Announcements

Lucigen mate-pair data are available – check the wiki

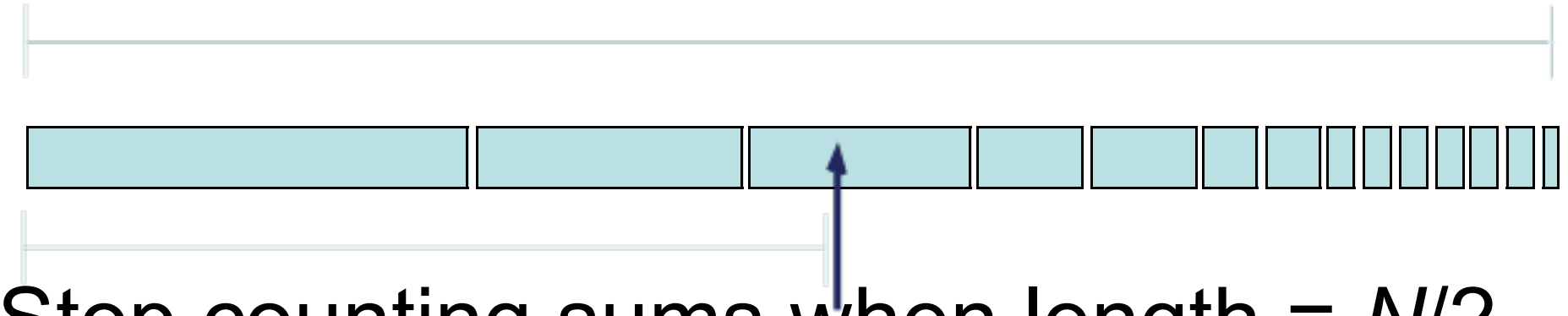
Need to decide what additional data to collect

Your first presentations:
- How does the assembler work (theory)?
- What are the theoretical advantages of the approach implemented in your assembler?
- What was your user experience?
- Some results

Grading

Contiguity

Total genome length = *N*



Stop counting sums when length = *N*/2

N50 is length of this contig/scaffold

Limitations of N50 (or Nx)

