

# *De novo* assembly of short reads using Velvet

Adapted from Nick Loman  
University of Birmingham

# Velvet

- One of the first short read assemblers
- Developed by Daniel Zerbino of EBI
- *A de Bruijn* graph assembler, like SPAdes

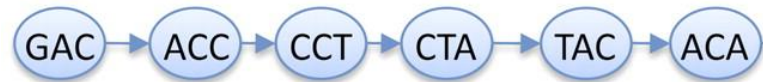
# K-mers again

Read    GACCTACA

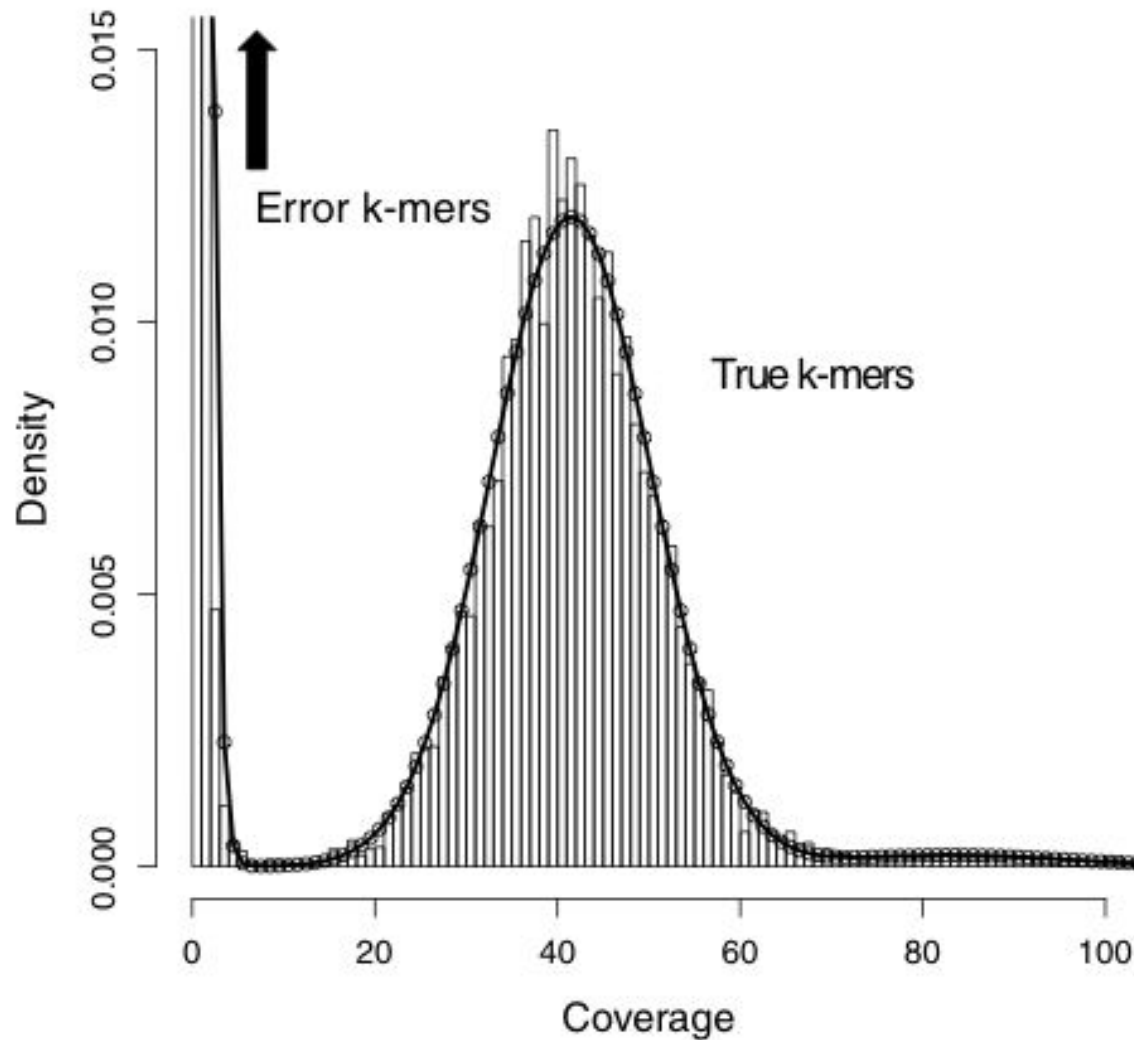
K-mers (K=3)    GAC  
                  ACC  
                  CCT  
                  CTA  
                  TAC  
                  ACA

