ABySS (Assembly by Short Sequencing)

* distributed representation of a de Bruijn graph (allows parallel computation of the assembly algorithm across a network of commodity computers)
* Generate all k-mers
* Remove Read errors (Follow method of EULER-SR, Velvet, Edena)
  + Remove short “dead-end” branches
  + Complex errors that form small divergent bubbles in the graph are removed and recorded in a subsequent step (repetitive regions, allelic variation in diploid genomes). Remove lower read coverage path, or in case of repeats, collapse into one representative sequence.
* Build Initial Contigs
* Extend contigs by using mate pair information to resolve ambiguities in contig overlaps
* *Found 8.9 Mb of sequence not represented in human ref assembly, 2.4 Mb of which is also not present in alternate human assemblies*

ALLPATHS-LG

* Assembly graph whose edges are sequences, branches represent alternate choices
* Collapses repeats of length >K
* For every 24-mer, compare to all reads containing the 24-mer. Individual reads may be edited if they differ from the consensus of this stack. Conflicting votes on a given base means it is not changed (don’t want incorrect error correction)
* For low coverage regions, use a smaller value of K (go as low as 15, where reads are bounded to lie within a short gap between two other sequences)
* …ATC{A,T}GGTTTTTTT{,T, TT}ACAC … .
* shows SNP and a homopolymer

**GAGE**: results show that data quality, rather than the assembler, has a dramatic effect on the quality of an assembled genome

**Chastity filter**: Illumina’s QC metric. Measure of signal to noise ratio. Ratio of the highest of the four base type intensities to the sum of the highest two (<0.6 is bad). Applied to the first 20 position in the read (regardless of read length), and atmost 1 base is allowed to fail to meet the 0.6 ratio threshold. QC Failure = too many N’s in your reads.

Degree of Contiguity (longer reads have better contiguity): Want assembly with largest contigs and scaffolds.

Once found, this unitig set can be used to correct the spanning long read. Because we are aligning unitigs, rather than short-reads, a consensus approach can no longer be applied to correct the long reads. Instead, the error correction selects a set of short-read unitigs that best covers each long read and uses it as a back- bone for correction.

Correctness of an Assembly

* Find base calling error reads: K-mers occurring only once or twice in a dataset
* Replace the lowest quality base with another base

Scaffold misjoins

* Velvet has a high error rate in this area
* ALLPATHS-LG has both high contiguity and high accuracy

Palindromes: Circular Consensus Sequencing**Repeats can only be resolved by a spanning read or read pair that is uniquely anchored on both sides.**