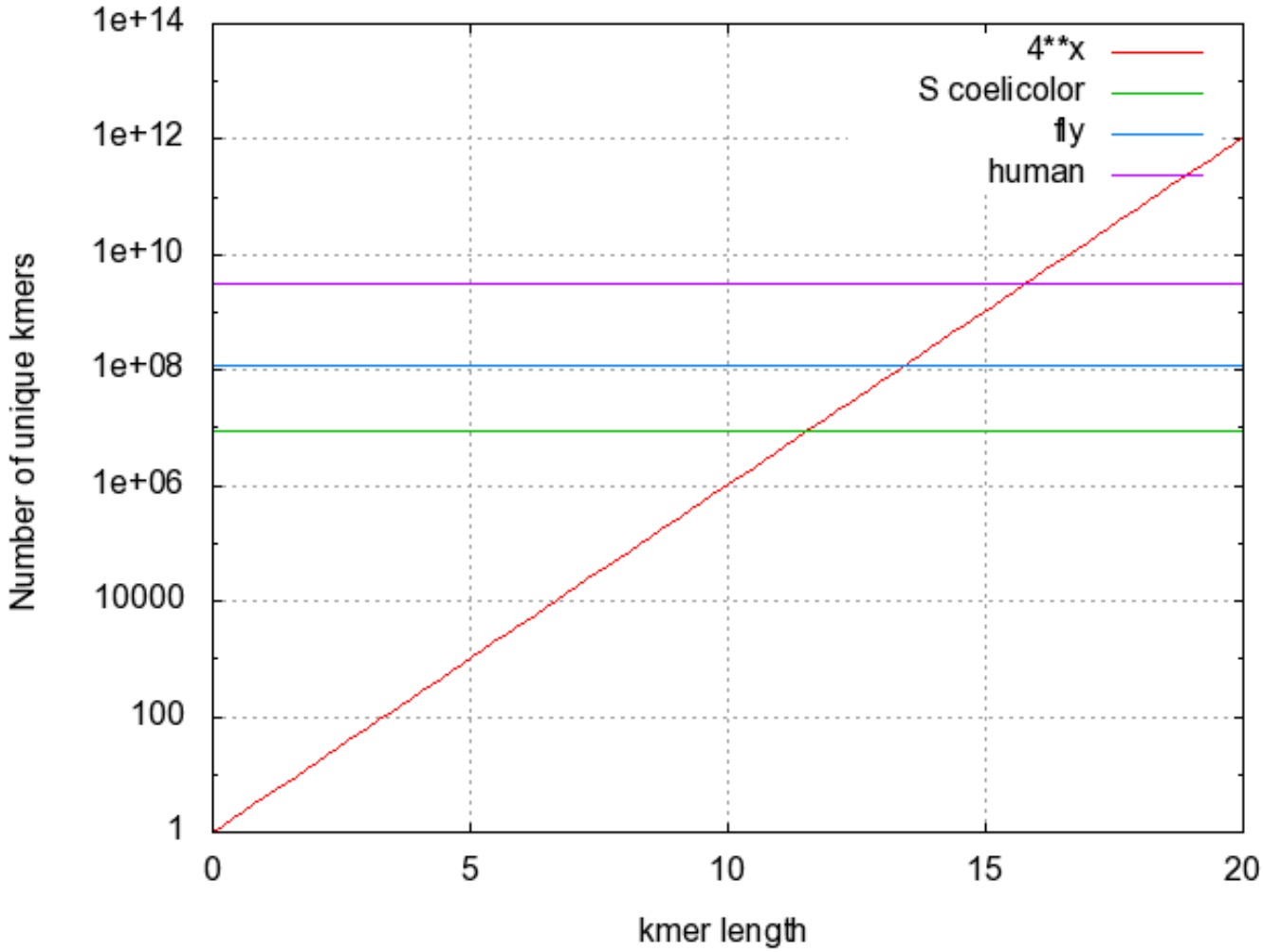
Join everything!

How to check for false scaffolding:
Gene contiguity correctness from RNAseq or other data
Paired-end mapping
Synteny analysis

