ngskit4b SNP Cultivar Marker Association Application Note

Release 0.9.6

Overview

Note: In this application note the term species is used interchangeably with the term cultivar, cultivar is preferenced, but no distinction is made between their usage.

ngskit4b offers a processing module 'ngskit4b snpmarkers' allowing cultivar specific SNPs identified by the 'ngskit4b align' processing module (application note 'ngskit4b SNP Calling Application Note') can be associated by SNP loci using user defined thresholding. However, the intent is that the resultant generated SNP marker loci association file be further processed with experimenter downstream scripting, and marker loci not meeting the experimental requirements be removed. To facilitate this downstream processing, another ngskit4b module 'ngskit4b csvm2sqlite' provides for the import of the SNP marker loci association file directly into a SQLite database thus enabling the utilisation of SQL queries by the experimenter in the filtering process. This application note primarily covers the 'snpmarkers' processing module.

SNP Cultivar Marker Loci Association Parameterisation

With the 'ngskit4b snpmarkers' processing module, the primary inputs are expected to be 'ngskit4b align' generated SAM/BAM alignment and CSV SNP call files for each cultivar. It is expected that each cultivar of interest has been independently aligned to a common reference species assembly using the same set of alignment SNP thresholding parameterisation. The aligner generated SAM/BAM alignments and SNP calls are loaded into 'ngskit4b snpmarkers' with the experimenter being provided with access to the following thresholding for controlling the association and subsequent reporting of cultivar association marker loci –

a) "-b<mincovbases>"
   Only consider as an accepted cultivar specific alignment reported SNP if there were at least this number of covering bases. The default is 5, with an accepted experimenter specified range of between 1 and 10,000.
b) "-p<maxpvalue>"
   Only consider as accepted an alignment reported SNP if the SNP has a P-Value which is at most this P-Value. The default is 0.05, but the experimenter can specify a range of 0.0 (no limit) up to a maximum of 0.25.
c) "-P<snpmajorpc>"
   Only accept as a putative SNP loci for a given cultivar if the most abundant base at that loci is at least this percentage of the total cultivar coverage without consideration of the lesser abundance bases at that same loci. Defaults to 50% but can be specified to be between 15 and 90%.
d) "-z<mintotcntthres>"
   Only report an otherwise accepted SNP marker loci if all cultivars have at least this number of total bases covering the SNP marker loci. The default is 0 (any coverage is accepted), and can be experimenter specified up to a maximum of 10,000.

e) “-a<altspeciesmaxcnt>”
Only report an otherwise accepted SNP marker loci if no other cultivar has more than this number of counts at the SNP marker loci. Defaults to 0 for no limit, and can be specified to be up to 10,000 maximum.

f) “-Z<mincovspecies>”
Do not report an otherwise accepted SNP marker loci unless at least this number of cultivars have an alignment SNP at the marker loci. By default at least 1 species is required to have alignment generated SNP, and the maximum required can be specified up to the number of cultivars being processed.

g) -m, --mode=<int>
The researcher can request that marker loci be reported for all nucleotide polymorphic variations either inter-cultivar or relative to reference; or only report variations where the SNPs are inter-cultivar.
Marker reporting mode:
0 - SNP markers if either inter-species/cultivar or relative to reference (default)
1 - Report SNP markers only if inter-species/cultivar differences

h) -R, --relgenomes=<relgenomes>
Each aligned cultivar is required to be specified as individual triplet parameterisations; ‘-R’ is used to specify the name of the sequenced cultivar or species, ‘-I’ the resultant alignments file generated, and ‘-i’ the resultant SNPs file from the alignment of a specific cultivar readset against a common referene. For example if two cultivars (‘Cult1’ and ‘Cult2’) had been aligned to a reference (‘Ref1’) then the command line parameters would include the following:
“-rRef1 –RCult1 –iCult1.snps –ICult1.bam –RCult2 –iCult2.snps –ICult2.bam”
Note that a maximum of 1000 cultivars are supported

i) -i, --insnps=<file>
Load SNPs from this file for each individual cultivar. The cultivar specific SNPs file must be in the default CSV format as generated by ‘ngskit4b align’.

j) -I, --inaligns=<file>
Load alignments from this individual cultivar file. Alignment file must be in ‘SAM’ or ‘BAM’ format.

