NORWEGIAN SEQUENCING CENTRE

De novo assembly of short reads using Velvet

Adapted from Nick Loman

University of Birmingham

https://github.com/lexnederbragt/denovo-assembly-tutorial

CEES
Centre for Ecological and Evolutionary Synthesis

Velvet

- One of the first short read assemblers
- Developed by Daniel Zerbino of EBI
- A *de Bruijn* graph assembler, like:
  - SOAPdenovo
  - ABYSS
  - ALLPATHS
  - Etc.

K-mers again

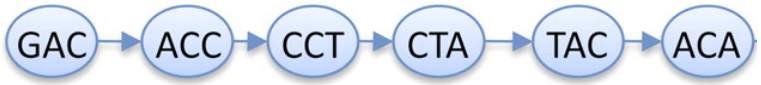
Read GACCTACA

K-mers (K=3)

```
GAC
 ACC
  CCT
   CTA
    TAC
     ACA
```

K-1 bases overlap

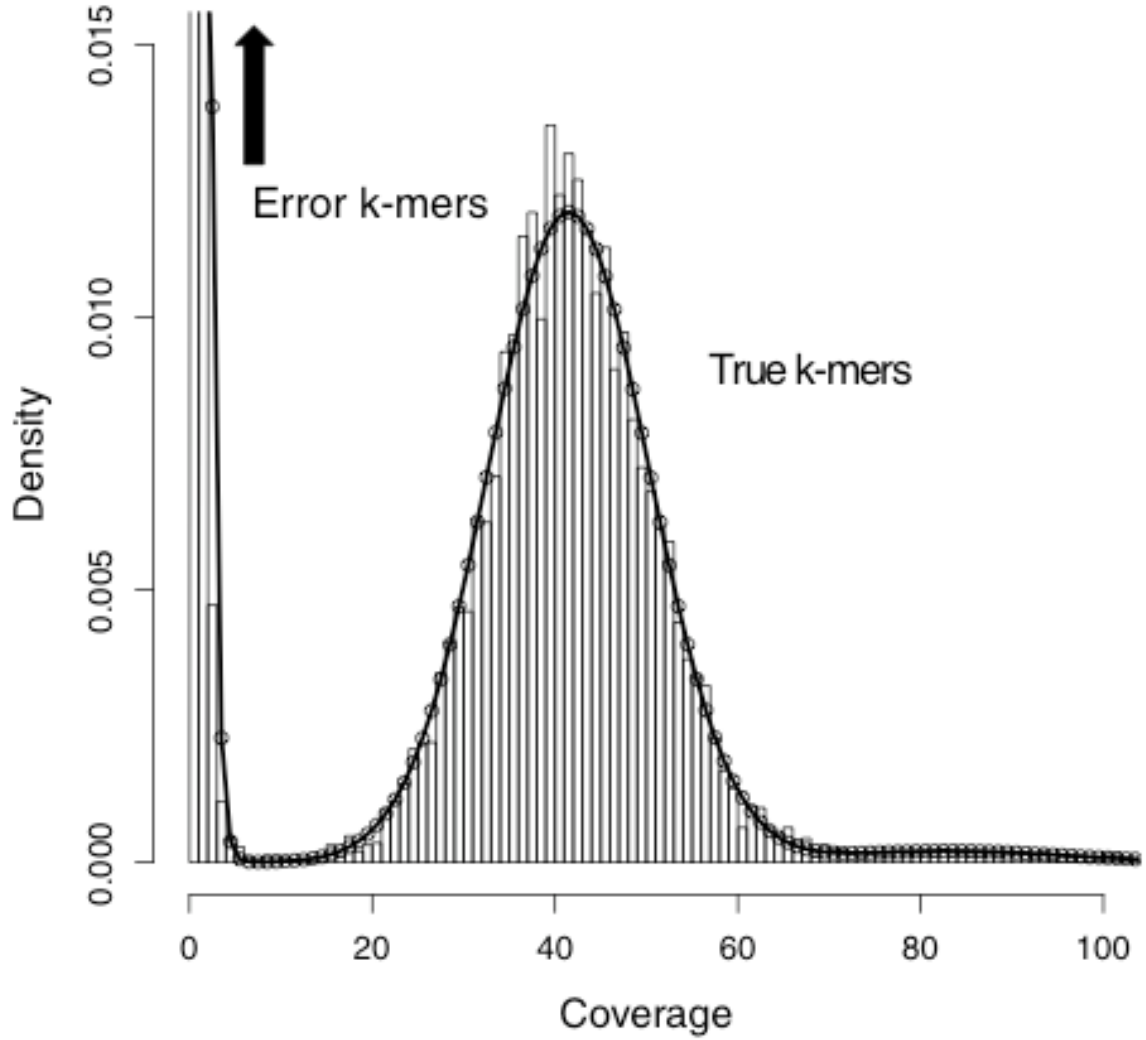
de Bruijn graph



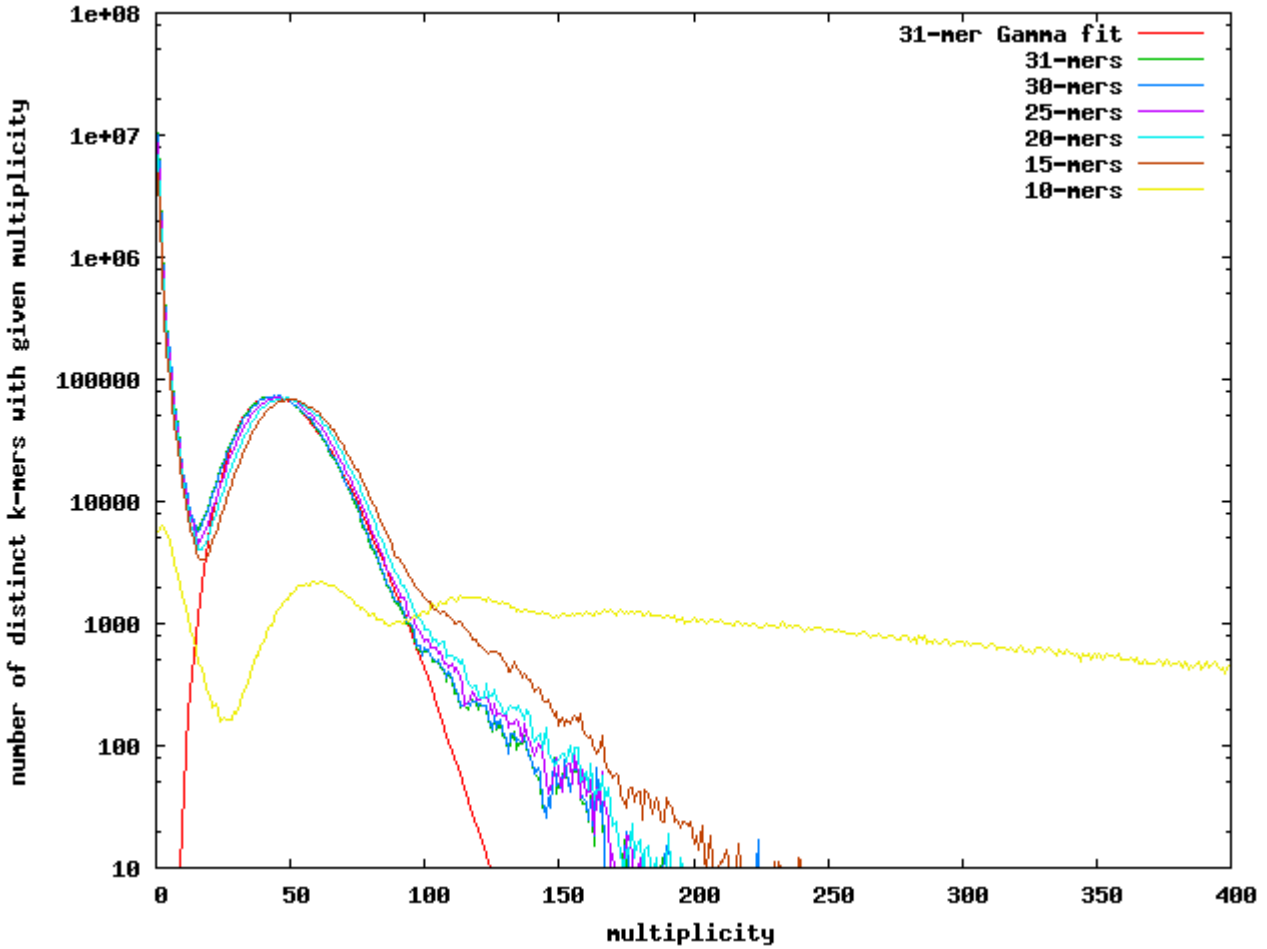
Counting k-mers

- Plotting k-mer frequencies is a quick and easy way of:
  - Estimating genome size
  - Seeing copy number variation in genome
  - Estimating sequence read error
  - Planning a short-read assembly

