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

Paul Harrison* and Torsten Seemann
Victorian Bioinformatics Consortium, Monash University

* paul.harrison@med.monash.edu.au

http://vicbioinformatics.com/software.shtml

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.

Example applications

- A spontaneous mutation of a strain of *Pasteuralla multocida* which lacked a polysaccharide capsule was investigated by Jason Steen 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 *Fis*. A plamid containing an intact copy of *Fis* 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]

From alignments to evidence tallies

Alignment: Nesoni uses the **SHRiMP** read aligner [2], which uses Smith-Waterman alignment around potential sites identified by hits to spaced seeds. SHRiMP is able to align reads in the presence of substitutions, insertions, and deletions.

↓

Fidelity filter: For each read, Nesoni discards all alignments scoring less than a specified percentage of the highest scoring alignment. (If reads are paired then, instead of individual alignments, Nesoni considers all possible validly oriented and spaced pairs of alignments.)

↓

Monogamy filter (optional): If more than one alignment of a read passes the fidelity filter, the mapping of the read to the reference is ambiguous, and all alignments from that read are discarded.

↓

Evidence tallying: From the chosen alignments, Nesoni collects the number of times each base (possibly a deletion) was seen at each position in the reference. Nesoni also collects the number of times an insertion (of one or more bases) was seen between one position and the next in the reference.

From evidence tallies to consensus, or not

For each position in the reference, the most frequently seen base (possibly a deletion) is called as the consensus, unless:

Depth filter: If the depth of coverage is below a user specified cutoff, a consensus is not called.

Purity filter: If the most frequent base was seen less than a user specified proportion of the time, an IUPAC ambiguity code is given.

The same procedure is used to call a consensus on insertions.

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** [4], allowing phylogenetic analysis.

Directly comparing evidence tallies

For each position position in the reference, the mixture of bases (and possibly deletions) in two runs can be compared using Fisher's Exact Test. The mixture of insertions seen between positions in the reference can be similarly compared.

This has two useful features: it produces significance levels, and it can detect differences in mixture as well as clear substitutions, insertions, and deletions.

Sequence	Position in reference	Change type	Reference	Strain 1	Strain 2	p-value
...						
gi|57650036|ref|NC_002951.2|	608428	substitution	C	"T"x302 "C"x1 "G"x1	"C"x387 "A"x5	1.06E-203
gi|57650036|ref|NC_002951.2|	1241187	insertion-before		"CAA"x60 "-"x50	"-"x152	6.38E-029
gi|57650036|ref|NC_002951.2|	1399358	insertion-before		"TGT"x93 "-"x50	"-"x235	5.89E-052
gi|57650036|ref|NC_002951.2|	1719916	substitution	A	"T"x176	"A"x228 "C"x1 "G"x1	5.56E-120
gi|57650036|ref|NC_002951.2|	1906875	substitution	A	"A"x72	"G"x105 "A"x78 "T"x2	2.94E-021
gi|57650036|ref|NC_002951.2|	1906883	substitution	G	"G"x121	"A"x217 "G"x119 "T"x18	1.02E-044
gi|57650036|ref|NC_002951.2|	1906913	different mix	T	"T"x197 "A"x1	"T"x215 "-"x131 "G"x3	1.26E-031
...						

Fig 2. Example output from direct comparison of tallies.

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.

Depth of coverage plots

These plots are useful for identifying missing sequences in the sequenced strain and identifying copy number variations.

If paired end reads are used, a second plot is produced which includes the span between the reads.

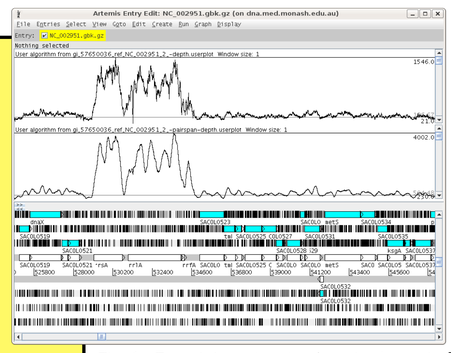

Fig 1. Examining depth plots in Artemis [3]

References

[1] Gao, W., Chua, K., Seemann T., Harrison P.F., Newton, H., Hartland, E.L., Holmes, N., Davies, J.K., Strinear, T.P., and Howden, B.P. (2009) New Mechanism of Small Colony Variant Formation in a Clinical *Staphylococcus aureus* Associated with Persistent Infection. Poster presented at BacPath 10.
[2] http://compbio.cs.toronto.edu/shrimp/
[3] http://www.sanger.ac.uk/Software/Artemis/
[4] http://www.splitstree.org/

Victorian Bioinformatics Consortium

MONASH University