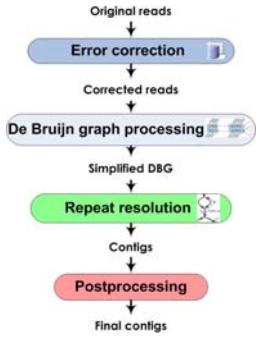
Velvet and SPAdes assembler

How assembly works in practice

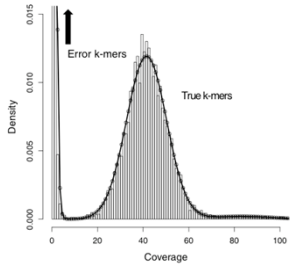


Velvet vs SPAdes

- SPAdes developed to be able to assemble single-cell sequence data
- Single-cell data:
  - Not uniform coverage
- Three main differences between Velvet and SPAdes
  - Error correction
  - Graph construction
  - Graph simplification/resolution
- Other differences too, but won't go into that here

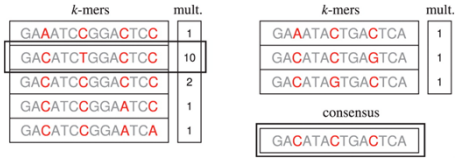
Velvet error correction

- Velvet: expects uniform coverage
- Uses high coverage k-mers to error correct low coverage k-mers



SPAdes error correcton - Hammer

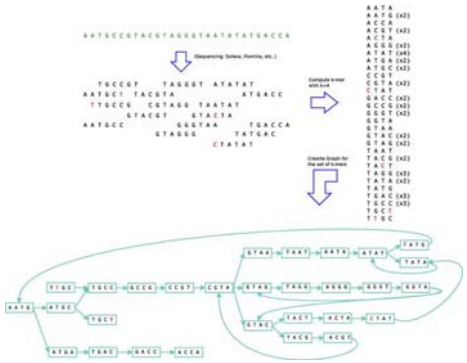
- BayesHammer (Illumina) or IonHammer (IonTorrent)
- How Hammer works:



BayesHammer does the same as Hammer, but looks at the problem probabilistically

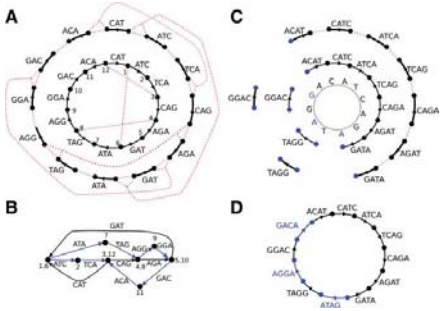
On the left is an example of a typical cluster with good coverage. There are five *k*-mers clustered together, with five loci having mis-alignments. We compute the consensus string (taking multiplicities into account), which we find is already in the cluster (boxed in). All the *k*-mers are then corrected to the consensus. On the right is an example of a common cluster in low coverage regions. The generating *k*-mer was sequenced three times but each time with a single error. There are three *k*-mers in the cluster, but the consensus (boxed in) has not been sequenced and therefore is not in the cluster. Nevertheless, we correct all the *k*-mers to the consensus, allowing Hammer to reconstruct new *k*-mers that are not present in the original data.
Medvedev *et. al.*, Bioinformatics. 2011

Velvet graph construction – fixed size *k-mer*



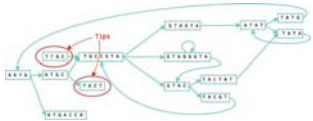
SPAdes: multisized graphs

- Uses several different *k-mer* sizes

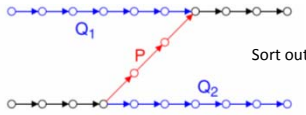


Standard and multisized de Bruijn graph. A circular GENOME CATCAGATAGGA is covered by a set READS consisting of nine 4-mers, {ACAT, CATC, ATCA, TCAG, CAGA, AGAT, GATA, TAGG, GGAC}. Three out of 12 possible 4-mers from GENOME are missing from READS (namely {ATAG,AGGA,GACA}), but all 3-mers from GENOME are present in READS. (**A**) The outside circle shows a separate black edge for each 3-mer from READS. Dotted red lines indicate vertices that will be glued. The inner circle shows the result of applying some of the glues. (**B**) The graph DB(READS, 3) resulting from all the glues is tangled. The three h-paths of length 2 in this graph (shown in blue) correspond to h-reads ATAG, AGGA, and GACA. Thus $READS_{3,4}$ contains all 4-mers from GENOME. (**C**) The outside circle shows a separate edge for each of the nine 4-mer reads. The next inner circle shows the graph DB(READS, 4), and the innermost circle represents the GENOME. The graph DB(READS, 4) is fragmented into 3 connected components. (**D**) The multisized de Bruijn graph DB(READS, 3, 4).
Bankevitch *et. al.*,.J Comput Biol. 2012

Graph simplification



Trim tips

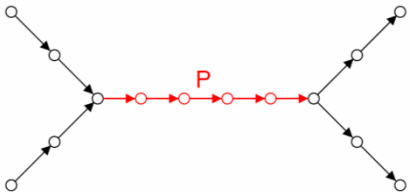


Sort out chimeras

Sort out bulges

- Velvet and SPAdes do these things in similar ways, but: SPAdes needs to keep track of these things in case there is an alternate path in the single cell data

Repeat resolution



How do you sort out which end goes which which end?

Repeat resolution

- Velvet
  - Looks at the reads connecting longer contigs
  - Uses paired read information to “straighten” out the repeats
- SPAdes
  - Uses read pair information
  - Creates a paired de Bruijn graph – each node a pair
  - Much sparser than the “normal” graph