Discussion

The SNP calls generated by ‘ngskit4b align’ for a given cultivar NGS readset are at targeted species loci sites deemed as heterozygotic according to the alignment SNP threshold parameterisations utilised. The reported alignment SNPs contain the loci of the SNP relative to the targeted species assembly sequence designation (chromosome or some other sequence identifier) plus the counts of each nucleotide (‘A’, ’C’, ’G’, ’T’, ’N’) stacking at the reported SNP loci. The ‘ngskit4b snpmarkers’ processing module is accepting the alignment SNPs plus read alignments for the experimenters cultivars of interest as the primary inputs, and subject to the ‘ngskit4b snpmarkers’ module thresholds specified, will generate a SNP marker association file containing at each accepted SNP marker loci the base counts for each input cultivar at that species loci. Thus a given generated output row will contain the targeted sequence loci plus the base ‘A’, ‘C’, ‘G’, ‘T’, ‘U’ counts for each cultivar for which there were one or bases from aligned reads stacking at that targeted sequence loci. If the input alignments from

'ngskit4b align' were not in SAM or BAM format, perhaps they were in BED format, then the base counts for a non-SNP cultivar at a given loci are inferred as being the coverage at that loci with the coverage attributed to the same base as the targeted reference loci base. If the input cultivar alignments were in SAM or BAM format which was containing the aligned read sequences, then the base counts for non-SNP cultivars will be counts derived from the actual reads stacking at the targeted reference loci. The 'ngskit4b snpmarker' generated CSV file contains columns which identify for a given reported loci the source of each cultivar specific base counts – 'S' if from actual SNP call, 'I' if imputed from non-SNP aligner called read coverage or actual counts if SAM/BAM alignments were utilised.

Kit4b SNP Marker Association Internal Processing

a) Individual cultivar specific aligner generated SNP files ("-i, --insnps=<file>") are parsed and loaded. SNPs from an individual cultivar are only retained if they meet the following thresholds:
   a) The total number of bases at the SNP loci is at least the "-b<mincovbases>" threshold which defaults to 5 bases required.
   b) "-p<maxpvalue>"
      The SNP has a P-Value which is at most the "-p<maxpvalue" threshold which defaults to 0.05.

Next the cultivar specific alignments ("-I, --inaligns=<file>") are parsed and loaded. For each SNP loci which was retained (sufficient covering bases and P-value accepted) then if none of retained SNPs (one or more cultivars have the same SNP loci) was one accepted from the current cultivar being processed then the coverage and base counts for the current cultivar are imputed from the cultivar specific alignments which may result in 0 base counts if there were no alignments.

When all imputed base counts have been determined then counts are checked to ensure they meet the following SNP marker association reporting thresholds –

a) All cultivars at a given SNP loci must have total coverage in base counts which are at least that specified with "-z<mintotcntthres>", by default this is 0 so even if only one cultivar (must have been an alignment SNP called cultivar) has any attributed counts then that loci will be reported subject to meeting the other thresholds.
b) All non-SNP cultivars at an otherwise accepted SNP marker loci must have no more than "-a<altspeciesmaxcnt>" base coverage. Defaults to 0 for no limit.
c) There must be at least "-Z<mincovspecies>" cultivars with accepted aligner reported SNPs at a putative marker loci. By default just 1 cultivar is required to have an alignment generated SNP at any marker loci for that loci to be reported provided all cultivars meet the other reporting thresholding requirements.

Expected SNP Loci Input File Format

ngskit4b alignment generated SNP files are generated in the following CSV format and this is the expected format for SNP files processed 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-value = 1.0 – 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

Generated SNP Marker Association Report File

The generated SNP marker association report file is in CSV (Comma Separated Format) with a header row identifying the columns and subsequent rows for each reported SNP marker. In addition to the targeted sequence loci at which a SNP marker has been called there is, for each cultivar, specific meta-data supplied which post-processing scripts can utilise if applying more stringent filtering to gain more confidence and specificity in the SNP marker calls.

Column	Description
<Name>:TargSeq	The name of the target assembly sequence (could be chromosomal, transcript, contig name etc.) containing the loci at which the SNP Marker is being called
Loci	The 0 based loci on <Name>:TargSeq at which the SNP marker is being called
TargBase	The target assembly sequence base at the Loci
NumSpeciesWithCnts	Number of cultivars with one or more read bases covering the Loci
	Following repeats for each cultivar; <Cultivar> identifies the cultivar
<Cultivar>:CntsSrc	Source of the cultivar specific base counts: 'S' SNP call, 'I' imputed from read alignments
<Cultivar>:Base	The cultivar canonical base at Loci; canonical base as being that base which has/ties the highest proportion of base counts. If two or more bases tie for the highest proportion then the canonical base is randomly chosen from these bases.
<Cultivar>:LSER	Local sequencing error rate
<Cultivar>:BaseCntTot	Total count of all bases for this cultivar which are covering the loci
<Cultivar>:BaseCntA	Count of 'A' bases
<Cultivar>:BaseCntC	Count of 'C' bases
<Cultivar>:BaseCntG	Count of 'G' bases
<Cultivar>:BaseCntT	Count of 'T' bases
<Cultivar>:BaseCntN	Count of 'N' or indeterminate bases
	Next (maximum of 4000) cultivars follows