Theoretical genome uniqueness

K-mers:
words of length *k*

How many *k-mers*
of length *n*?
4n

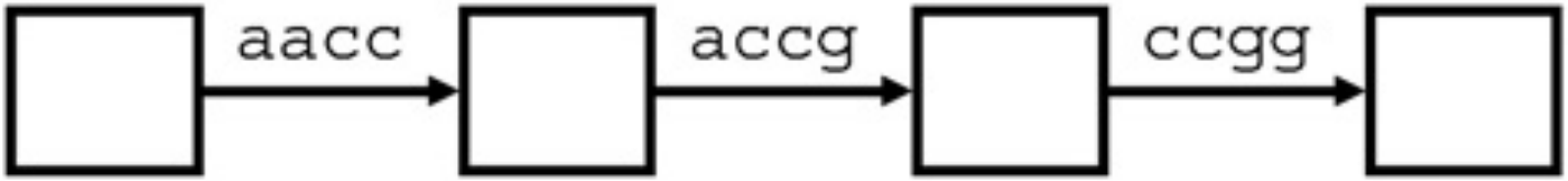


Read 1:AACGTAGCTAC
Read 2:AGCTACGG
Read 3:CGTAGCT

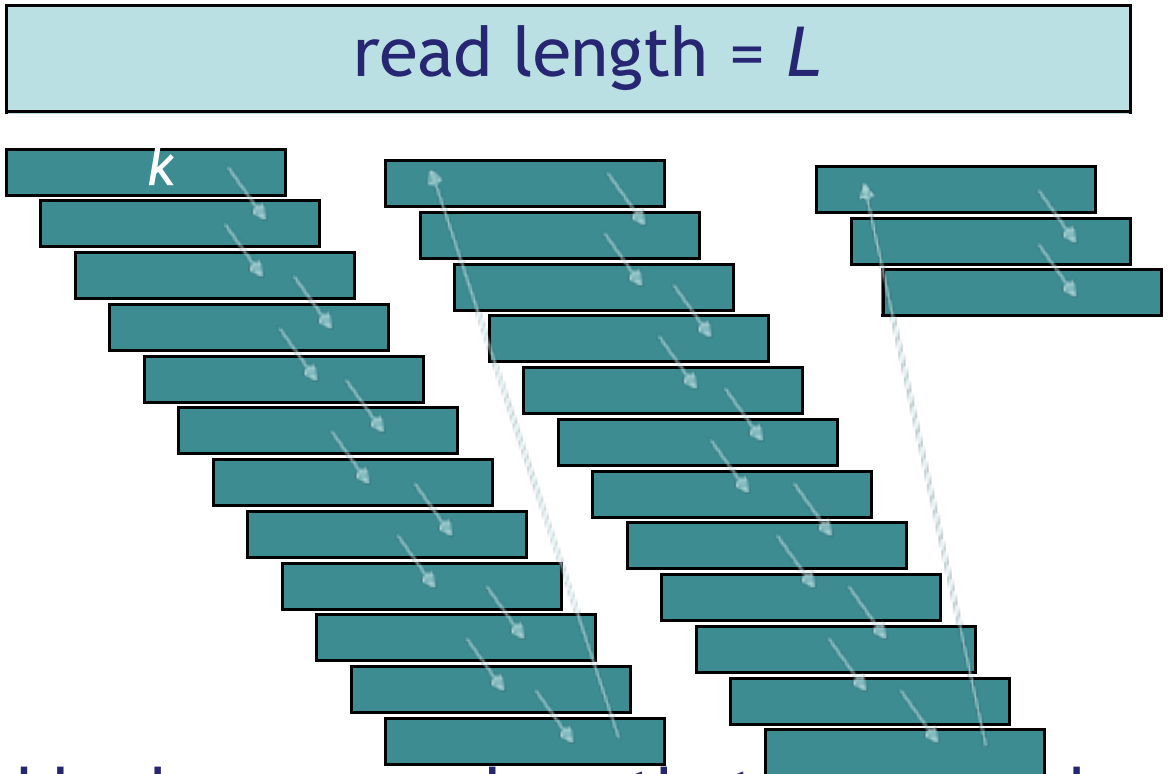
de Bruijn graph

(a) aaccgg

(b) 

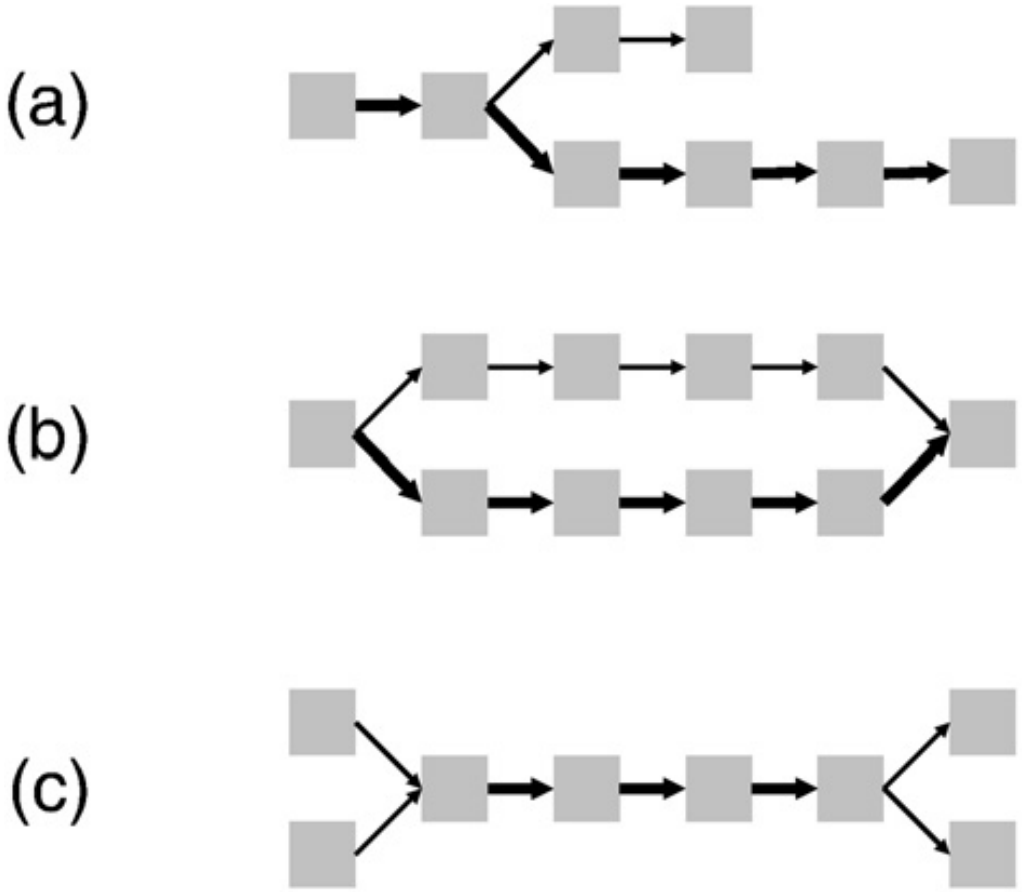
(c) 