K-1 bases overlap

de Bruijn graph



# Idealised k-mer plot



# 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

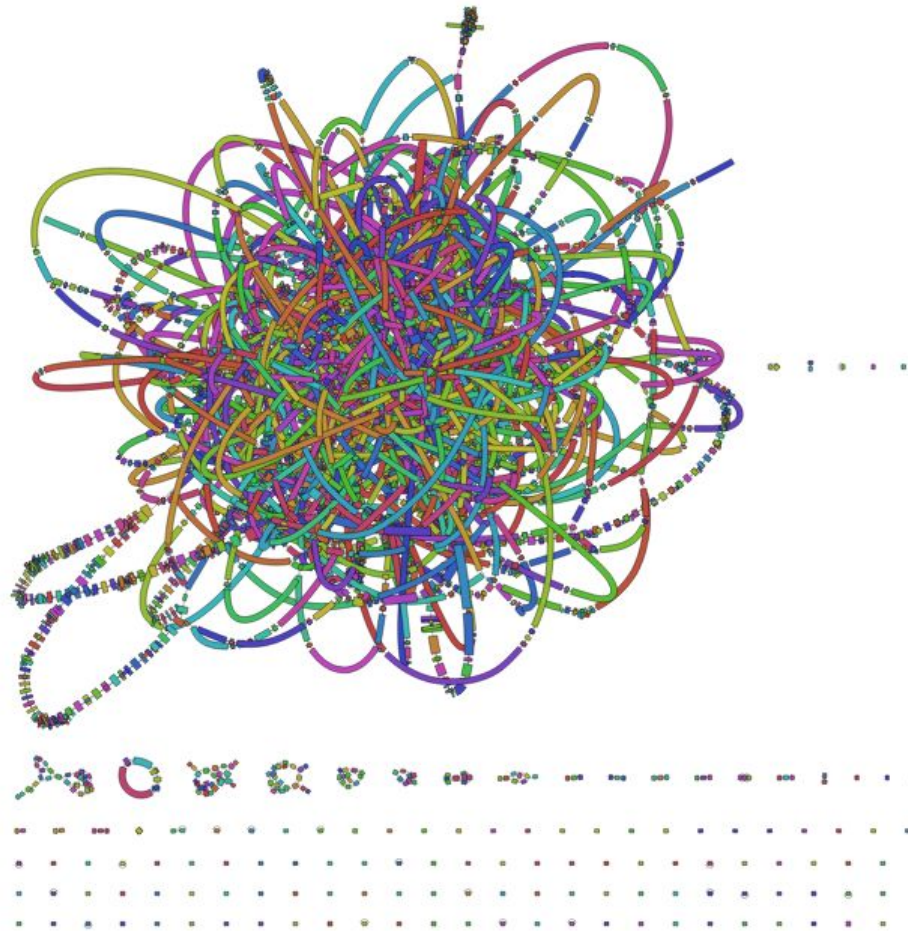
# *K-mers and K*

- No magical value of k
- Depends on
  - read length
  - sequencing error
  - rate of polymorphism
  - coverage
- Some rules:
  - k must be less than the read length
  - k can't be an even number (can produce palindromes)

