Understanding the *ngsconfig.json* File

MixtureAce Configuration Guide

NicheVision Forensics, LLC

Understanding the *ngsconfig.json* file

MixtureAce Configuration Guide

OVERVIEW

MixtureAce uses information contained within the ngsconfig.json file to configure the analysis of MPS data. This file is located in the following directory:

C:\Program Files\NicheVision Inc\MixtureAce\data\plugins\import\ngs_import

This configuration file must be named “ngsconfig.json”.

NGS KIT DEFINITION

The ngsconfig.json file works with a companion kit definition file. Kit definition files define the arrangement of loci in the MixtureAce bar chart display. There can be more than one kit definition in the ngsconfig.json file. They occur immediately prior to the DefaultStutter and DefaultSequencingError for each kit (Figure 1).

```
  "kits": [
    {
      "Name":"MainstAYd_DPMD_ASTR",
      "DefaultStutter": [
        {
          "PhrThreshold":  0.2,
          "NPlus": 1, "NMinus":
          2
        }
      ],
      "DefaultSequencingError": [
        {
          "IntensityThreshold": 0.2
        }
```

```
  ],

...

     "Name": "MainstAY_Ya",
      "DefaultStutter": [
        {
          "PhrThreshold":  0.2,
          "NPlus": 1, "NMinus":
          2
        }
      ],
      "DefaultSequencingError": [
        {
          "IntensityThreshold": 0.1
        }
      ],
```

Figure 1. Example of two kit definition names in the same ngsconfig.json file.

Kit definitions appearing in the ngsconfig.json file appear as options in the MixtureAce data input window (Figure 2). When a default kit is specified, that kit will be pre-selected in the data input window (Figure 2).

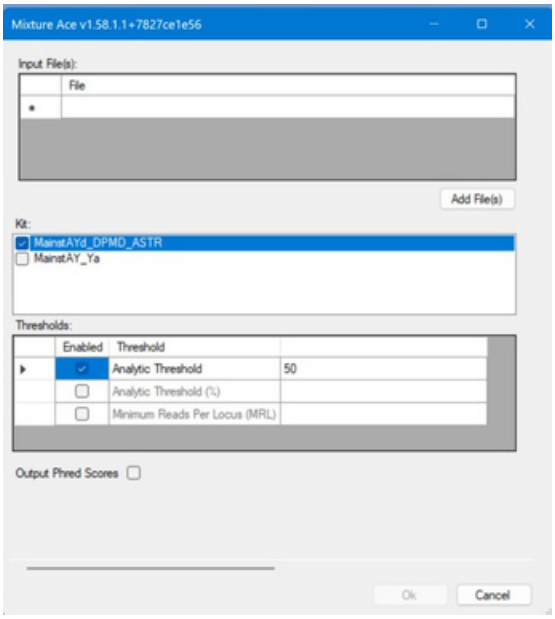

Figure 2. MixtureAce data input window.

The default kit can be specified near the top of the ngsconfig.jsonfile.

```
{
  "AbundanceThresholdInput": 50,
  "AbundanceThresholdOutput": 50,
  "OutputRawSequences": false,
  "DefaultKit": "MainstAYd_DPMD_ASTR",
  "OutputAbundantAlleleSequences": 0,
```

Figure 3. Header lines in the ngsconfig.json file.

The kit definition names used in the ngsconfig.json file must match the corresponding kit definition file name. Kit definition files are text files that can be edited by users, and are in the following directory:

C:\Program Files\NicheVision Inc\MixtureAce\data\plugins\kits

If the names do not match, or the kit definition file is missing, then MixtureAce will attempt to place all loci into a single channel in the bar chart display.