Implication of *k*

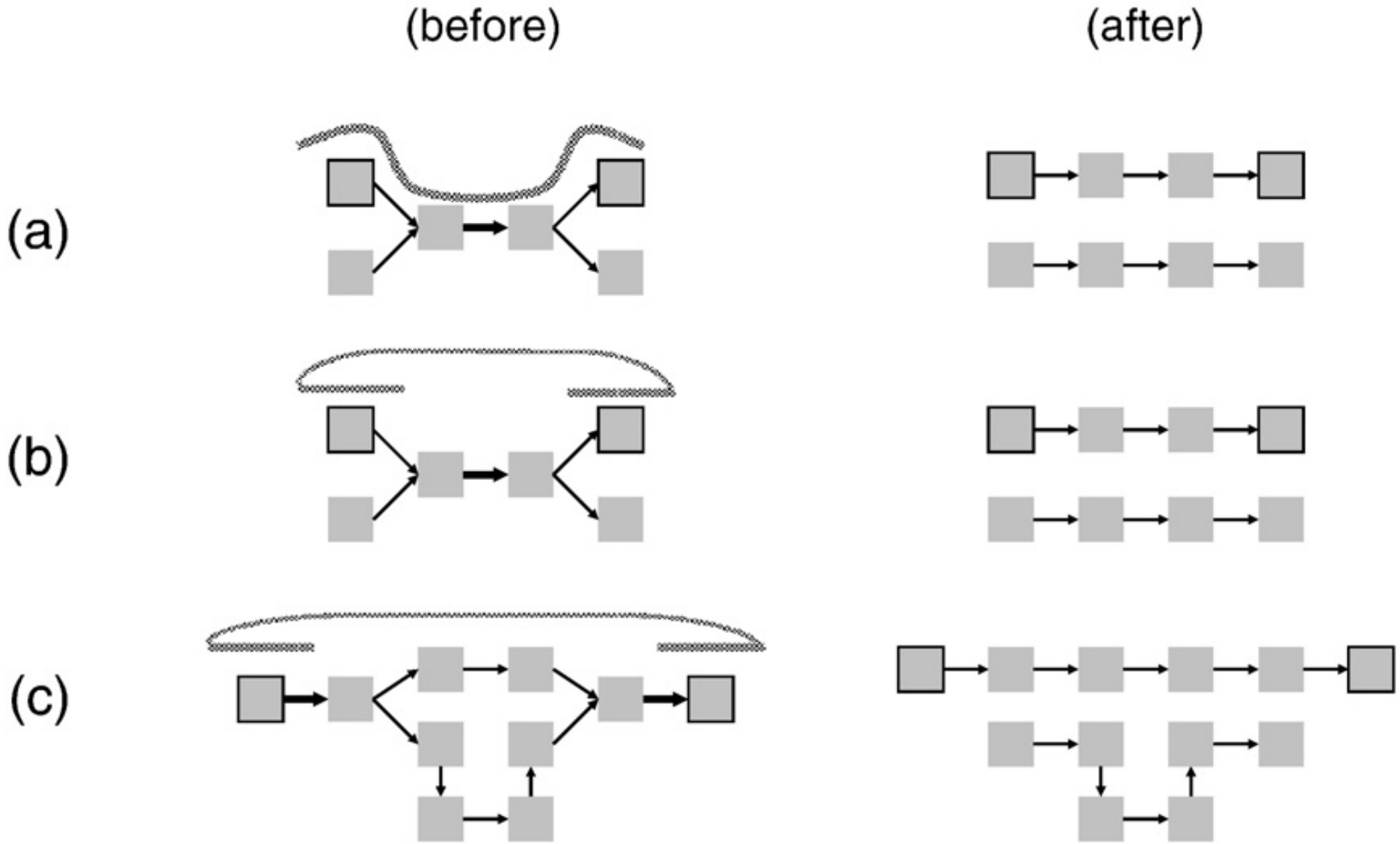


K should be long enough so that most single-copy genome regions
are **unique**
Number of *k-mers* = *L* - *k* + 1
Number of arcs = *L* - *k*
In case of sequencing error: k-mers affected = *k*

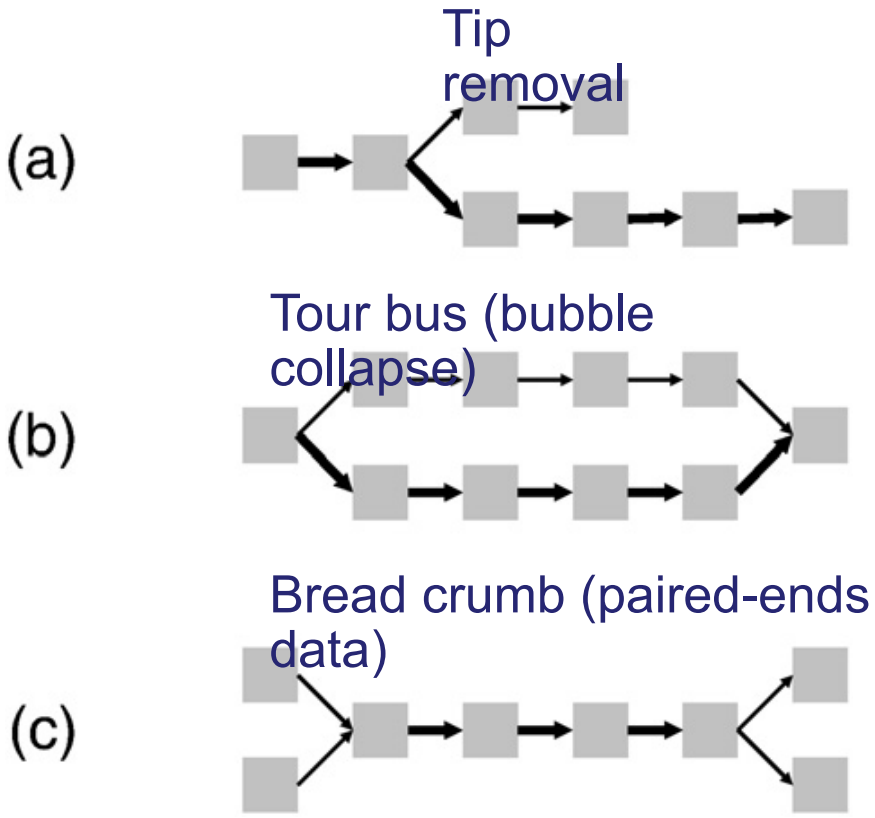
What could possibly go wrong?