# K-mer value effects

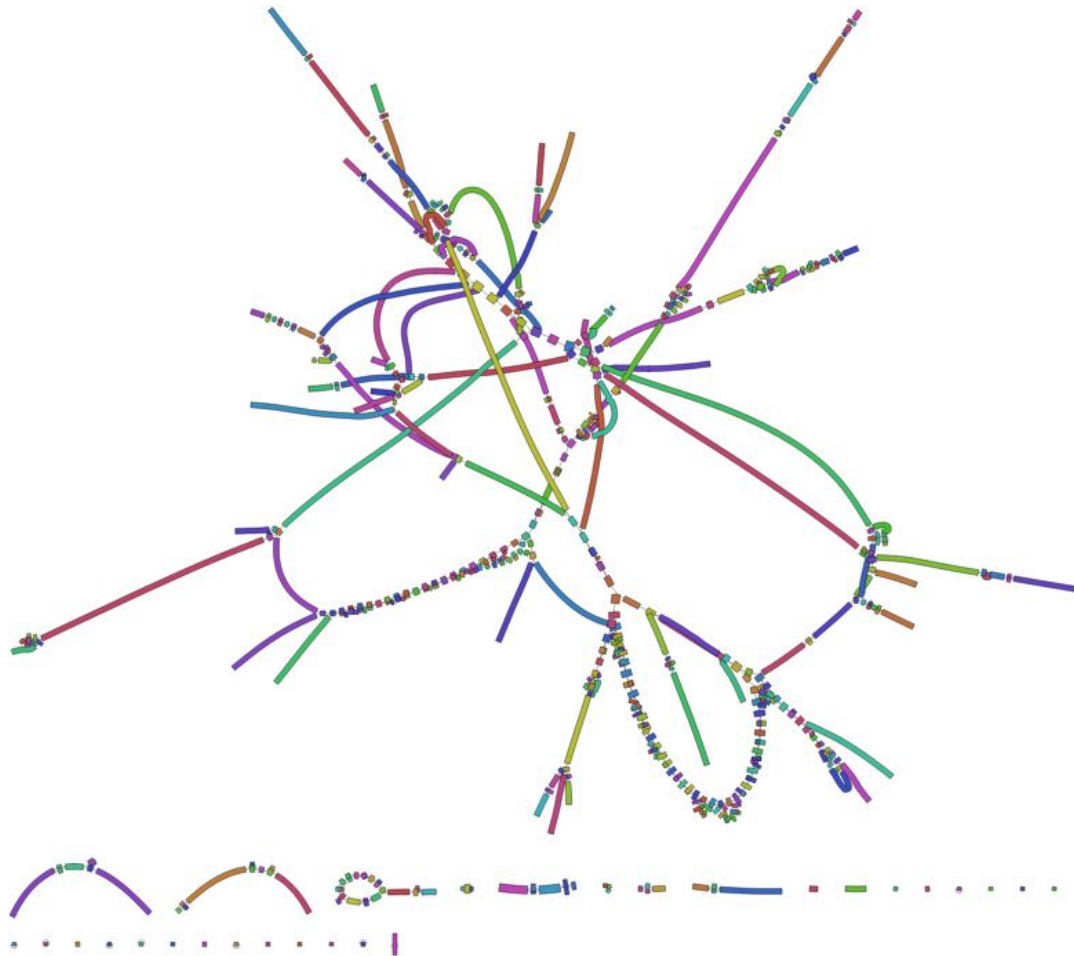
- Bigger K:
  - Solves more repeats
  - Fewer overlaps
  - Lower k-mer coverage
- Smaller k:
  - More overlaps
  - Higher k-mer coverage
- Larger: longer contigs, fewer connections
- Smaller: short contigs with lots of connections

