Hot RAD Manual

Lauren Assour, Nicholas LaRosa
University of Notre Dame

September 24, 2014

Hot RAD is a pipeline application for the analysis of RAD tag (also known as RAD-seq) data. It provides users with modular actions that can be preformed on raw Illumina data. All steps in the pipeline can be replaced with any outside applications desired, and all steps provide standardized output for ease of use. All actions can be run through the GUI or via the command line[1].

1 Actions

Hot RAD currently provides three basic actions to the user:

Trimmer:
The trimmer options takes raw Illumina reads and demultiplexes them.

Build Reference:
Build reference has two components - an alignment and a filtering. The aligner built into Hot RAD can be replaced with another *de novo* assembly tool, and the filter can then be run on the resulting contigs.

2 Trimmer

Trims barcodes and cut sites from sequences. Renames sequences to reflect individual from which read originated.

2.1 Input

Barcode File:
Tab separated in the format BARCODE MID. MID can be any unique identifier not containing underscores. Barcodes may be uppercase or lowercase, or a mix. Barcodes are not required to be all the same length. Example is provided in testfiles folder.

[1]Please see README for additional information on running Hot RAD through the command line

Sequence File(s) :

FASTQ format file(s), all identifier/sequence/quality lines should be on a single line. This is the file with tagged reads. Bases should be uppercase. Example is provided in testfiles folder.

Paired Sequence File(s) :

FASTQ format file, all identifier/sequence/quality lines should be on a single line. This is the secondary read file (untagged). Bases should be uppercase. Paired files should appear in same order as sequence files. Example is provided in testfiles folder.

2.2 Options

Cut Site:

The remaining end of the cut site, following the barcode. Default is AATTC (EcoR1).

Min Cut Site Match:

Minimum number of bases (from front to end) of the cut site that must be matched to count as the cut site being located. Default is length of cut site - 1 (Ex. for AATTC, min match = 4, AATT must be found within 4 bases of the barcode).

Individual Files:

Generate a separate FASTQ output file for each MID in each file. Note that built-in assembly tool only takes a single FASTQ file as input.

Same Length:

Cut all sequences down to the same provided length. This is primarily for prepping data for programs such as Stacks (`http://creskolab.uoregon.edu/stacks/`).

Minimum Length:

Minimum length of a sequence after trimming to retain read.

Maximum N Count:

Maximum number of Ns allowed in a sequence. Otherwise, the sequence is thrown out. Default is an infinite number of Ns (no sequences are removed).

N String Length:

Replace paired reads with N String Length number of Ns if they are at least in part a reverse complement of the forward read. Set to 0 to not replace.

2.3 Output

Barcode File.info:

Unless -q/–quiet flag set, information about the input files. This includes total number of reads in file, total number of reads retained (after removing unidentifiable barcodes/reads without the given cut site, or too many Ns), number of reads

removed due to too many Ns, no cut site, or bad barcodes, number of reads with barcodes corrected, average number of reads per barcode, average read length after trimming, the length distribution of sequences, and the counts per barcode.

Sequence File.trimmed.fq:
FASTQ format file with barcodes and cut sites trimmed. Naming convention is #_MID_FILENAME, where # is the number of sequences seen thus far in a given MID (first sequence is 0), MID is the MID linked to the located barcode, and FILENAME is the name of the ILLUMINA_FILE input (in case of MID overlap between files). Sequences with no identifiable barcode (after 1 bp correction) or no identifiable cut site will no be in this file. Quality ID lines (+ lines) have__MID appended unless the line was blank, in which case it is left blank. (Ex. input qual id line: +Thisismyqualidline, output qual id line: +Thisismyqualidline__MID).

Paired Sequence.trimmed.fq:
FASTQ format file. Naming convention is _2_#_MID_FILENAME, where # is the number of sequences seen thus far in a given MID (first sequence is 0), MID is the MID linked to the located barcode in the first read, and FILENAME is the name of the ILLUMINA_FILE input, not the PARIED_FILE input filename. Quality ID lines (+ lines) have __MID appended unless the line was blank, in which case it is left blank. (Ex. input qual id line: +Thisismyqualidline, output qual id line: +Thisismyqualidline__MID).