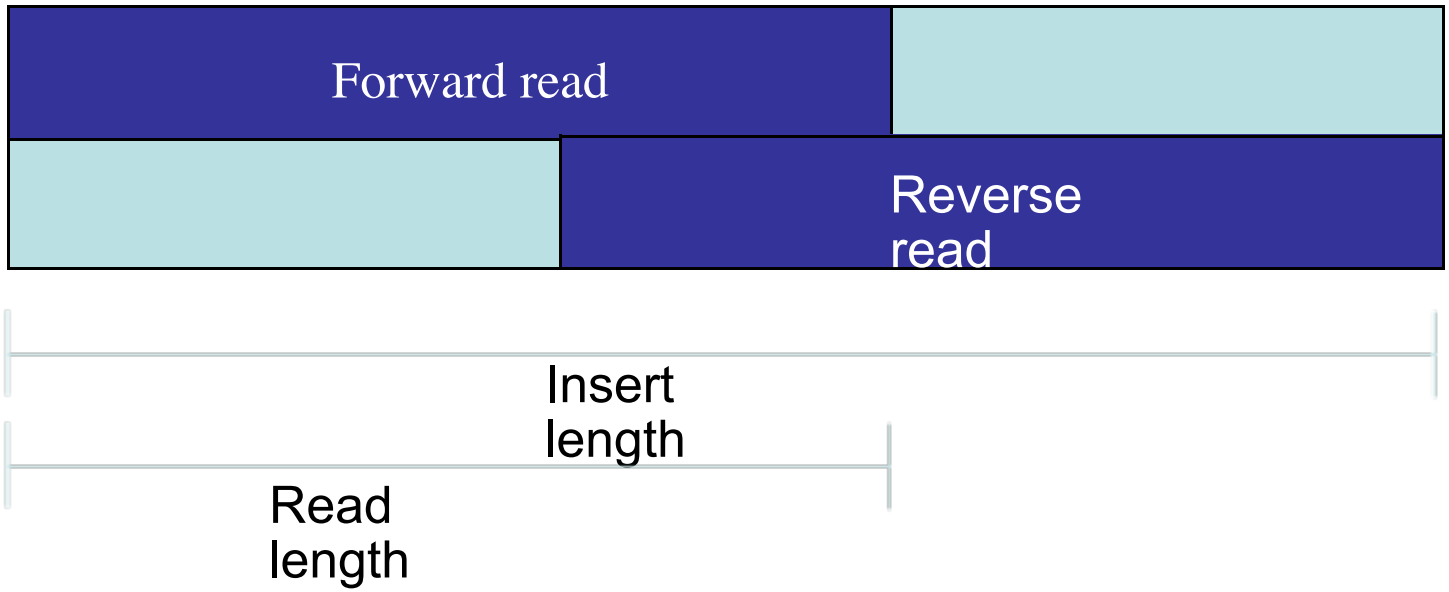
Long reads and/or paired reads can resolve ambiguities



