From high-throughput sequencing read alignments to confident, biologically relevant conclusions with Nesoni

Nesoni is open source software for analysis of high-throughput sequencing data based on alignment to a reference. We use this software for analysis of Illumina, 454, and SOLID sequencing data, largely from prokaryotes. Prokaryotic genomes are smaller than those of eukaryotes, but there is greater within-species diversity, and a more rapid rate of mutation. When studying prokaryotes we find we are more often interested in the differences between two newly sequenced strains than in the differences between a sequenced strain and a well polished reference sequence. Nesoni can detect base substitutions, insertions and deletions between between two or more sequenced strains.

Nesoni includes a series of checks to ensure read alignments and consensus calls are unambiguous, allowing confidence that any differences it finds are real. Per-base evidence tallies are also carried through the various steps, allowing a manual assessment of the trustworthiness of any differences found.

N-way comparison of consensii

Any number of sequencing runs may be compared amongst themselves or with the reference sequence. Where a consensus was not called, no difference is reported. For example, if an insufficiently pure mixture of bases was seen in one or other strain, a difference will not be called even if the most frequent base is different between the two strains. The result of an n-way comparison may be output in a format that can be read by the program SplitsTree4, allowing phylogenetic analysis.

Protein level consequences

Given a genome annotation in Genbank format, Nesoni can produce a list of protein level changes between the reference and a sequenced strain. These might include amino acid changes and changes to start and stop codons, and might be due to substitutions, or insertions or deletions (whether frame-shifting or frame-preserving).

Only changes where a consensus was called are reported, avoiding reporting of spurious changes where the depth or purity were insufficient.

Example applications

- A spontaneous mutation of a strain of Pasteurella multocida which lacked a polysaccharide capsule was investigated by Jason Sheen and John Boyce at Monash University. The parent and mutant strains were sequenced using an Illumina GAII sequencer, and aligned to the PM70 reference sequence. Comparison of evidence tallies using Fisher’s Exact Test as described below identified three significant SNPs, two of which were silent, the third being a mutation to the regulatory gene Pus. A plasmid containing an intact copy of Pus was used to transform the mutant strain, restoring the capsule.

- Two clinical isolates of Staphylococcus aureus were obtained from a patient, one of which was a Small Colony Variant (SCV). Both strains were sequenced using an Illumina GAII sequencer. Aligning against the reference strain COL, and again using Fisher’s Exact Test, several significant SNPs and insertions were found. One of these SNPs was introduced into the non-SCV strain, and the modified strain was found to have some but not all of the phenotypic features of the SCV strain [1].

References