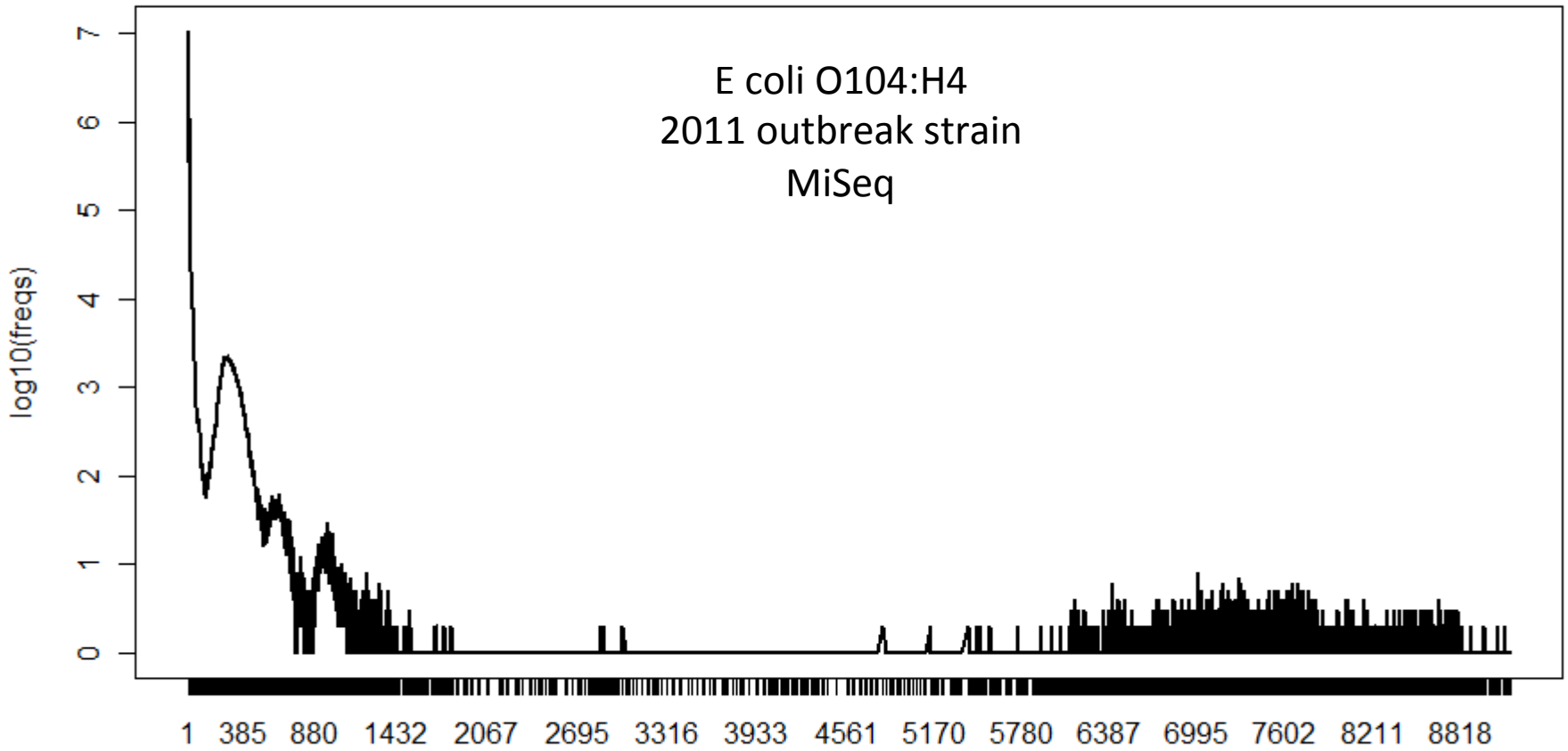
Idealised k-mer plot



A real K-mer plot



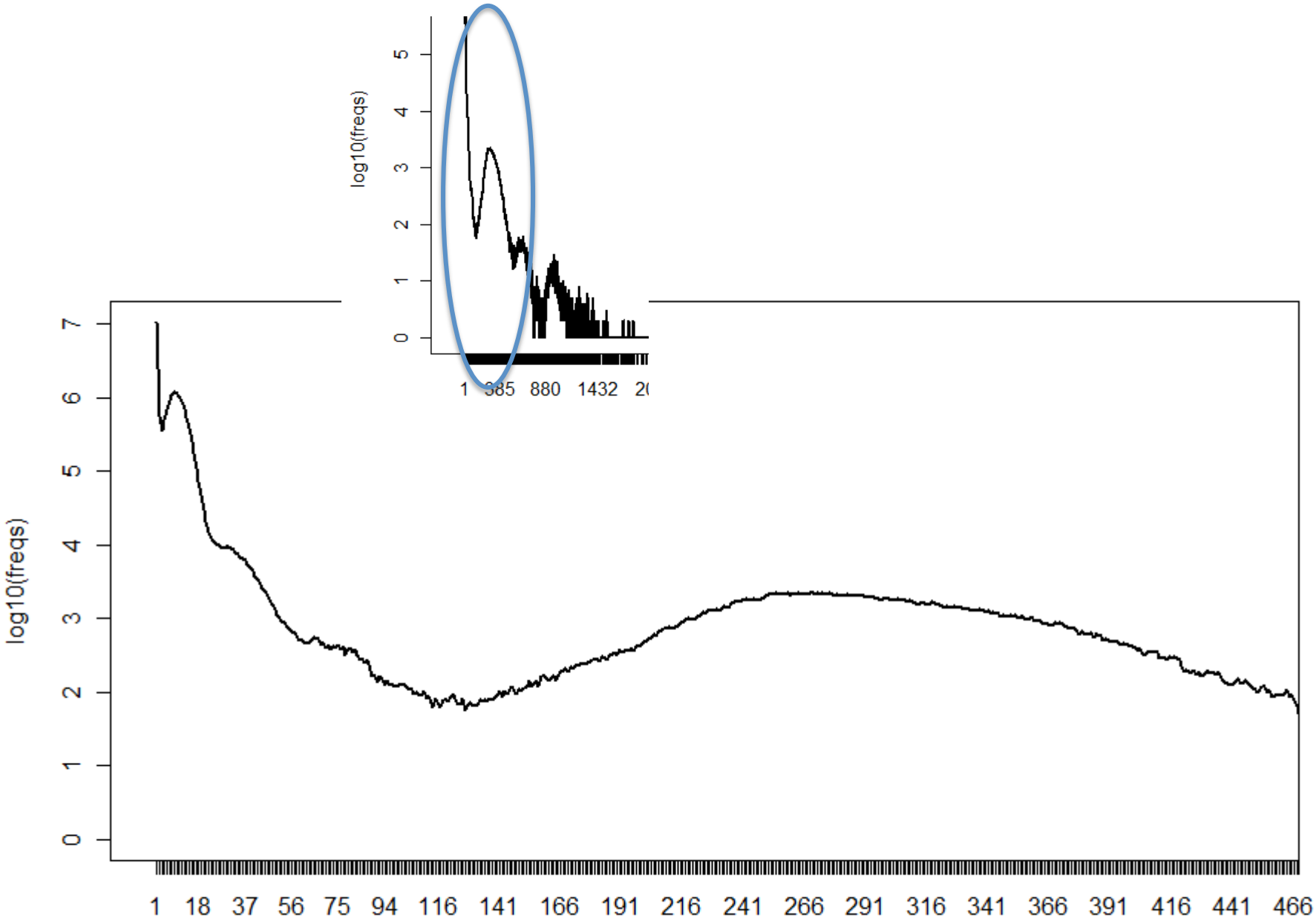
https://banana-slug.soe.ucsc.edu/bioinformatic_tools:jellyfish



Plasmids!

	TY2482	S282	S283	S540	S541
Chromosome depth	300	20	24	27	21
pTY1 (IncI) depth	310	496	563	550	472
pTY1 copy number	***1.0***	***24.8***	***23.5***	***20.4***	***22.5***
pTY2 (EAEC) depth	200	477	417	435	392
pTY2 copy number	***0.7***	***23.9***	***17.4***	***16.1***	***18.7***
pTY3 (mini plasmid) depth	2658	11418	9972	8420	8403
pTY3 copy number	***9***	***571***	***416***	***312***	***400***

Mean read depth and mean plasmid copy number for outbreak strains


log10(freqs)
0
1
2
3
4
5
6
7
1
18
37
56
75
94
116
141
166
191
216
241
266
291
316
341
366
391
416
441
466
log10(freqs)
0
1
2
3
4
5
1
385
880
1432

K-mers and K

no magical value of k

Depends on

- read length
- sequencing error
- rate of polymorphism
- coverage

K-mers and K

Some rules:

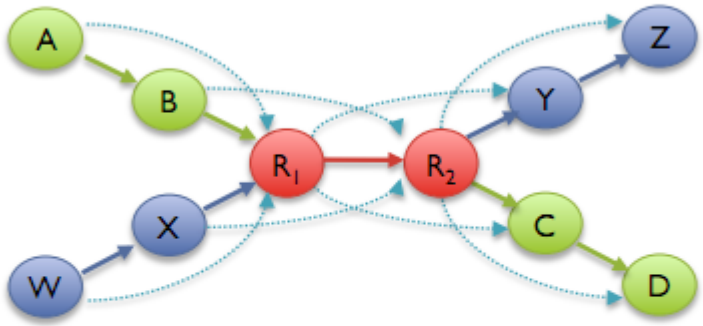
- k must be less than the read length
- k can't be an even number (can produce palindromes)

K-mers and K

- Bigger *k*
  - Solves more repeats
  - fewer overlaps
  - lower k-mer coverage
- Lower *k*
  - more overlaps
  - higher k-mer coverage

de Bruijn Graphs

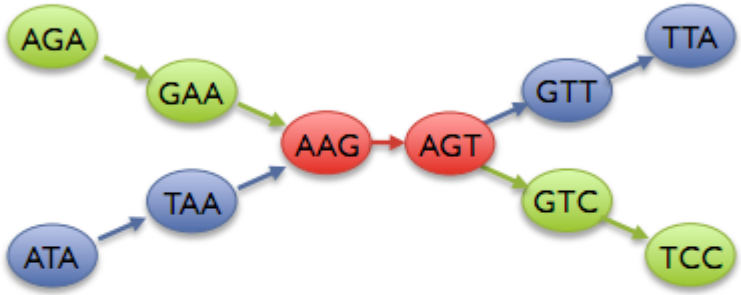
Overlap Graph



Long read assemblers

- Repeats depends on read length
- Read coherency, placements kept
- Tangled by high coverage

de Bruijn Graph

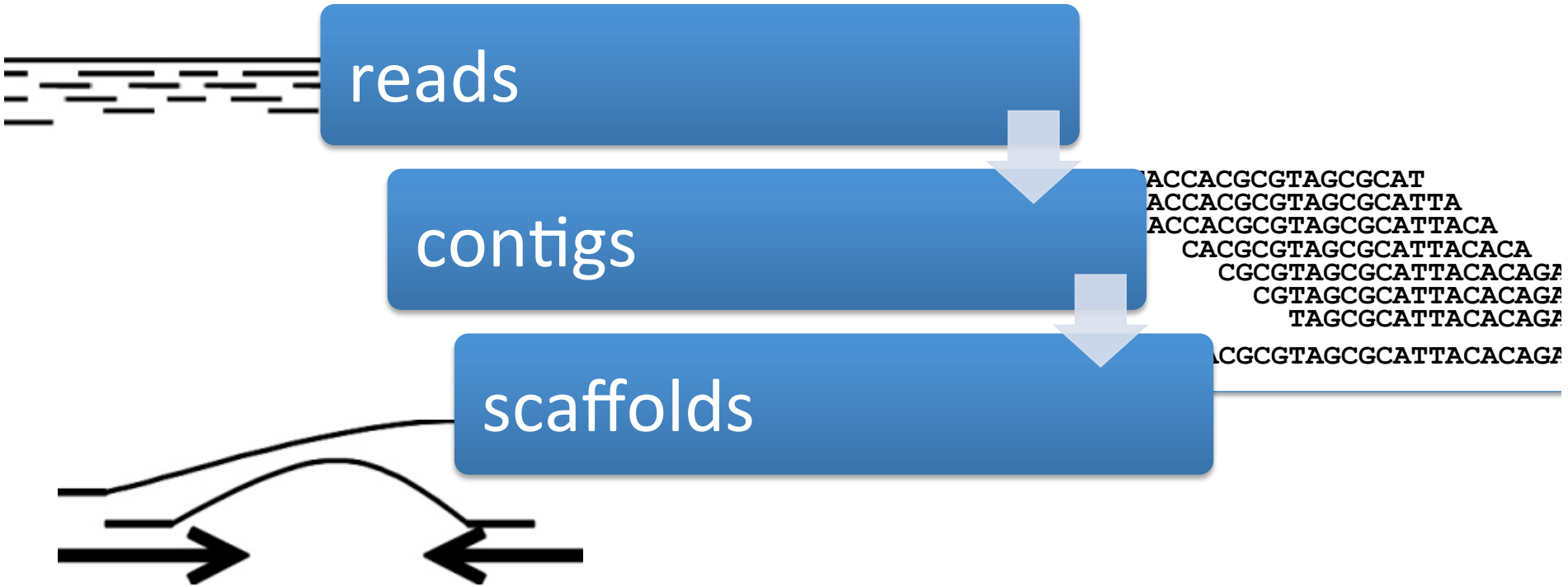


Short read assemblers

- Repeats depends on word length
- Read coherency, placements lost
- Robust to high coverage

Mike Schatz

What do you get?

