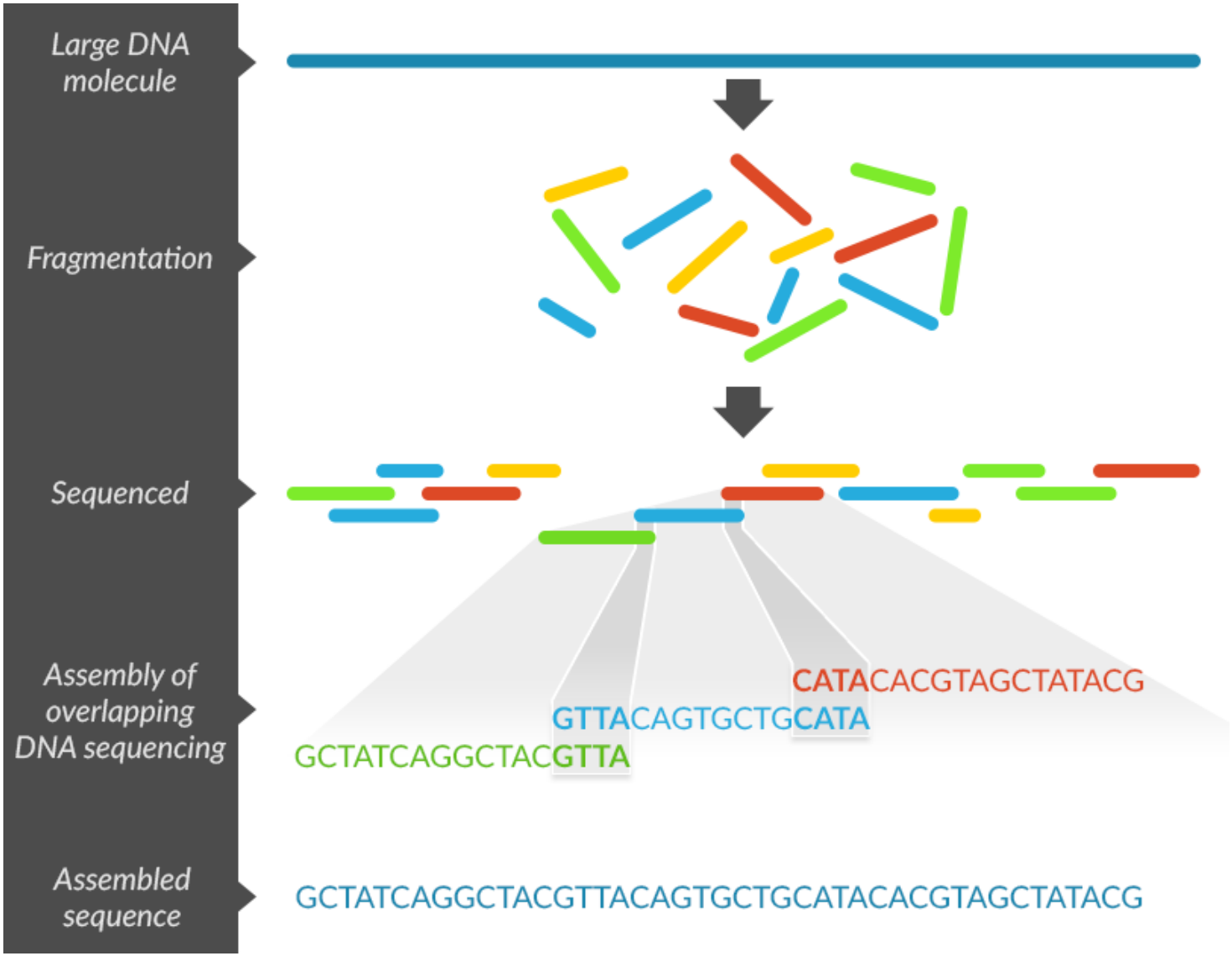
Overview

- Current state of sequencing technologies
- De novo assembly paradigms
- Coverage
- Repeats
- Metagenomics assembly:
  - Coassembly
  - Strains
- Assembly software

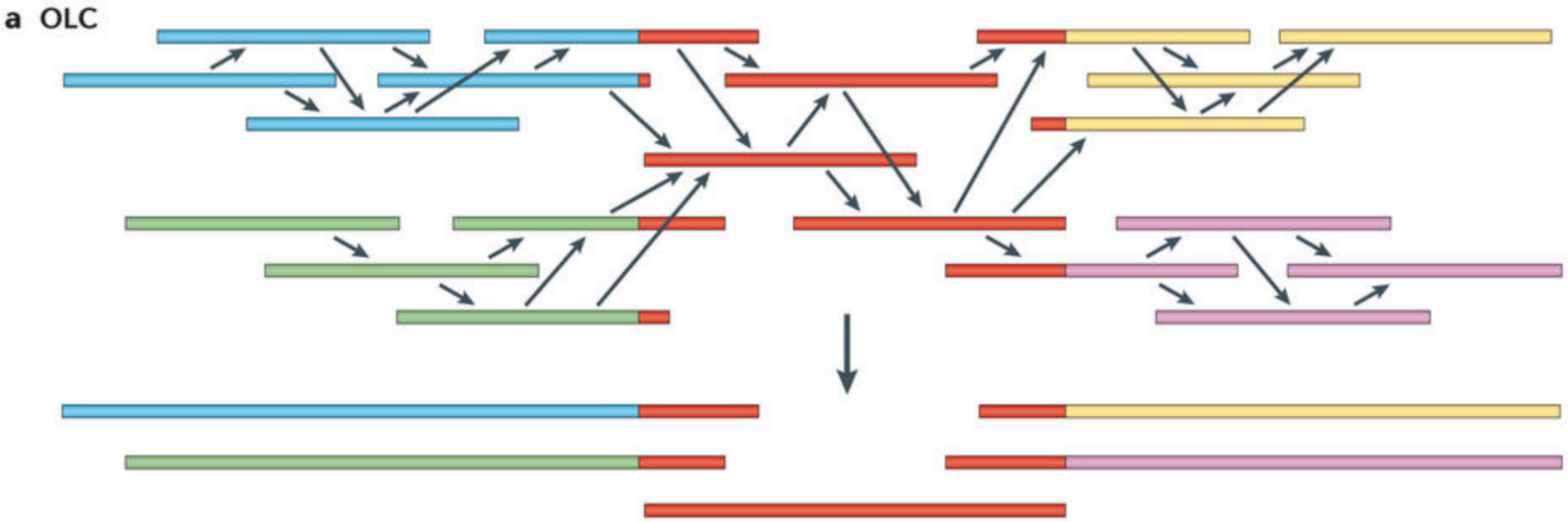


Lex Nederbragt
https://figshare.com/articles/developments_in_NGS/100940

What is de novo sequence assembly?