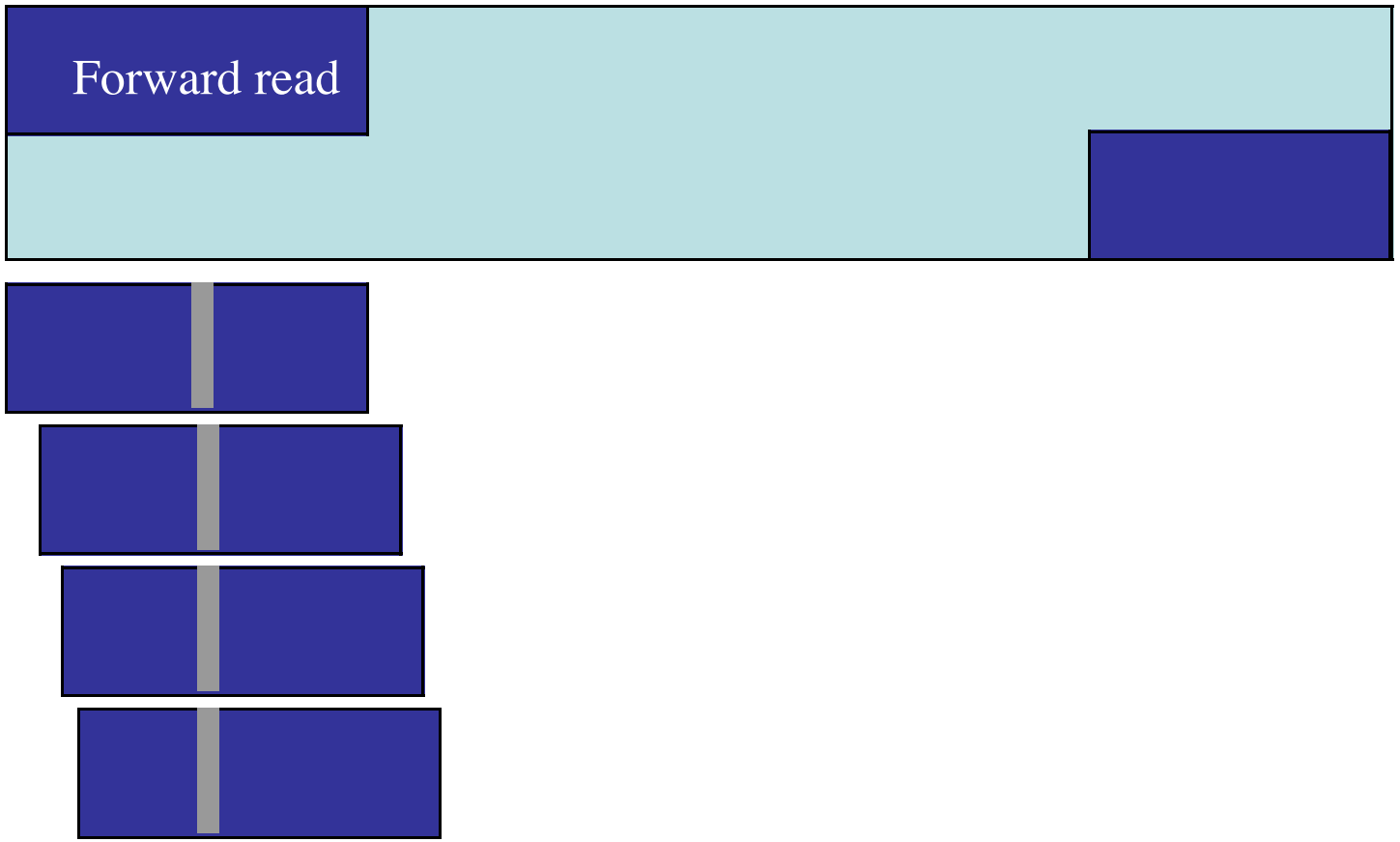
Velvet: a de Bruijn graph assembler



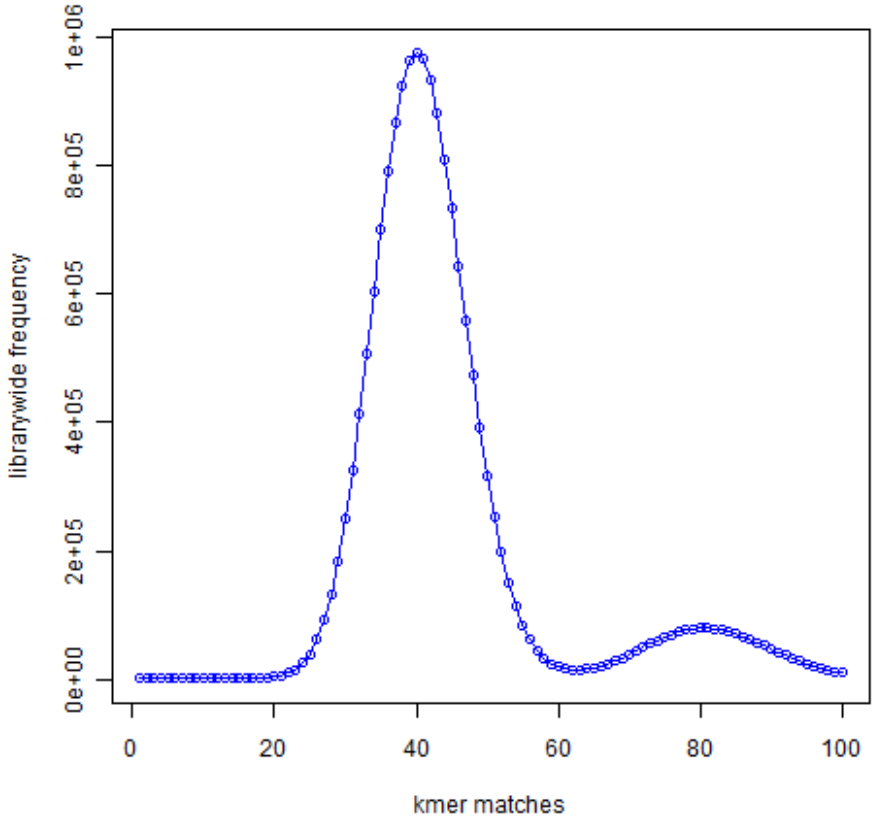
Overlapping read pairs for *de facto* longer reads