STUTTER THRESHOLDS

Once the correct ngsconfig.json file is in place with regard to the NGS kit, Mixture Ace will use the marker definitions contained within this file for analyzing data. While the markers in this file are pre-defined based on the MPS kit, the stutter thresholds are able to be changed by the user. For example, if internal validation results support changing the Back 1 stutter threshold at D3 from the default value of 0.15 (i.e. 15%) to 0.08 (i.e. 8%), then the user can:

- find the desired “Marker” in the ngsconfig.json file
- find the “Stutter” section below
- choose which type of stutter to update
  - N+1 (or Forward 1)
  - N-1 (or Back 1)
  - N-2 (or Back 2)
- change the orange text in “PhrThreshold” from 0.15 to 0.08 (for this example)
- save the ngsconfig.json file

```
"Marker": "D3S1358",
],
"Stutter": [
"PhrThreshold": 0.03,
"NPlus": 1,
"NMinus": 0
"PhrThreshold": 0.15,
"NPlus": 0,
"NMinus": 1
"PhrThreshold": 0.02,
"NPlus": 0,
"NMinus": 2
```

This change will now cause Mixture Ace to filter any peaks that are <8% of a parent peak as stutter, and label it as such; and any peaks that are >8% of a parent peak will be labeled as an allele.

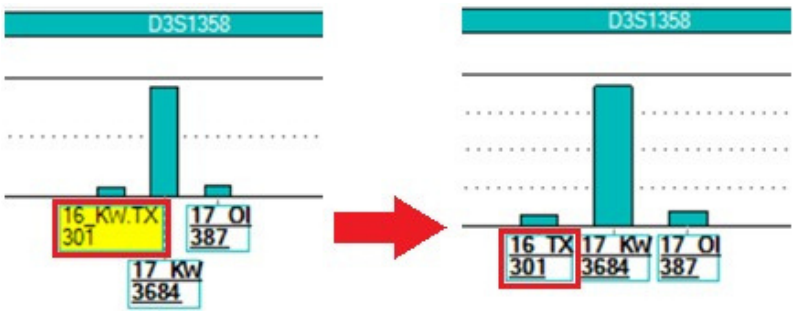

Figure 4. The stutter peak 16_KW.TX is 8.17% of the parent peak 17_KW. When the stutter threshold is 0.15 (15%), this peak is labeled as stutter with a yellow box (left). However, when the stutter threshold is changed to 0.08 (8%), this peak is labeled as an allele.

DEFAULT STUTTER

If no stutter threshold has been assigned to a marker, a default value will be applied to peaks within that marker. Default stutter thresholds are found near the top of the ngsconfig.json file, and users can also edit this value.

```
"DefaultStutter": [
  {
    "PhrThreshold": 0.2,
    "NPlus": 1,
    "NMinus": 2
  }
```

REPEAT MOTIFS

Each STR marker has its repeat motifs defined under the “RepeatMotifs” header, located below the “Marker” header. This displays repeat motifs for bracketing. For example, the repeat motifs selected for bracketing at D3S1358 are both tetramers: one being [TCTA] and the other being [TCTG].

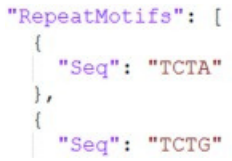

Figure 5. Specification of motifs for bracketing on a per-marker basis.

MixtureAce can bracket motifs of different lengths at the same marker. Order matters as the length of the first motif specified is the length used for calculating allele numbers.

SEQUENCE TRIMMING AND FORMATTING

Genomic positions at which sequencing reads are trimmed for analysis and display are specified as short sequence tags in the ngsconfig.json file.

EXAMINED REGION SEQUENCES (MIXTUREACE SEQ2)

Two sequence tags define the total genomic extent displayed for examination in MixtureAce output including the bar chart and the Excel and CSV reports. The total genomic extent for examination is termed SEQ2 in the MixtureAce Excel and CSV reports. Sequence tags labeled “PlusSequence” and “MinusSequence” define this examined region. These tags are used for alignment and typically are ≥15nt in length. PlusSequence and MinusSequence tags must be unique within the fastq file (but not necessarily unique within the genome). Trimming of the examined region occurs at the inside edge of these tags. Moving the tags or changing the length of the tags will change the reported allele number unless this is corrected (see Tech Note on allele numbers).