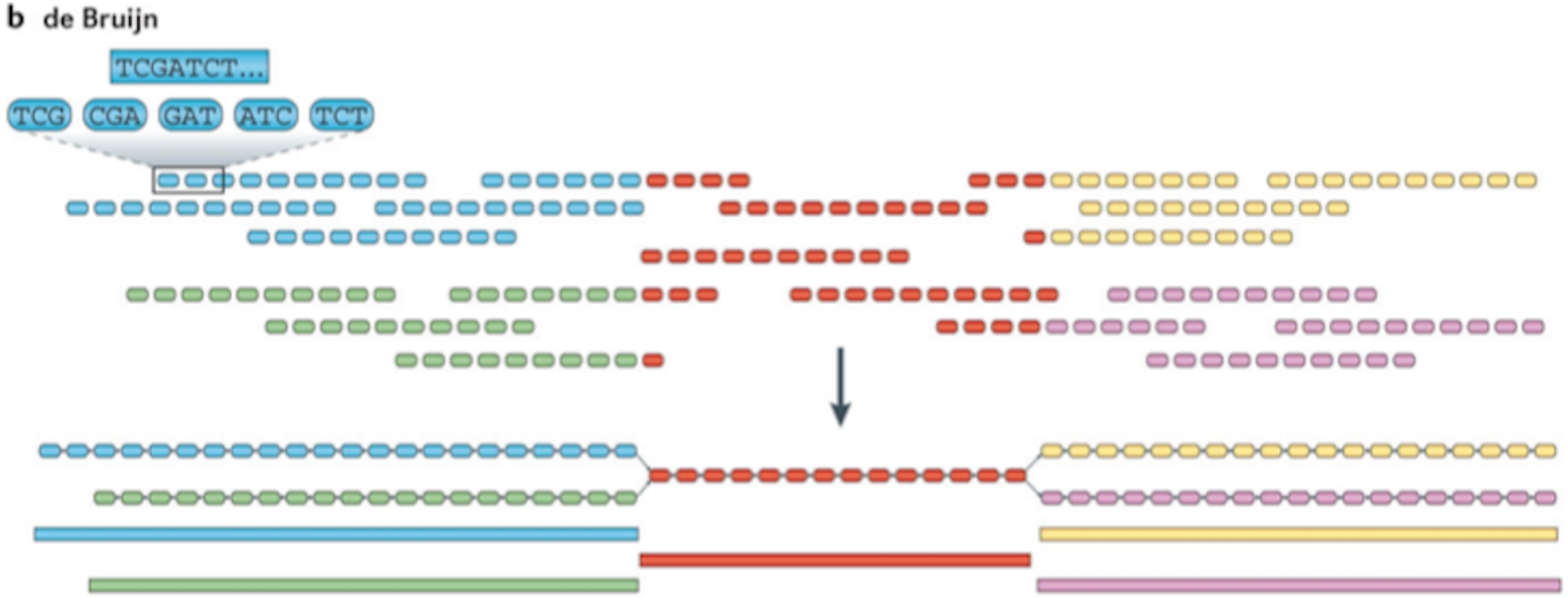
Overlap layout consensus



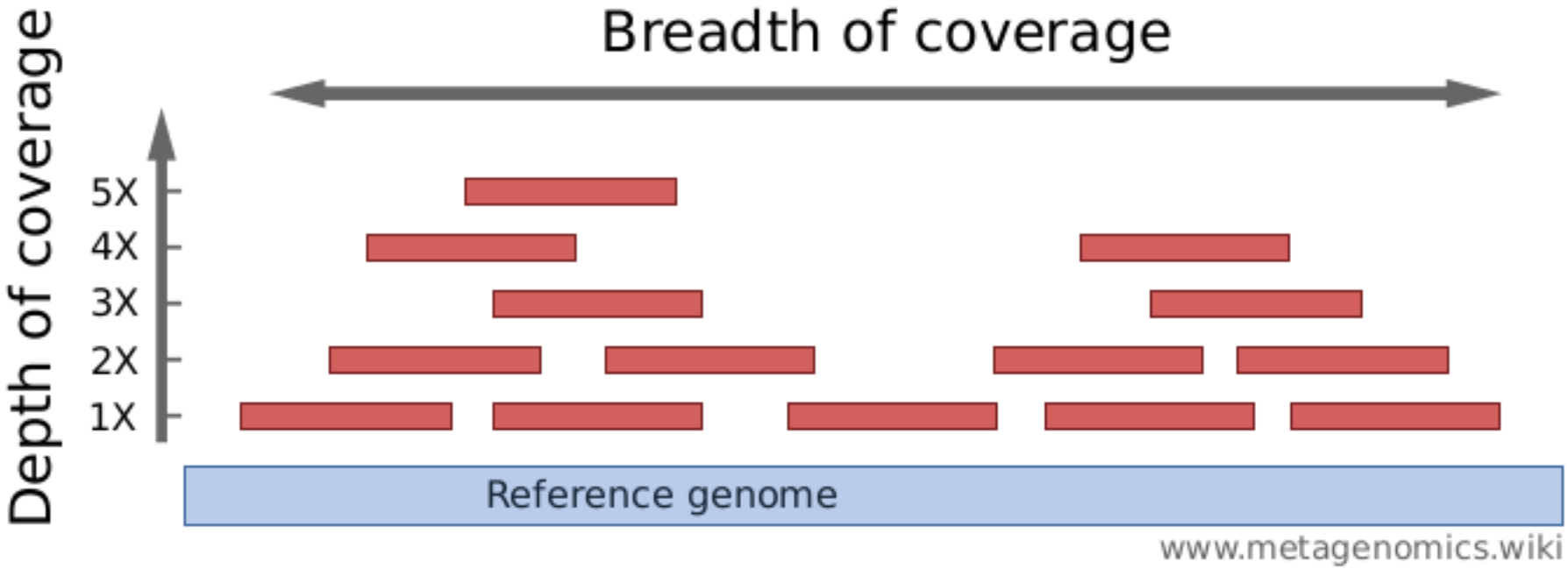
- OLC slow because of pair-wise comparisons
- Renaissance with long read technologies

Genetic variation and the de novo assembly of human genomes
Mark J. P. Chaisson, Richard K. Wilson & Evan E. Eichler
Nature Reviews Genetics 16, 627–640 (2015) doi:10.1038/nrg3933

De Bruijn graph assembly



- Fast but effectively fixed length exact overlaps
- Still default for short read next generation