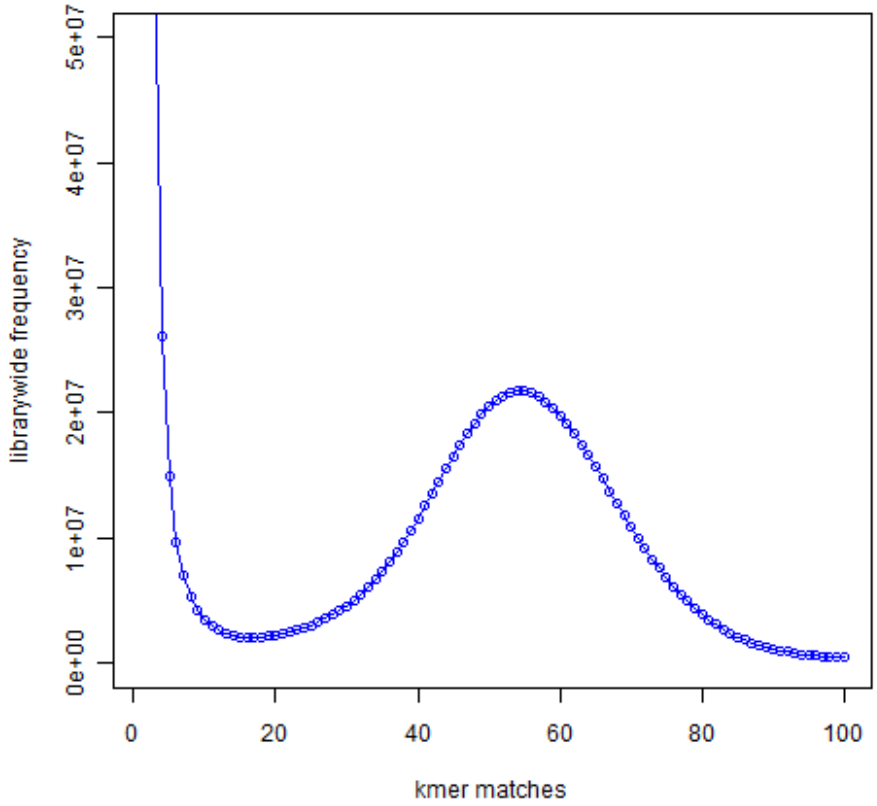
Local error correction by k-mer analysis



K-mer spectra from shotgun genomic data



Error-free data



Real data