3 Build Reference

3.1 Alignment Options

The options for Align portion of the Build Reference section of Hot RAD are controlled by a configuration file that built by the GUI. Hot RAD's aligner saves all files to a folder specified by the user, with the end file saved as final.txt in the output folder. If Filter Only is selected no alignment will be performed, but the provided sequence file will be filtered using a script built for use with DNAStar's SeqMan NGen. Local will run the scripts directly as opposed to building a makeflow.

 # Segments
Number of segments to split the input file into, as Hot RAD can be run distributed using makeFlow.

Percent Identity
Percent identity of sequences to count as sufficiently aligned. Positive integer between 0 and 100.

FASTA/FASTQ Input
Specify either Fasta or Fastq sequence file. Ideally this input file has already been demultiplexed either using Trimmer or some other tool.

K Band

Banded alignment refers to a method of performing Smith-Waterman alignments where only the center band of the alignment matrix is filled; this limits the number of insertions/deletions. The k band determines the width of this band from the center.

Max Alignment Len

Number of bases to align. Sequences are deemed sufficiently aligned if their percentIdent is great enough after aligning maxAlign bases. Positive integer greater than or equal to 1. If maxAlign ¿ sequenceLength, the whole sequence is used.

Match Score

Value assigned to a match between bases in typical Smith-Waterman fashion. Positive integer (including 0).

Mismatch Score

Value assigned to a mismatch between bases in typical Smith-Waterman fashion. Negative integer (including 0).

Gap Score

Value assigned to a gap in a sequence in typical Smith-Waterman fashion. Negative integer (including 0).

 # Bases to Compare

Number of bases to compare base counts. This is directly related to one filtration option in Hot RAD, which counts the number of each base in the two sequences to be aligned. If the difference between their base counts is under a certain threshold, they are aligned. This is the number of bases to compare base counts with. Positive integer, greater than or equal to 0. If countLen ¿ sequenceLength, the whole sequence is used.

Max Difference

The maximum difference allowed between base counts, as related to the above value. Positive integer greater than or equal to 0 and less than or equal to countLen.

Minimum Group Size

Minimum group size to continue from distributed assembly. After the initial distributed step, sequence groupings that are less than Min Group Size will be removed and no longer considered. Positive integer greater than or equal to zero.

Test Area

Number of bases to test for initial alignment. This is directly related to a second filtration option in Hot RAD, which does base counting for the first Test Area bases similar to # Bases to Compare/Max Diff above. However, it directly compares bases to one another (base 0 in sequence 0 to base 0 in sequence 1, etc.), and allows for an offset (of size overhang, below) between the two sequences where they are

offset to each other by a maximum of overhang bases. If the two sequences have a minimum number of bases that match to one another, they will be aligned. Positive integer greater than or equal to zero.

Overhang
Maximum number of bases to shift over sequences in regards to each other for Test Area comparisons.

Min Initial Matching
Minimum number of bases to match in test area. Positive integer greater than or equal to zero. Maximum overhang length (Section X.X of manual)

3.2 Alignment Output

final.txt
Sequences split into similar clusters, where the first sequence in each group is deemed the most representative using center star tree alignment. Each group is separated by a blank line.

3.3 Filtering Options

The filtering section filters out contigs created by the aligner/assembly tool that are undesireable due to their length or coverage, and can create a single, long contig as opposed to multiple contigs to use as a reference.

Minimum Coverage
Minimum number of reads used to create a contig.

Minimum Length
Minimum length of a contig.

Maximum Length
Maximum length of a contig.

Single Contig
Output as a single, long contig padded by Ns.

Padding Ns
If Single Contig selected, the number of Ns to pad between contigs.

3.4 Filtering Output

Contig File
A file containing contigs that passed the filtering options. If Hot RAD's aligner was used this will be the most representative sequence of a given cluster. By default they are all separate contigs, but if Single Contig is used then one contig named 'FakeGenome' is output.