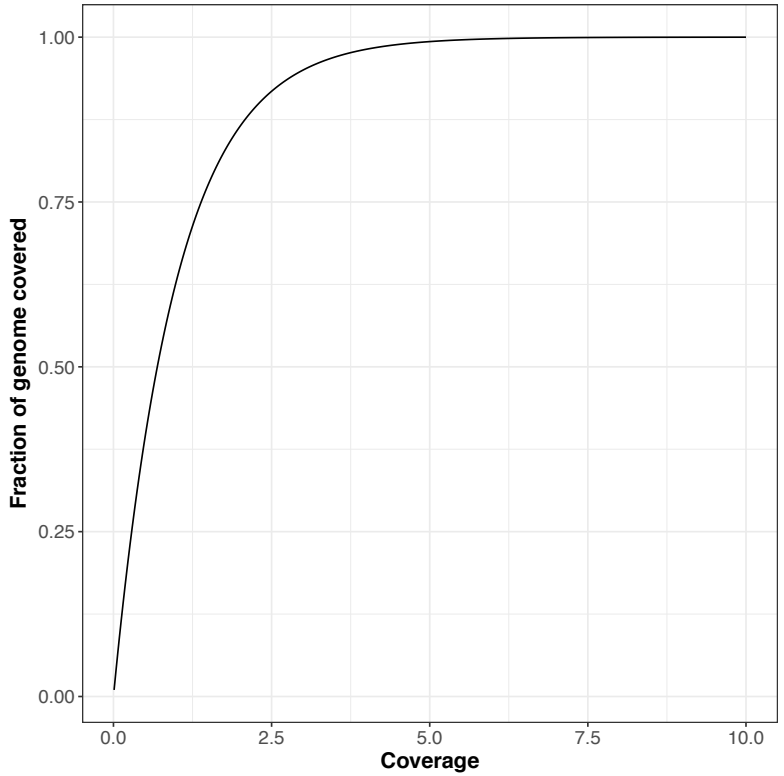


Coverage depth of taxa n $$c_n = \frac{\rho_n RL}{G_n}$$

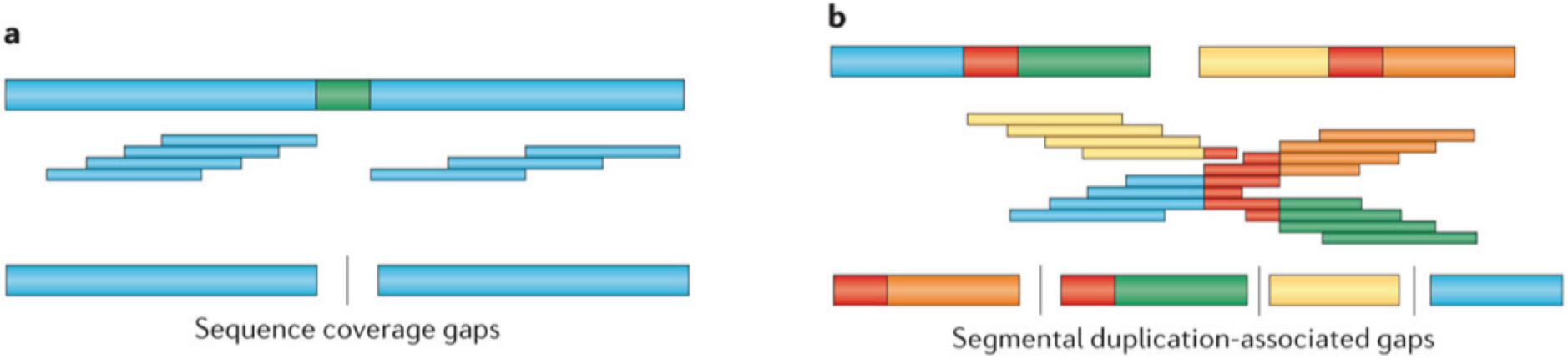
- Where ρ_n is genome relative frequency of taxa n
- R number of reads in library
- L read length
- G_n is genome length of taxa n

Lander-Waterman statistics

- Assume reads randomly distributed across genome
- Fraction of genome covered: $1 - e^{-c_n}$
- Mean number of contigs: Re^{-c_n}



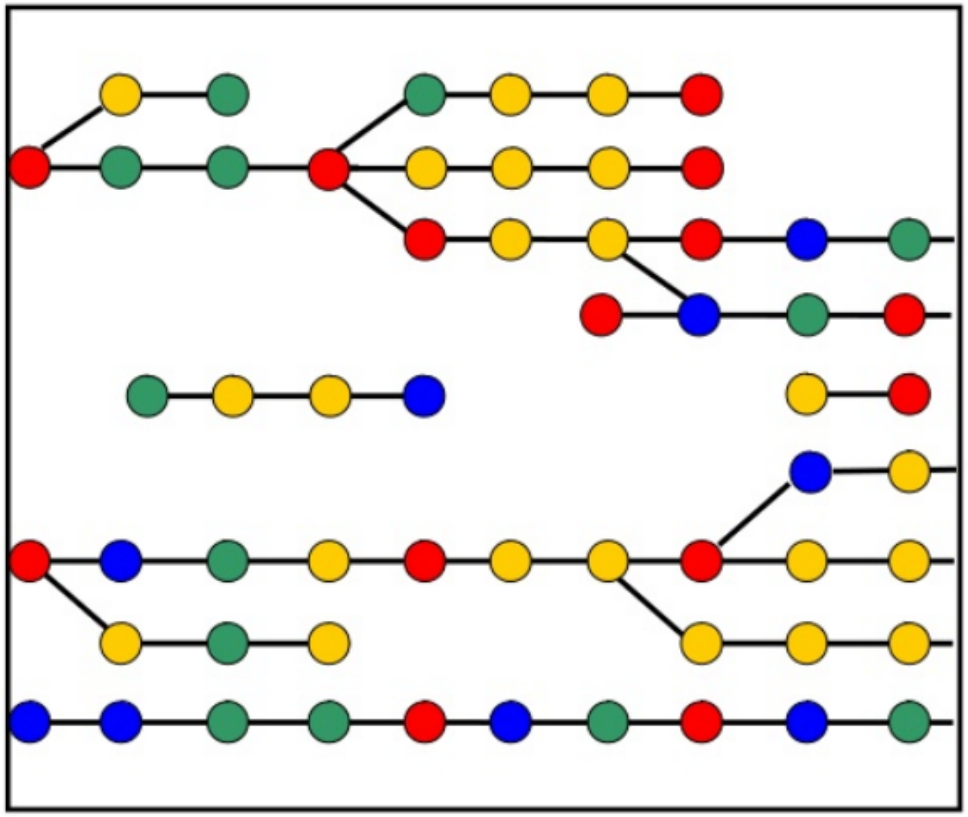
Gaps in assembly



- A contig is just an unambiguously assembled genome fragment
- What is the optimal kmer length for de Bruijn graph assembly?

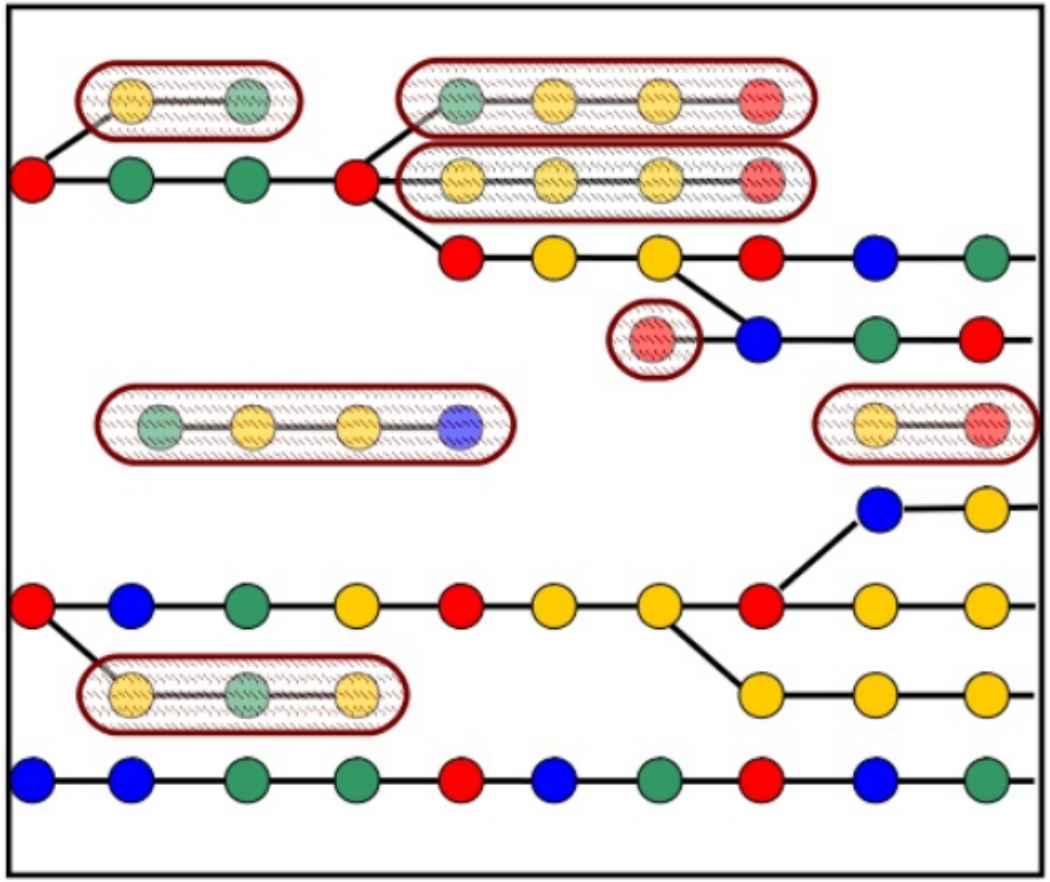
Genome assembly in practice

- Read errors generate tips that are pruned:



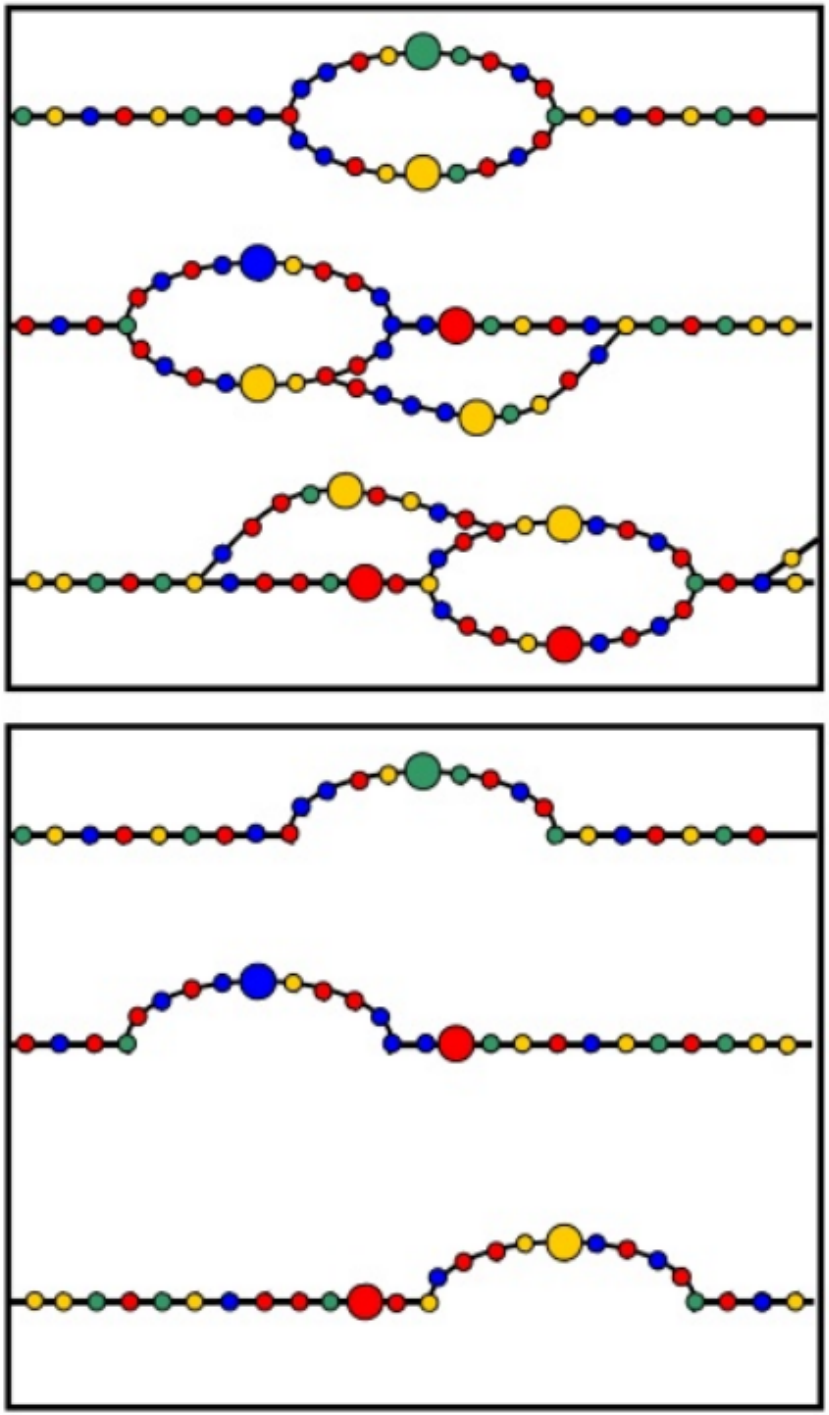
Genome assembly in practice

- Read errors generate tips that are pruned:



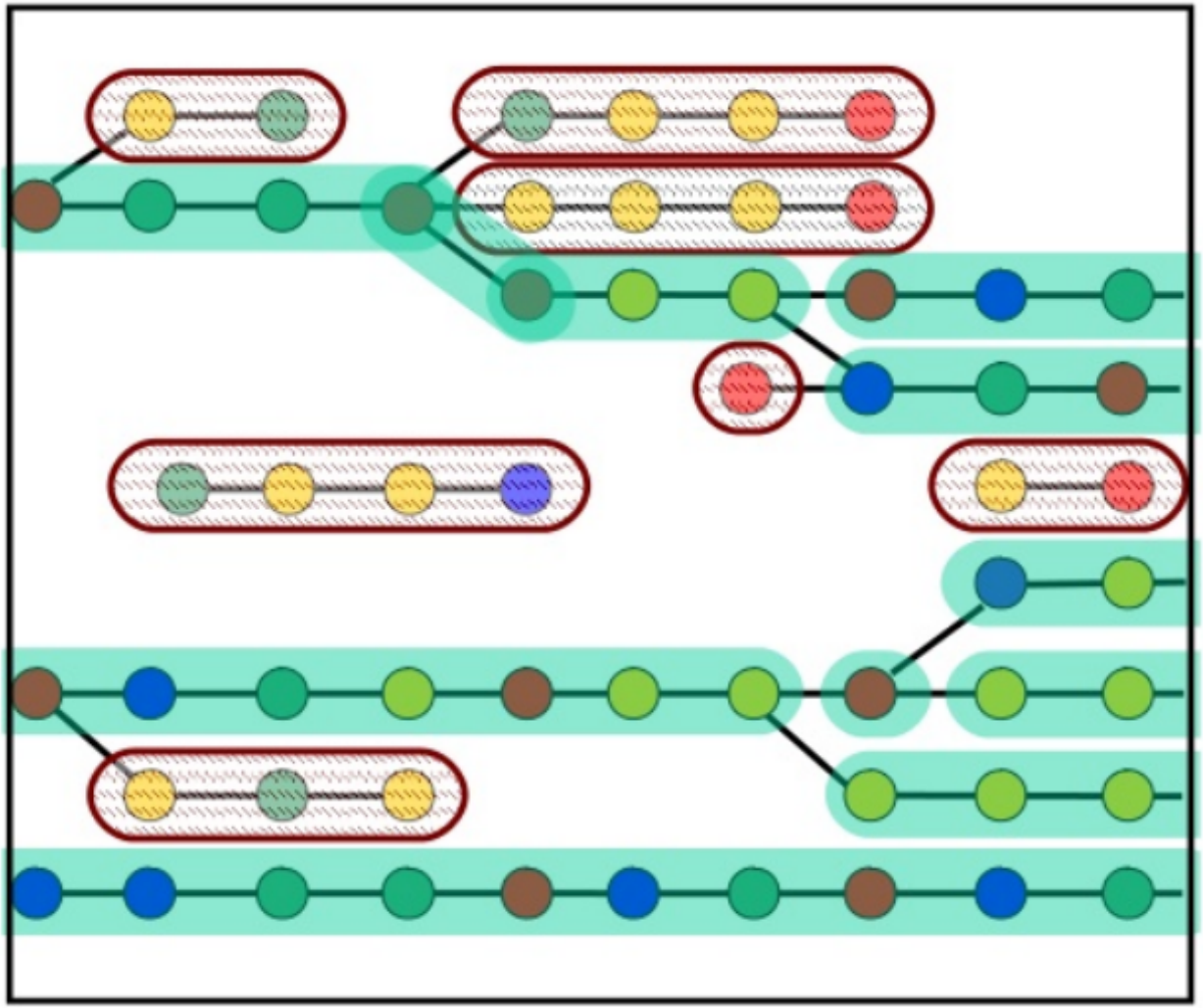
Bubbles

- Errors can also generate bubbles but these can also be real variants
- These are popped
- Removing real variants from assembly



Contigs

- Remove ambiguous edges output linear portions of graph:



Variable k-mer length dBG assemblers

- Small k more branches
- Larger k more gaps
- Approach pioneered in IDBA is to iterate through kmer lengths to get best of both
- Used in most dBG assemblers now: idba, idba_ud, megahit, minia

