

ngskit4b Aligner SNP Detection Application Note

Release 1.0.0

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 then 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 ($1.0 - \text{binomial}(n, k, p)$) 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.

Standard SNP CSV Format

Standard ngskit4b alignment generated SNP files are generated in the following CSV format and this is the expected format for SNP files processed as input by 'ngskit4b snpmarkers':

| Column Header | Meaning | Example |
|-----------------------|----------------------------------------------------------------------------------------------------------------|---------------|
| "SNP_ID" | Monotonically increasing unique integer identifier | 1234 |
| "ElType" | Element Type – currently always "SNP" | "SNP" |
| "Species" | Targeted species | "GSS Wheat" |
| "Chrom" | Name of chrom/contig/sequence on which SNP was identified | "Chr1AL_5678" |
| "StartLoci" | 0 based loci on chrom/contig/sequence on at which SNP was identified | 13579 |
| "EndLoci" | Same as StartLoci for SNPs | 13579 |
| "Len" | SNP so length always 1 | 1 |
| "Strand" | SNP is reported relative to sense strand | "+" |
| "Rank" | Benjamini-Hochberg rank, highest confidence assigned rank 999 and lowest confidence assigned rank 1 | 679 |
| "PValue" | Probability of SNP as false positive using $P\text{-value} = 1.0 - \text{binomial}(n,k,p)$ | 0.000000 |
| "Bases" | Total number of bases stacking at the SNP loci from all reads covering that loci | 2091 |
| "Mismatches" | Of the bases stacking, the number which were not matching the targeted reference sequence | 941 |
| "RefBase" | The nucleotide in the reference sequence at the SNP loci | "T" |
| "MMBaseA" | Number of mismatch stacking A bases | 1 |
| "MMBaseC" | Number of mismatch stacking C bases | 934 |
| "MMBaseG" | Number of mismatch stacking G bases | 6 |
| "MMBaseT" | Number of mismatch stacking T bases - in this example "T" is the reference but still reported as if mismatches | 1150 |
| "MMBaseN" | Number of mismatch stacking N bases | 0 |
| "BackgroundSubRate" | Local sequencing error rate (LSER) | 0.015647 |
| "TotWinBases" | Total number of bases over which LSER was calculated | 113054 |
| "TotWinMismatches" | Total number of mismatch bases in LSER | 1769 |
| "MarkerID" | Not used in SNP reporting | 0 |
| "NumPolymorphicSites" | Not used in SNP reporting | 0 |

Extended SNP CSV Format

An extension to the standard SNP CSV format is now also available (since release 1.0.0), in this format codon 3x frame shifted counts are also reported enabling down stream applications to analyse codon mixtures at SNP sites.

Each frame shift consists of 65 fields, the first frame starts at -2 relative to the SNP loci, second frame at -1, and 3rd frame at 0. The first field in a frame contains the 3 reference codon bases for that frame. The next 64 fields in the same frame contain counts of all codon tri-mers starting from AAA up to TTT discovered in reads aligned at that frame shift loci. The first 23 fields in the SNP file remain the same in both the standard and extended SNP CSV files, the extended fields follow on from the standard.

It is important to note that the total number of codon numbers in any frame may be less than the number of bases covering the SNP loci – this is because frames containing any indeterminate 'N' bases are not included in codon counts nor are frames spanning over read ends.

To enable this new feature, specify the output SNP file as having a '.csvx' file extension to 'ngskit4b kalign', resultant SNP CSV files will be renamed back to having '.csv' extensions so down stream processing modules will not be impacted.

Standard SNP CSV files contain 23 fields, the new extended SNP CSV files contain 218 fields (23 + 3x65) fields. It is this difference in number of fields which ngskit4b processes use to discriminate between the standard and extended SNP CSV formats. For example, 'ngskit4b snpmarkers' will accept either the standard or extended CSV format but only uses the first 23 fields.