The PlusSequence and MinusSequence tags also control the read direction reported by MixtureAce. When the PlusSequence is in forward orientation (i.e. GenBank top strand, GenBank forward strand) and the MinusSequence is in reverse complement orientation, then the read direction reported by MixtureAce is GenBank forward direction. Reversing the tag orientations will cause MixtureAce to report the GenBank reverse direction (see Tech Note on examined region reporting).

STR MARKER SEQUENCES (SEQ1)

Two sequence tags define the STR region reported in bracketed format in MixtureAce bar charts, and Excel and CSV reports. The reported STR region is termed SEQ1 in MixtureAce. The region considered the STR is configurable by specifying the Flank5p and Flank3p sequence tags, both of which are in the reporting orientation selected by the PlusSequence and MinusSequence tags. Typically, this is GenBank forward orientation. The reported STR region is trimmed at the inside edge of these sequence tags. There is some variability in the genomic extent considered to be the forensic STR. For example, extents specified in recent ISFG guidance differ from how the markers have traditionally been defined (see Tech Note on sequence-based STR markers). In addition, users can choose to include additional sequence to the SEQ1 extent for interpretation purposes. For example, additional uncounted STR repeats are near some forensic STR markers. Including these repeats in the STR (SEQ1) definition can allow stutter artifacts of those uncounted STRs to be accounted for.

The Flank3p and Flank5p sequence tags are typically ≥15nt and must be inside or equal to the genomic positions used for the PlusSequence and MinusSequence tags. When they are equal to the PlusSequence and MinusSequence tags, then the reported genomic extent for SEQ1 is the same as for SEQ2.

ALLELE NUMBER REPORTING

MixtureAce reports allele numbers concordant with CE methods by calculating allele numbers using the sequence-based equivalent to allele sizing by CE. MixtureAce does not calculate allele numbers by bracketing.

In CE, alleles are sized by sizing standards designed to “subtract out” the flanks and other base pairs that do not contribute to the allele number. In similar fashion, MixtureAce subtracts out uncounted sequencing read nucleotides. These uncounted nucleotides include STR flanking sequence out to the outside edge of the PlusSequence and MinusSequence tags. In addition, some STR markers have uncounted nucleotides inside the STR motifs themselves. For example, D21S11 has 11 uncounted nucleotides inside the STR motifs. The uncounted nucleotide count is specified in the AlleleOffset and SizeStandard values (Figure 6). This division is not mandatory since the two numbers are summed prior to allele number calculation. After subtracting the uncounted nucleotides, the remaining nucleotides included in the SEQ2 sequence are divided by the repeat motif length. The repeat motif length is not explicitly specified in the ngsconfig.json file, but rather is determined by the length of the first repeat motif listed in the “RepeatMotifs” specification (Figure 5).

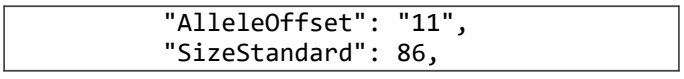

Figure 3. Specification of uncounted reads for allele number calculations.

Appendix A

Trim Protocols

Trim protocol A: Reports the STR region only.

Trim protocol B: Fixed Trim. Reports the sequence-based allele when the PCR primers are unknown. Some NGS kit vendors do not reveal their PCR primer extents.

Trim protocol C: Examined Region. Reports the sequence-based allele inside of the PCR primers. In other words, this is the sequence-based allele.

Trim protocol D: Intersection. Reports the overlap between the examined region of the NGS kit, and the regions represented in the allele frequency database.

Trim protocol E: NIST Minimum Range. Reports the region defined by NIST that includes the STR region plus the minimal adjacent flanking region.

For more detailed information about trim protocols, refer to this document: Technical Note Mixture Ace Trim Configurations.docx