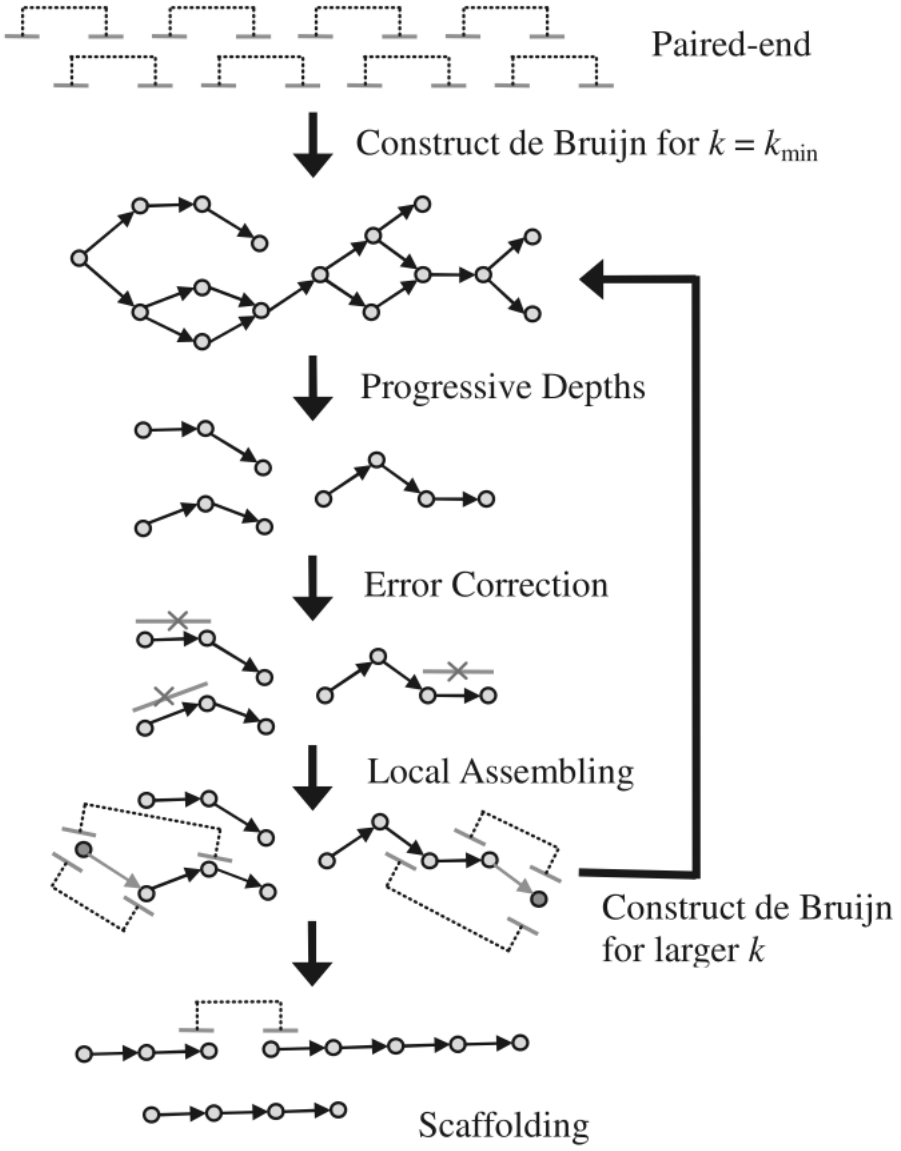


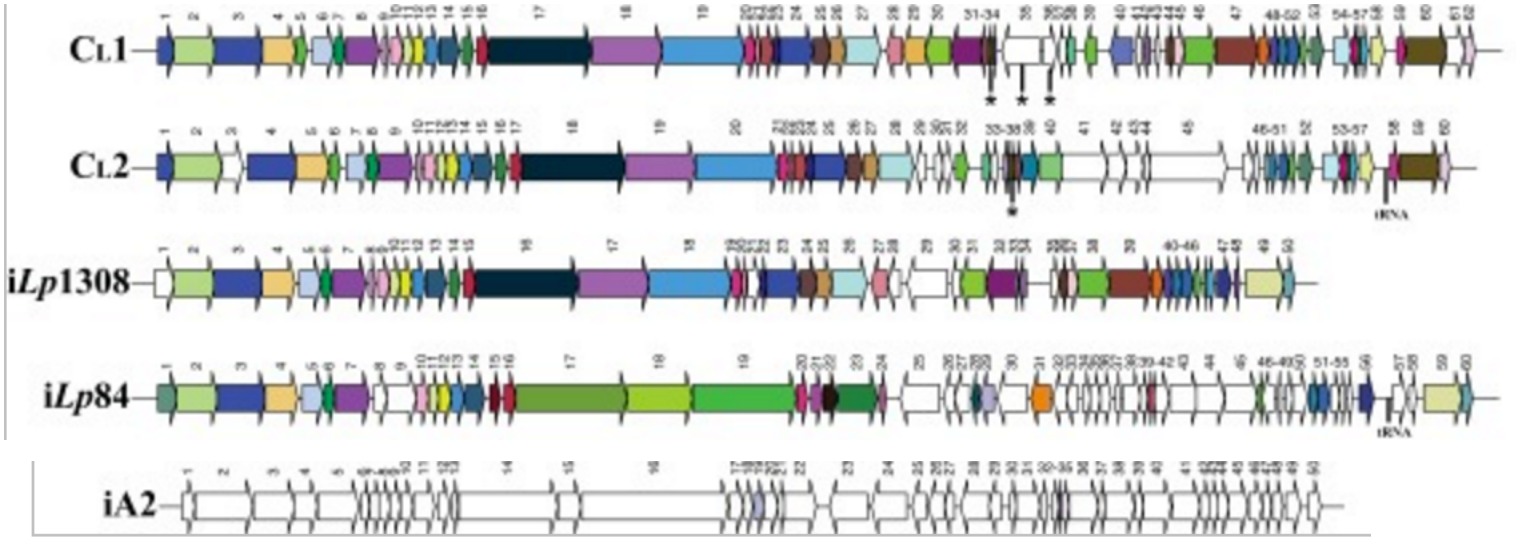
Fig. 1. Flowchart of IDBA-UD

Evaluating assemblies

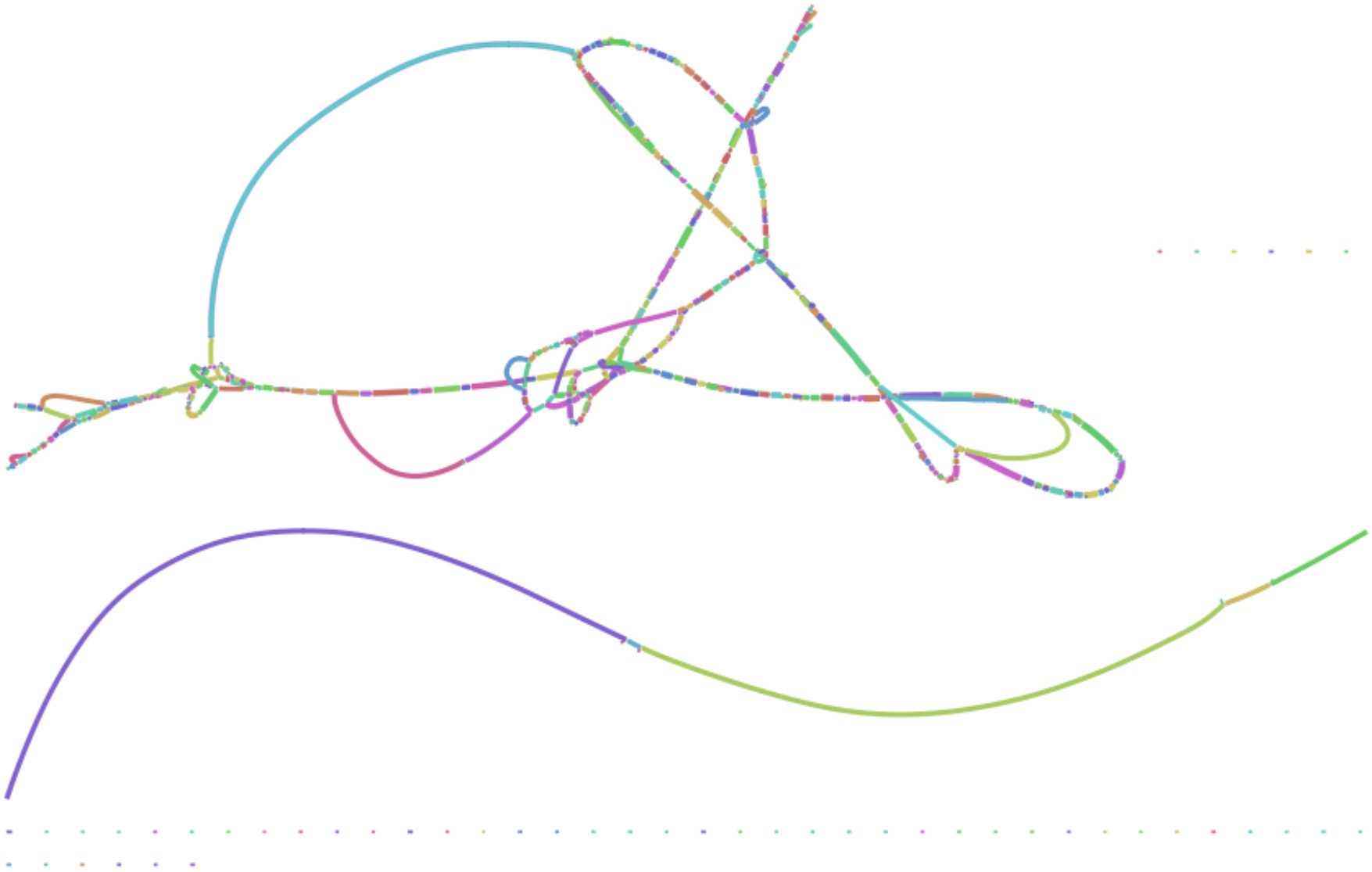
- Total size of assembly
- Number of contigs
- N50 – minimum contig length in which at least 50% of bases are contained
- But contigs can be chimeric (Quast)

Why is metagenomics assembly hard?

- Took five *Lactococcus paracasei* temperate phage genomes: CL1, CL2, iLp84, iLp1308, and iA2
- The genome lengths ranged from 34,155 bp (iA2) to 39,474 bp (CL1)
- Generated 10,000 synthetic 2X150bp reads for 16 samples

