RASER: Reads Aligner for SNPs and Editing sites of RNA (version 0.51)

Manual

July 02, 2015

Index

1. System requirement and how to download RASER source code

Currently RASER supports Linux 64bit (Ubuntu, Red Hat, SUSE and others) system.

Download raser_v0.51.tar.gz from https://github.com/JaegyoonAhn/RASER/releases/tag/v0.5.1 or https://www.ibp.ucla.edu/research/xiao/RASER.html

2. Installation

'g++' and 'make' are required to compile RASER.

(1) Uncompress raser_v0.51.tar.gz

(2) Go to uncompressed raser directory, and compile it by running

./make

(3) If successfully compiled, you can find two binary files, 'raser' and 'raserIdx'.

Note: If the size of your genome is above 4.2 Gb (=2^{32}), run '**./make MODE=-DS64**' instead. Please refer to http://en.wikipedia.org/wiki/Genome#Genome_size to get the size of your genome.

3. Making index files

Before you running RASER, you should generate an index file using 'raserIdx', or download pre-built index file.

1) Downloading pre-built index files

You can download pre-built index files for human from https://www.ibp.ucla.edu/research/xiao/RASER.html

(1) hg19.ridx – an index file for hg19 only
(2) hg19_ensembl.ridx – an index file for both hg19 and ensembl transcriptome

Note: if you compiled raser using -DS64 option, do not use above index files (you should generate a new index file using raserIdx compailed with -DS64 option)

2) Generating an index file

(1) Using only genome sequences

./raserIdx -r *GENOME_FASTA_FILE* **[-x** *YOUR_INDEX_NAME***] [***optional parameters***]**

GENOME_FASTA_FILE : Genome sequence in FASTA format

YOUR_INDEX_NAME : Name of the index file. If not set, the name is set to default = 'a.ridx'.

optional parameters : Refer to following 3-2)-(4) optional parameters

(2) Using only transcriptome sequences

./raserIdx -r *mRNA_FASTA_FILE* **-l** *ANNOTATION_FILE* **[-p** *mRNA_name_prefix***] [-x** *YOUR_INDEX_NAME***] [***optional parameters***]**

mRNA_FASTA_FILE : transcriptome sequences in FASTA format. Its format is like below:

```
>ENST00000237247 range=chr1:67000042-67208778 5'pad=0 3'pad=0 strand=+ repeatMasking=none
ATGATGGAAGGATTGAAAAAACGTACAAGGAAGGCCTTTGGAATACGGAA
GAAAGAAAAGGACACTGATTCTACAGGTTCACCAGATAGAGATGGAATTC
AGGGGAAAAAAAAGACCCAGAAGACTCAGCTTCTCCTCACCTCTTGCTTC
…
```

ANNOTATION_FILE : annotation table which is downloadable from UCSC genome browser. Its format is like below:

```
#bin  name  chrom  strand  txStart txEnd  cdsStart  cdsEnd  exonCount  exonStarts  exonEnds  score  name2  cdsStartStat  cdsEndStat  exonFrames
0  ENST00000237247 chr1  +  66999065  67210057  67000041  67208778  27  66999065,66999928,67091529,67098752,67099762,
67105459,67108492,67109226,67126195,67133212,67136677,67137626,67138963,67142686,67145360,67147551,67149789,67154830,67155872,67161116,67184976,67194946,67199430,67
205017,67206340,67206954,67208755,  6999090,67000051,67091593,67098777,67099846,67105516,67108547,67109402,67126207,67133224,67136702,67137678,67139049,67142779,
67145435,67148052,67149870,67154958,67155999,67161176,67185088,67195102,67199563,67205220,67206405,67207119,67210057,  0  ENSG00000118473 cmpl  cmpl
-1,0,1,2,0,0,0,1,0,0,0,1,2,1,1,1,1,1,0,1,1,2,2,0,2,1,1,
0  ENST00000371039 chr1  +  66999274  67210768  67000041  67208778  22  66999274,66999928,67091529,67098752,67105459,67108492,
67109226,67136677,67137626,67138963,67142686,67145360,67154830,67155872,67160121,67184976,67194946,  67199430,67205017,67206340,67206954,67208755,
66999355,67000051,67091593,67098777,67105516,67108547,67109402,67136702,67137678,67139049,67142779,67145435,
67154958,67155999,67160187,67185088,67195102,67199563,67205220,67206405,67207119,67210768,  0  ENSG00000118473 cmpl  cmpl  -  1,0,1,2,0,0,1,0,1,2,1,1,1,0,1,1,2,2,0,2,1,1,
…
```

mRNA_name_prefix : if mRNA names in *ANNOTATION_FILE* and *mRNA_FASTA_FILE* are different, *mRNA_name_prefix* should be set. For example, mRNA names in *ANNOTATION_FILE* and *mRNA_FASTA_FILE* are 'ENSMUST00000086465' and 'mm10_ensGene_ENSMUST00000086465' respectively, *mRNA_name_prefix* is 'ENSMUST'.

