ngskit4b Aligner SNP Detection Application Note

Release 0.4.6

CAUTION – this document has not been updated to include new functionality added since original cloning of the BioKanga 4.2.0 documentation

Overview

When aligning, ngskit4b can also process the alignments for single nucleotide polymorphic (SNP) calling using several different experimenter specified thresholds. These are :

a) Minimum coverage (number of reads with alignments covering a potential SNP loci)
b) Minimum percentage of bases at a putative SNP loci which are non-reference
c) Max allowed P-value derived by summing Pr(k=k) accounting for the local sequencing error rate
d) Using Benjamini-Hochberg QValue to rank the called SNPs as a FDR control

The processing flow for SNP detection and acceptance is as follows:

Aligned reads are stacked in sense orientation (antisense aligned reads are reverse complemented) by ascending alignment loci, so at any given loci both the coverage and stacked base composition counts can be easily determined.

For each chromosome a global sequencing error rate (GSER) is calculated as being the total number of read alignment required substitutions (TotChromReadBasesSubs) divided by the total length of all aligned reads (TotChromReadBases) to that chromosome with 0.01 as the floor.

Each loci along the length of the chromosome is then iterated and the following processing on that loci is executed:

If the coverage at the loci currently being processed is less than that specified by the experimenter then that loci is skipped and next loci will be processed.

If the proportion of non-reference bases at the loci currently being processed is less than that specified by the experimenter than that loci will be skipped.

A local sequencing error rate (LSER) is calculated over a 101bp window bracketing (50bp 5' and 50bp 3' relative to loci being processed for SNP but excluding the putative SNP loci). This LSER is calculated by dividing the total number of read alignment required substitutions in the window by the total length of all aligned bases within the window (excluding bases stacking at the putative SNP loci).

If the LSER is more than 0.2 then the currently processed loci is skipped (local context too noisy) and next loci will be processed.

The P-value (1.0 – binomial(n,k,p)) is then calculated for the current loci using the sum of Pr(K = k) as nCk * p^k * q^(n-k) for K = 0 up to K = k where k = number of non-reference bases, n = total bases, and p = LSER.

If the P-value is above that specified by the user then that loci is skipped and the next loci will be processed.

Putative SNP loci which meet the forgoing criteria are then deemed as accepted and will be reported.

When all SNPs have been accepted for a chromosome, these SNPs are then ranked using Benjamini-Hochberg with the highest rank = 999 and the lowest = 1.

.