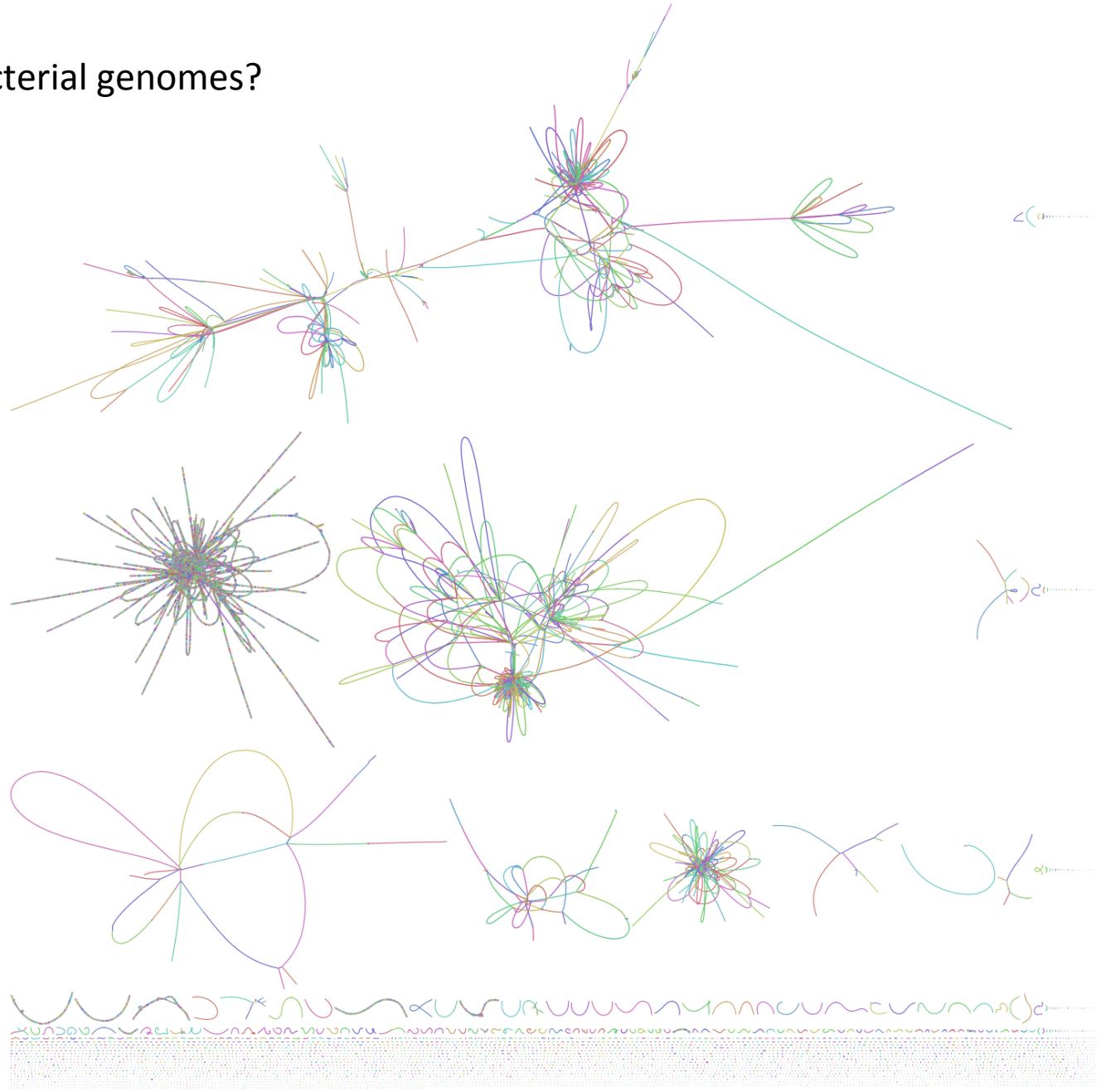


Mercanti et al. Genomic Diversity of Phages Infecting Probiotic Strains of Lactobacillus paracasei AEM 2016



De Bruijn graph Bcalm assembly with kmer length 71

20 bacterial genomes?



Use contig binning to cluster contigs back into strain/species genomes

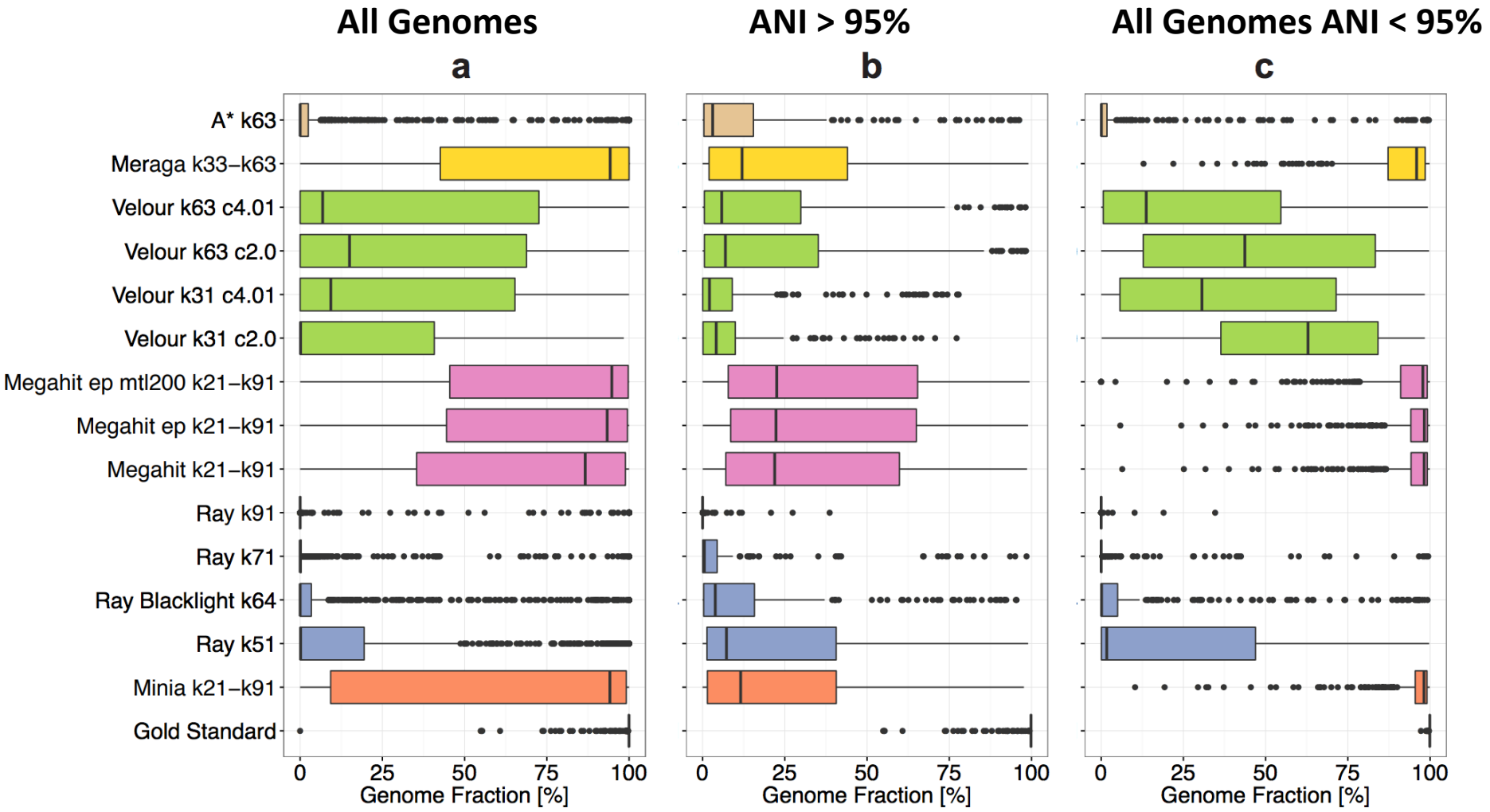
Critical assessment of metagenome interpretation

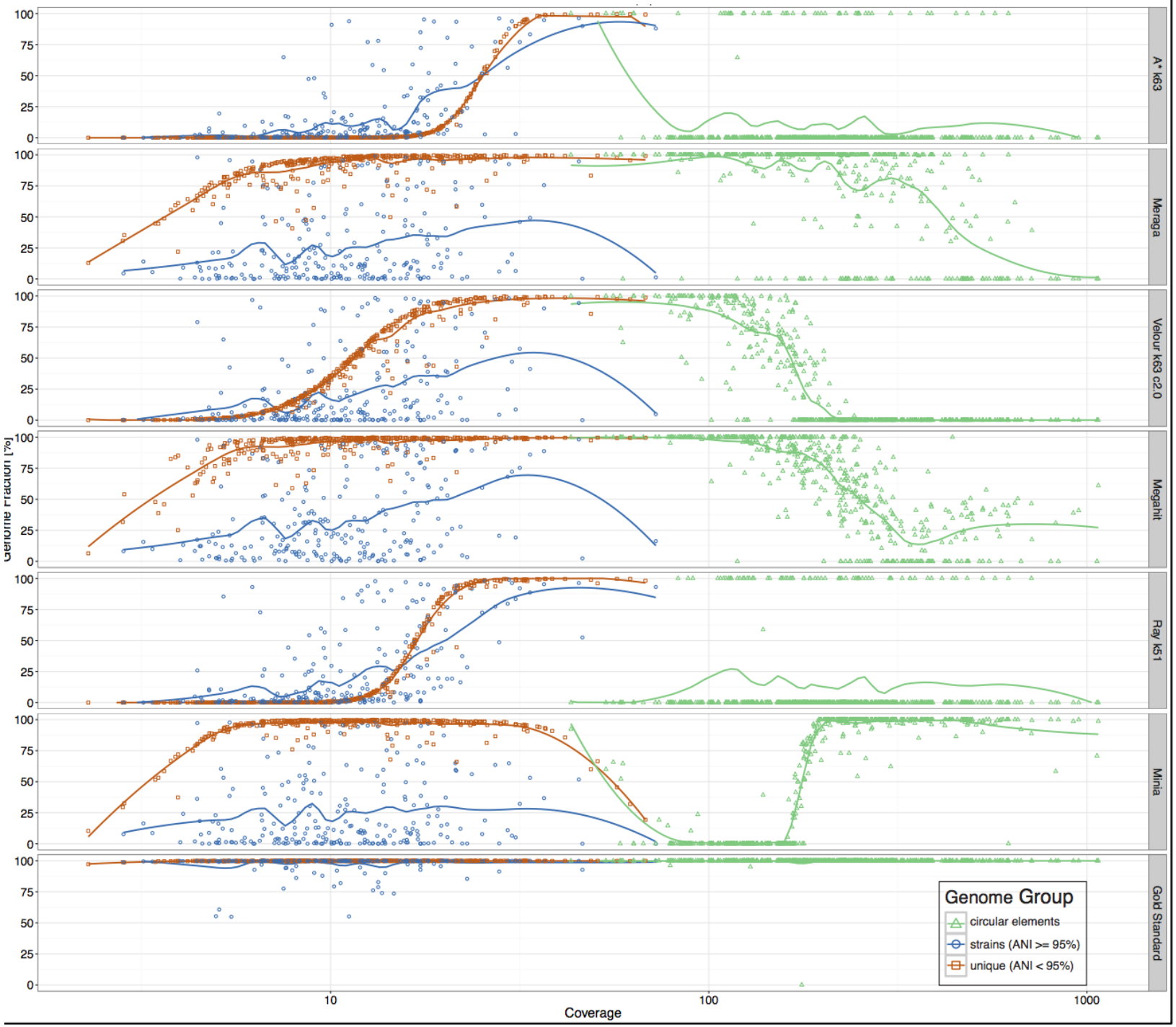
(https://www.biorxiv.org/content/biorxiv/early/2017/01/09/099127.full.pdf)