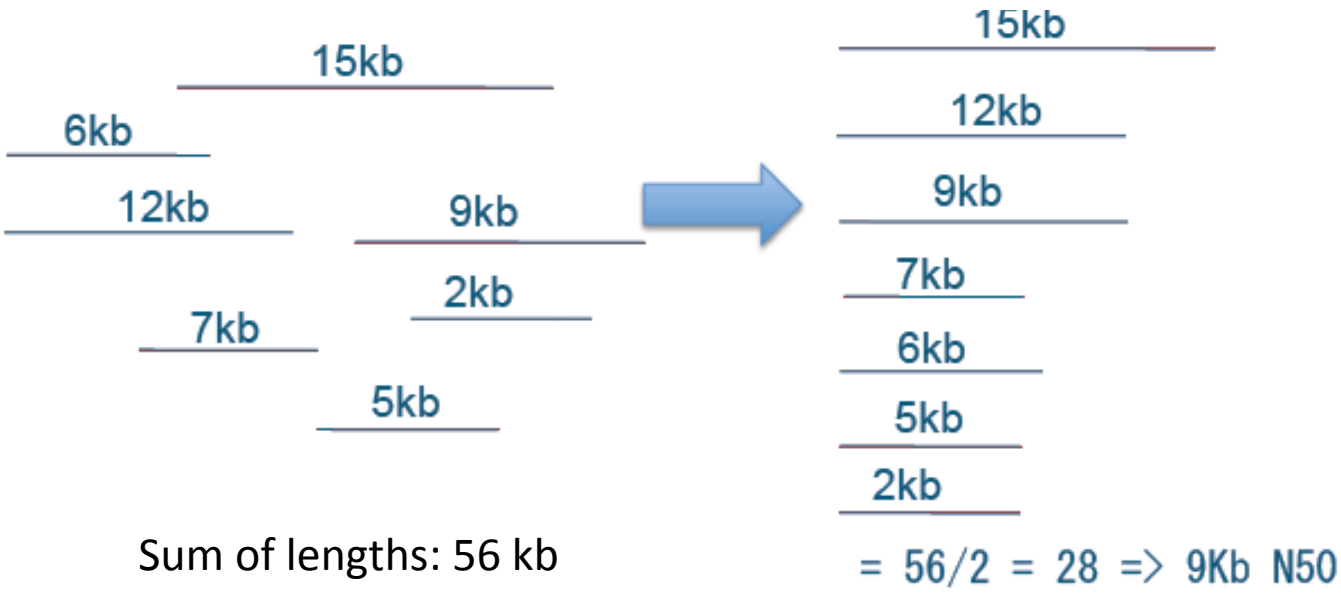


Metrics

- contigs
  - how many
  - total bases
  - N50
- scaffolds
  - how many
  - total bases
  - N50
  - how many gaps
  - total gap bases

N50

- Size of contig such that 50% of total bases are in contigs of this length or more



Sum of lengths: 56 kb

= 56/2 = 28 => 9Kb N50

Thomas Keane and Jan Aerts, Wellcome Trust Sanger

N50

- Size of contig such that 50% of total bases are in contigs of this length or more

- OR

- Shortest of the longest contigs that together make up 50% of the assembly

N50

- Size of contig such that 50% of total bases are in contigs of this length or more

→longer N50 is better

- N50 count:
- number of contigs of at least N50 size

N50 – NG50

- N50:
- Size of contig such that 50% of total bases are in contigs of this length or more

- NG50:
- Replace 'total bases' with 'genome length'

N50

- Minimum contig length influences N50

- Take away shorter contigs → N50 goes up

N50

- High N50 → better assembly

- BUT
- says nothing of quality

Insert size

- Experimental evidence
- Allow Velvet to guess
- Map reads and calculate

Mate-pair data

- Orientation different
- Read contamination

PE => insert <=
mate-pair <= insert =>