# K-mer size effect: Salmonella, 51

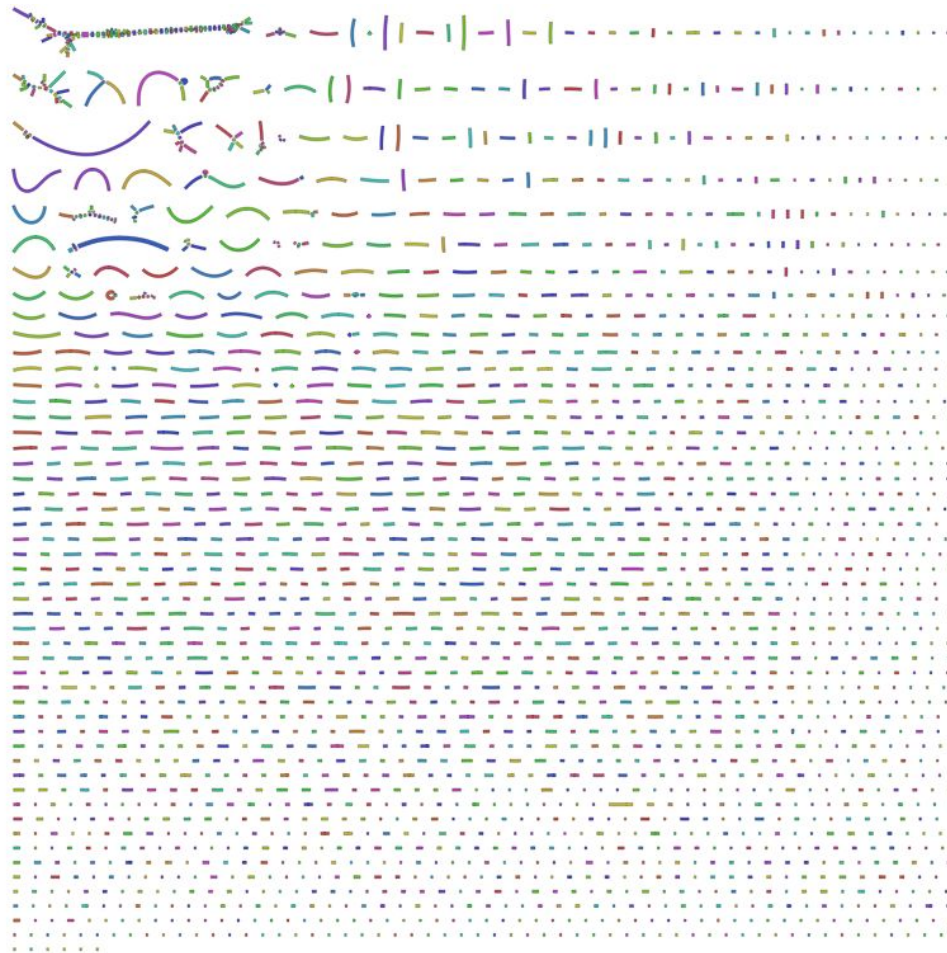




# K-mer size effect: Salmonella, 71



# K-mer size effect: Salmonella, 91



# What do you get?

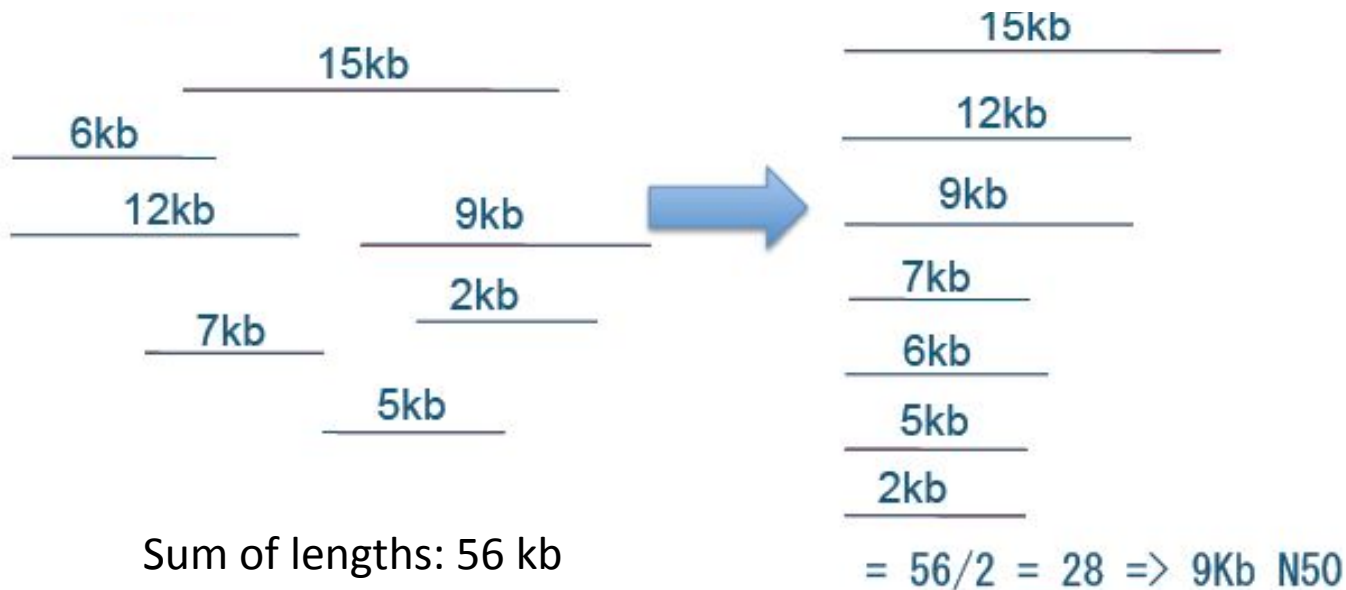


# Metrics

- 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



# 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

- > fewer is better

# 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

- Note: minimum contig length influences N50
- If you 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 =>