- Three datasets:
  - 15 Gbp – single sample low complexity (40 genomes and 20 circular)
  - 40 Gbp – differential abundance dataset with two samples of a medium complexity community (132 genomes and 100 circular elements)
  - 75 Gbp time series dataset with five samples from a high complexity community (596 genomes and 478 circular elements).

No Spades

Software	Description
Assemblers	
Megahit v.0.2.2	Metagenome assembler using multiple k-mer sizes and succinct de Bruijn graphs
Ray Meta v2.3.2	Distributed de Bruijn graph metagenome assembler
Meraga v2.0.4	Meraculous + Megahit
Minia 2 and Minia 3	De Bruijn graph assembler based on a Bloom filter
A*	OperaMS Scaffolder using SOAPde novo2 on medium complexity and Ray assemblies on low and high complexity data sets
Velour	De Bruijn graph genome assembler

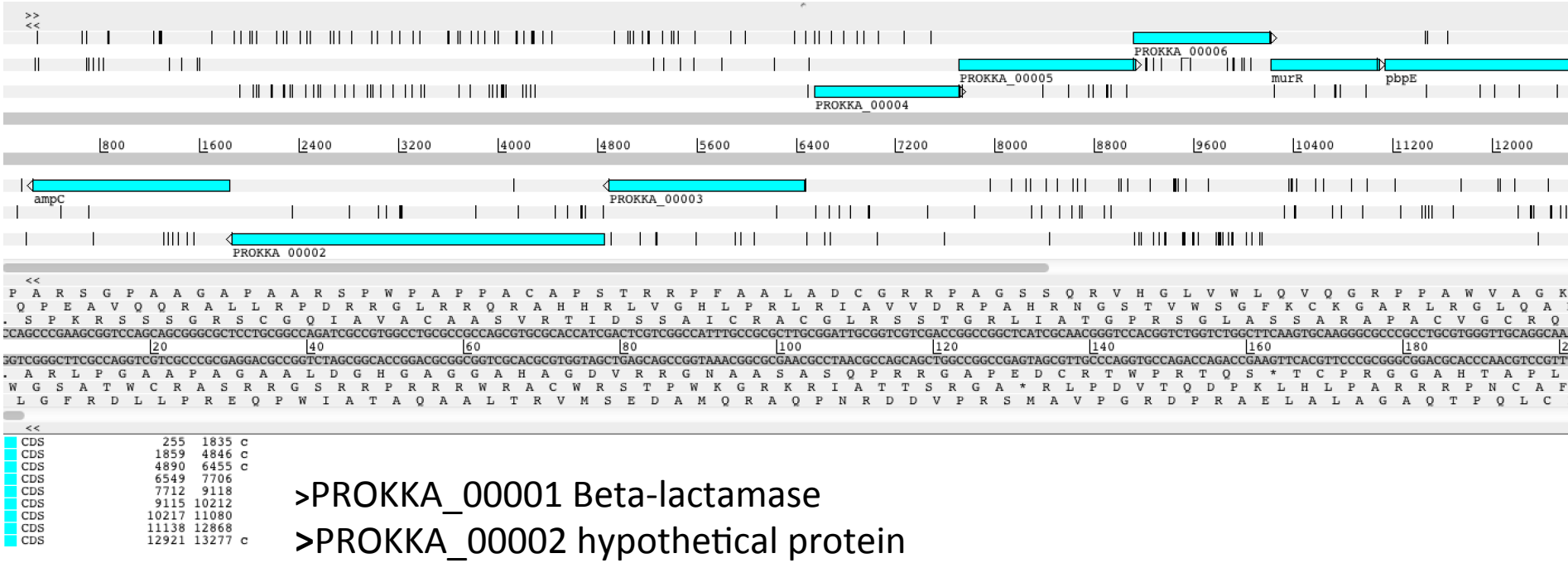

All Genomes
a
ANI > 95%
b
All Genomes ANI < 95%
c
A* k63
Meraga k33–k63
Velour k63 c4.01
Velour k63 c2.0
Velour k31 c4.01
Velour k31 c2.0
Megahit ep mtl200 k21–k91
Megahit ep k21–k91
Megahit k21–k91
Ray k91
Ray k71
Ray Blacklight k64
Ray k51
Minia k21–k91
Gold Standard
0
25
50
75
100
Genome Fraction [%]


A* k63
Meraga
Velour k63 c2.0
Megahit
Ray k51
Minia
Gold Standard
Genome Fraction [%]
Coverage
10
100
1000
0
25
50
75
100
Genome Group
circular elements
strains (ANI >= 95%)
unique (ANI < 95%)

Coassembly

- In a metagenome study we may have multiple studies from similar environments with the same organisms:
  - Time series e.g. reactor
  - Horizontal sampling
- Combining samples increases coverage of rare organisms and enables binning
- Cost is increased strain confusion
- Are contigs useful?

Can look at context in contig e.g. AMR



>PROKKA_00001 Beta-lactamase
>PROKKA_00002 hypothetical protein
>PROKKA_00003 hypothetical protein
>PROKKA_00004 hypothetical protein
>PROKKA_00005 hypothetical protein
>PROKKA_00006 L-Ala-D/L-Glu epimerase
>PROKKA_00007 HTH-type transcriptional regulator MurR
>PROKKA_00008 Penicillin-binding protein 4*
>PROKKA_00009 hypothetical protein

Top blast hit Lysobacter antibioticus strain ATCC 29479 genome 78% 75%

Summary

- De novo assembly of genomes is a powerful way to extract biological information from data set
- We will end up with very fragmented assemblies
- These are still useful and can be made even more useful with binning