**DM-tRNA-seq output for samples with reference genomes**

*Data alignment and quality:*

1. For faster alignment, first generate reference tRNA sequences for each organism:

(i) Get tRNA sequences in fasta format from Genomic tRNA database, <http://gtrnadb.ucsc.edu/>. For mitochondrial sequences, go to <http://mttrna.bioinf.uni-leipzig.de/mtDataOutput/>.

(ii) Remove intron sequences from non-mitochondrial tRNAs (in Genomic tRNA database these are highlighted).

(iii) Group all identical sequences together as one (in Genomic tRNA database all identical sequences have the exact same tRNAScan score).

(iv) Add 3’CCA to sequences if needed.

2. In raw sequencing files remove Illumina sequencing primers and adaptors (Trimmomatic, or other standard off the shelf software).

3. Align each read to the reference tRNA sequences, make sure that each read is assigned only once to a seed sequence (e.g. use JNP code). *Standard alignment is to allow one mismatch from the seed sequence (this may be changed to two mismatches depending on the project). In the read assignment, 3’A is not included in mismatch allowance.*

A particular issue for isodecoder alignment is that shorter reads can align to multiple sequences. In JNP code, this is dealt with as follows:

(i) Calculate the number of reads of similar sequences that align uniquely to an isodecoder seed.

(ii) Proportionate the ambiguous reads to those from (i).

4. Isoacceptor/Isodecoder identifier:

(i) For chromosome encoded tRNAs, use Genomic tRNA database. Example:

>Homo\_sapiens\_chr6.trna95-AlaAGC (58141949-58141877) Ala (AGC) 73 bp Sc: 42.26

GGGGAATTAGCTCAAGCGGTAGAGCGCTCCCTTAGCATGCGAGAGGTAGCGGGATCGACGCCCCCATTCTCTACCA

Isodecoder identifier = AlaAGC\_c6t95

Mitochondrial tRNA identifier = mtAlaTGC

**5. Data quality output: for each sample, calculate the total counts and % reads that align to all reference tRNA sequences.**

>80% Outstanding

50-80% Very good, excellent

20-50% Good

<20% Poor

*Abundance output for a dataset (multiple samples):*

**#1. Fraction of tRNA isodecoders.**

(i) Each read is assigned an isodecoder identifier (point #4 above).

(ii) Get read counts for each isodecoder, separately for genomic and mitochondrial tRNAs.

(iii) Get the fraction of each isodecoder for each sample.

(iv) A single output file for the whole data set with these columns: (a) isodecoder identifier (mitochondrial tRNAs are separated); (b) data from (iii) for sample 1; (c)….. data from (iii) for subsequent samples.

**#2. Calculate fraction of tRNA acceptors.**

(i) Add up all isodecoder reads with the same anticodon, separately for genomic and mitochondrial tRNAs.

(ii) Get the fraction of each acceptor for each sample.

(iii) A single output file for the whole data set with these columns: (a) amino acid-anticodon, e.g. AlaAGC (mitochondrial tRNAs are separated); (b) data from (ii) for sample 1; (c)….. data from (ii) for subsequent samples.

*Charging output for a dataset (multiple samples):*

**Do this only for isodecoders that fulfill these criteria from abundance output:**

**(1) > 0.1% of the most abundant isodecoder in each sample.**

**(2) > 50 read counts.**

**(3) Specific isodecoder that fulfills (1)+(2) in some samples, but not others is still included.**

**Charging level of tRNA isodecoders.**

(i) For each isodecoder get read counts that end with 3’CCA and 3’CC.

(ii) For each isodecoder get the charging level as [CCA]/([CC]+[CCA]).

(iii) A single output file for the whole data set with these columns: (a) isodecoder identifier (mitochondrial tRNAs are separated); (b) data from (i), add the counts for 3’CCA+3’CC; (c) data from (ii) for sample 1; (d)(e)……, repeat for subsequent samples.

*Modification output for a dataset (multiple samples):*

**Do this only for isodecoders that fulfill these criteria from abundance output:**

**(1) > 0.1% of the most abundant isodecoder in each sample.**

**(2) > 50 read counts.**

**(3) Specific isodecoder that fulfills (1)+(2) in some samples, but not others is still included.**

**Calculate mut1, stop, mut2, and MI for each position.**

A single output file for each isodecoder, e.g. AlaAGC\_c6t95 with these columns:

Sample 1:

(a) Position # (5’ nucleotide of the seed sequence is 1).

(b) Seed sequence.

(c) Total count for each position.

(d) A count.

(e) C count.

(f) G count.

(g) T count.

(h) Mutation fraction (mut1): this only considers the reads past the mutated position.

(i) Stop fraction.

(j) Mutation fraction (mut2). this includes the reads that stopped at the mutation position.

(k) Modification index (MI, sum of (i)+(j)).

(l) empty = spacer.

Repeat these same columns for samples 2, 3…. etc.

9/15/2019:

**Additional three outputs for mut1, mut2, and Modification index.**

These are listed in separate files for each sequence meeting the above requirements.