Assembly evaluation

Karin Lagesen
karin.lagesen@vetinst.no

QUAST uses *nucmer* from MUMmer package

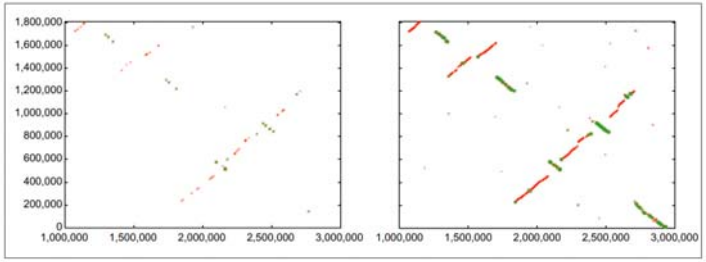


Figure 1
Dot-plot alignments of a 2.9 Mbp chromosome of *A. fumigatus* (x-axis) to a 2.1 Mbp scaffold of *A. nidulans* (y-axis). Left: nucleotide-based alignment with Nucmer. Right: amino-acid-based alignment with Promer. Aligned segments are represented as dots or lines, up to 3,000 bp long in the Nucmer alignment and up to 9,500 bp in the Promer alignment. These alignments were generated by the mummerplot script and the Unix program gnuplot.

QUAST statistics

- Contig size information
- Misassemblies and structural variations
- Genome features found in the assembly
- Variations on N50 statistics

- Visualization

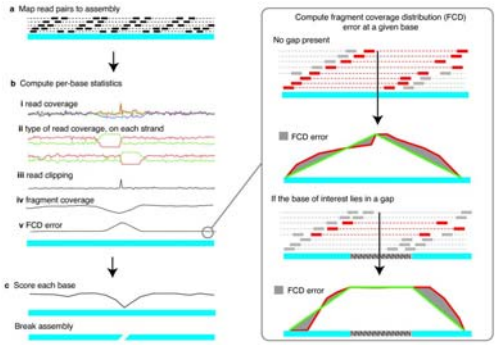
REAPR

- Assembly evaluation
- Assembly “fixing”
- Basic principle
  - Take assembly and reads used to create assembly
  - Map reads to the assembly
  - Examine support for each base
- Two types of errors
  - Local errors: indels/mutations inside a read
  - Structural changes

Fragment coverage distribution

- Reads mapped independently
- Locate “proper pairs”, i.e. reads that map with
  - a normal insert size (estimated from the read sets insert size distribution)
  - map in the “right” direction
- Measure per-base coverage from these fragments (edge to edge)
- Find actual coverage distribution
- Compare to theoretical distribution
- Break up genome in regions with high FCD error

REAPR method



Overview of the REAPR pipeline. (**a**) The input is a BAM file of read pairs mapped to the assembly. (**b**) Statistics are calculated at each base of the genome: (**i**) Read coverage per strand, and any perfect and uniquely mapped read coverage is incorporated; (**ii**) The type of read coverage on the forward (upper plots) and reverse (lower plots) strand: proportion of reads that are properly paired (red), orphaned (green), and in the wrong orientation or exceed the fragment size range (not shown); (**iii**) The number of reads soft-clipped at each base; (**iv**) The fragment coverage, determined by the properly paired reads; (**v**) FCD error, taking into account the presence of a gap. Boxed are: FCD calculation at a given base. The fragments covering that base, shown in red, are used to construct a fragment depth plot (red). The FCD error is the area (grey) between the observed (red) plot and ideal plot (green). Since no read can map to a gap in the assembly, the calculation is corrected when a gap is present. (**c**) The statistics at each base are used independently to assign a score to each base of the assembly and also to break the assembly at scaffolding errors.

Hunt *et. al.* Genome Biology 2013

Corrected assembly statistics

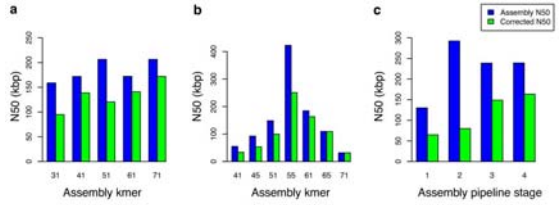


Figure 4**N50 statistics of various assemblies before and after correcting with REAPR**. Blue bars show the N50 of the assembly input to REAPR, green bars show the corrected N50. (**a**) *De novo* assemblies of *S. aureus*. (**b**) *P. falciparum de novo* assemblies. (**c**) *B. pahangi* assemblies at four different stages of the genome project.