YOUR_INDEX_NAME : Name of the index file. If not set, the name is set to default = 'a.ridx'.

optional parameters : Refer to following 3-2)-(4) optional parameters

(3) Using both genome and transcriptome sequences

./raserIdx -r *GENOME_FASTA_FILE mRNA_FASTA_FILE* **-l** *ANNOTATION_FILE* **[-p** *mRNA_name_prefix***] [-x** *YOUR_INDEX_NAME***] [***optional parameters***]**

GENOME_FASTA_FILE : Genome sequence in FASTA format

mRNA_FASTA_FILE : transcriptome sequences in FASTA format (its format is explained in 2.2)

ANNOTATION_FILE : annotation table downloaded from UCSC genome browser (its format is explained in 2.2)

mRNA_name_prefix : explained in 2.2

YOUR_INDEX_NAME : Name of the index file. If not set, the name is set to default = 'a.ridx'.

optional parameters : Refer to following 3-2)-(4) optional parameters

(4) Optional parameters

-k <int> : RASER indexes reference sequence of length (s + 16) at every k steps. Smaller k leads to larger index but may enable more complete mapping. Default value is 4, but you can try smaller k if too many reads are unmapped.

-s <int> : Minimum size of sequence indexed. For most eukaryote genomes, s = 8 (default value) is enough. For smaller

genomes, smaller value should be used. Recommended value of s is as follows:

Genome size (bps)	s
5,000,000,000 ~	9
800,000,000 ~ 5,000,000,000	8 (default)
100,000,000 ~ 800,000,000	7
15,000,000 ~ 100,000,000	6
2,500,000 ~ 15,000,000	5
400,000 ~ 2,500,000	4
~ 400,000	3

-h : prints out help page

-v : prints out version of raserIdx

4. Aligning

1) Basic usage

./raser [-x *index***] {-i** *read* **| -i1** *read_end1* **-i2** *read_end2***} [-o** *result***] [***optional parameters***]**

Parameter name	Description	Default value
index	(optional) Name of index file	a.ridx
read	Name of read file in FASTQ or FASTA format (single end)	N/A
read_end1/2	Read pair end1/2 file in FASTQ or FASTA format (paired end)	N/A
result	(optional) Result file in SAM format	a.sam
optional parameters	Refer to following 4-2) optional parameters	N/A

2) Optional parameters (options in bolds are important ones)

Parameter name	Description	Default value
-m <float>	maximum ratio of mismatches (including indels)	0.05
-p <int>	Reports <int> number of good alignments. Set 0 to get all good alignments. Set 1 for best alignments	0
-l <int>	maximum length of insertion (usually equal to maximum intron length)	200000
-d <float>	Set this to apply double filtering. Double filtering is mapping scheme that a read (pair) should be uniquely mapped with less than -m <float> threshold, and not mapped to anywhere with more than -d <float> threshold (refer to Bahn et al., 2012). It aims to predict SNP or RNA editing sites accurately, by minimizing false positives. Recommended value is 0.09, and should be in (m, 0.3). If -d is set, RASER reports best (unique) result (ignoring -p <int>).	Not set
-b <float>	**(highly recommended to set for SNP or editing site detection)** Set this to finds the obviously best mapping of a read (pair), whose mismatch score is less than -m <float>, and also less than mismatch scores of all other mappings by -b <float>. It aims to maximize the mapping rate, as well as to minimize the false positive rate. **Recommended value is 0.03**, and should be in (0, 0.2) (Refer to Ahn et al., 2015). If -b is set, RASER reports best (unique) result (ignoring -p <int>).	Not set

-g	Don't map spliced reads	Not set
--sanger	If read quality scores are based on Illumina v1.7 or less, this option will convert quality score to Sanger format (=Illumina v1.8 format)	Not set
-s <float>	allows <float> * read_length of soft-clipping. We recommend to use smaller s for shorter read (e.g. 0.1 for 36 bp reads)	0.25
-t <int>	Number of thread to use. Set 1 for no multi-processing	8
--idop	Indel open penalty	3
--idep	Indel extend penalty	1
--nxm	Don't add XM/XJ fields to output SAM files. Refer to 5. Output for details of XM/XJ fields	Not set
--nxt	Don't add XT field to output SAM files. Refer to 5. Output for details of XT field	Not set
-h	prints out help page	N/A
-v	prints out version of RASER	N/A

5. Output

If successfully finished, RASER will generate SAM and log files.
SAM file includes following extra fields (if options '--nxm' and/or '--nxt' are not set).

Column	Tag	Format	Description
16	XJ:Z:	chr,{*start*:*end*} : *start* and *end* are 1-base loci. If a read is spliced, there can be multiple '*start*:*end*' separated by comma	For easier parsing of read mapping positions within the reference
17	XM:Z:	{*ref_loc*-*read_loc*-*ABC*} : *ref_loc* is a reference locus of a mismatch (1-base). *read_loc* is a read position of a mismatch (1-base). *A* is a reference base of a mismatch. *B* is a read base of a mismatch. *C* is a read quality of a mismatch.	For easier parsing of mismatch information
18	XT:Z:	Transcript names separated by comma	Includes transcript names to which a read is mapped, if a transcriptome was indexed

6. Contact

Bug reporting, questions or any suggestions are highly appreciated.

Jaegyoon (Jay) Ahn (jgahn AT ucla.edu)
Xinshu (Grace) Xiao (gxxiao AT ucla.edu)

7. Reference

J. Bahn, J. Lee, G. Li et al., Accurate identification of A-to-I RNA editing in human by transcriptome sequencing, Genome Research, Vol.22, 2012

J. Ahn and X. Xiao, RASER: Reads aligner for SNPs and editing sites of RNA, under review, 2015