JACUSA2 manual

Michael Piechotta
michael.piechotta@gmail.com

29th, September, 2019

Contents

Todo list

1 Introduction

JAVA framework for accurate SNV assessment (JACUSA2) is a one-stop solution to detect single nucleotide variants (SNVs) and reverse transcriptase induced arrest events in Next-generation sequencing (NGS) samples.

replace SNV with something more general

michael

make a text out of this

- `https://github.com/dieterich-lab/JACUSA2/`JACUSA2 direct successor of `https://github.com/dieterich-lab/JACUSA/`JACUSA1 — JACUSA1 is hereby deprecated and won't be continued
- all methods from JACUSA1 are available in JACUSA2
- reverse transcriptase arrest events can be identified
- explain briefly rt-arrest and lrt-arrest
- stratification by read base changes
- number of deletion per site can be calculated
- command line changes: only one dash options, two dash options have been removed
- new architecture -> 3x faster than JACUSA1
- new filter(s): exclude/mark SNP/variants/regions
- some filters habe been move to JACUSA2helper
- htsjdk to parse BAM files

JACUSA2 employs a window-based approach to traverse provided BAM files featuring highly parallel processing and utilizing the new `https://github.com/samtools/htsjdk`htsjdk framework.

1.1 Variant calling

Robust identification of variants has proven to be a daunting task due to artefacts specific for NGS-data and employed mapping strategies. We implement various feature filters that reduce the number of false positives.

JACUSA2 has been extensively evaluated and optimized to identify RNA editing sites in RNA-DNA and RNA-RNA sequencing samples. JACUSA2 requires an operating JAVA environment and uses sorted and indexed BAM files as input.

1.2 Reverse transcriptase arrest events

michael

makr nice text

Reverse transcriptase arrest events can be induced during library preparation. They are identified by shorter than expected read length due to premature termination during first strand synthesis. Per site a vector of read through and read arrest can be calculated and compared between conditions. Read through and read arrest events are modelled by the Beta-Binomial distribution.

2 Download

The latest version of JACUSA2 can be obtained from `https://github.com/dieterich-lab/JACUSA2/`JACUSA2. Got to releases and pick the lastest release, currently: `https://github.com/dieterich-lab/JACUSA2/releases/down` `2.0.0-RC2/JACUSA_v2.0.0-RC3.jar`JACUSA2 2.0.0-RC3

michael
How to automate this?

2.1 Installation and requirements

JACUSA2 does not need any configuration but needs a correctly configured Java environment. We developed and tested JACUSA2 with Java v1.8. If you encounter any Java related problems please consider to change to Java v1.8.

2.2 Migrating from JACUSA1 to JACUSA2

michael
make a text out of this

- command line parameters can ONLY be provided by one dash “-” options, e.g.: “-c 10”
- all two dash options “–option [. . .]” have been removed and are NOT available anymore
- “–filterNH” and “–filterNM” been replaced in JACUSA2 with “-filterNH” and “-filterNM”
- the CLI format to provide library type for each condition has changed: JACUSA1: “-P Lib1,Lib2”, JACUSA2: “-P1 Lib1 -P2 Lib2”.
- JACUSA2 adds a “##” prefixed header line to the default output file format that contains command line options and used JACUSA2 version.

michael
I might support old format

2.3 Sample *in silico* data

2.3.1 Variant calling

You can choose between different setups and species where the later greatly influences the data size and running time to detect variants. The gDNA VS cDNA represents the typical data setup that is encountered in detection of RNA editing sites via comparing genomic and transcriptomic sequencing reads. In this setup, variants have been only imputed to the cDNA BAM file. The cDNA VS cDNA data setup can be interpreted as representing allele specific expression of single variants or differential RNA editing. In this setup, variants with pairwise different base frequency have been imputed into both cDNA BAM files. Additionally, to make the identification of variants more challenging SNPs with pairwise similar base frequencies have been included to both BAM files. This sites should not be identified as true positive sites.

gDNA data has been simulate with art[1] and cDNA reads have been simulated with flux[2]. Read simulations have been restricted to the corresponding first chromosome of the respective species. Sample data is available for *C. elegans* ce10 and *Homo sapien* hg19. Each archive consists of:

[1] `http://www.niehs.nih.gov/research/resources/software/biostatistics/art/`art

[2] `http://sammeth.net/confluence/display/SIM/Home`flux simulator

gDNA.bam, cDNA.bam BAM files: gDNA.bam and cDNA OR cDNA_1.bam and cDNA_2.bam

snps.txt Only available for cDNA VS cDNA. Coordinates of imputed SNPs. In both BAM files matching SNPs have the same target frequency but different effective or sampled frequency. The shape parameter determines how much the sampled frequency will deviate from the target frequency in each BAM file. The suffixes: _cdna_1 and _cdna_2 correspond to the respective BAM file

variants.txt Coordinates of imputed variants and their target and sample frequencies

Available sample data:

michael

we have to move the data to dieterichlab

-
- `https://data.dieterichlab.org/s/hg19_chr1_gDNA_VS_cDNA`hg19_chr1_gDNA_VS_cDNA

2.4 Reverse transcriptase arrest event

michael

What data should we provide here?

3 Input

All JACUSA2 methods require sorted and indexed `https://samtools.github.io/hts-specs/SAMv1.pdf`BAM files. BAM is a standardized file format for efficient storage of alignments. Furthermore, JACUSA2 requires that the reference sequence is available either through the “MD” `https://samtools.github.io/hts-specs/SAMtags.pdf`tag in BAM files or by providing the reference sequence in indexed FASTA format with the command line option “-R <reference.fasta>”. The “MD” contains mismatch information that allow to perform variant calling without providing.

Check the manuals of `http://samtools.sourceforge.net/`SAMtools/BCFtools and/or `http://broadinstitute.github.io/picard/`picard tools for how to use the respective tool to convert your alignment files to valid JACUSA2 input BAM.

3.1 Processing BAM files

In the following, commands for SAMtools are presented.

To sort and index your raw BAM files perform the following sequence of commands:

SAM → *BAM* `samtools view -Sb mapping.sam > mapping.bam`

sort BAM `samtools sort mapping.bam mapping.sorted`

index BAM `samtools index mapping.sorted.bam`

Check your BAM file for the “MD” `https://samtools.github.io/hts-specs/SAMtags.pdf`tag if you want to provide reference sequence information via this tag. When your BAM files do not have the “MD” tag set correctly use SAMtools:

```
samtools calmd mapping.sorted.bam reference.fasta > mapping.sorted.MD.bam
```

Table 1: Example of BED-like traverse file

contig	start	end
1	1000	1100
2	10000	10000

3.1.1 Remove duplicates for variant calling

It is a recommended pre-processing step to remove duplicate reads when identifying variants. Duplicated reads occur mostly due to PCR-artefacts. They are likely to harbour false variants and most statistical test require that reads are sampled independently. In the following, commands for picard tools are presented:

michael

rephrase paragraph: for call it is recommended, for rt/lrt-arrest absolutely not!

```
java -jar MarkDuplicates.jar \
 I=mapping.sorted.bam O=dedup_mapping.sorted.bam \
 M=duplication.info
```

Invoke JACUSA2 with the additional command line option "-F 1024" to filter reads that have been marked as duplicates.

3.1.2 Library type and strand information

JACUSA2 supports stranded paired end and single ends reads. With the command line parameter "-P <LIBRARY-TYPE> | -P1 <LIBRARY-TYPE> -P2 <LIBRARY-TYPE>" the user can choose the underlying library type:

RF-FIRSTSTRAND STRANDED library - first strand sequenced,

FR-SECONDSTRAND STRANDED library - second strand sequenced, and

UNSTRANDED UNSTRANDED library.

The UNSTRANDED library type is not available for rt/lrt-arrest because an arrsest site can not unambiguously be defined for this library type.

3.2 Traverse BED-like file

Identification of interesting sites can be restricted to specific regions of the genome or transcriptome. Provide a minimalistic BED-like file to limit the search to this region(s) or site(s). Remaining region(s) of the BAM files will not be considered.

In the following traverse file, the search is confined to a 100nt region on contig 1 starting at 1,000 and a single site on contig 2 at coordinates 10,000:

HINT: Many individual sites will slow down JACUSA2. If possible, try to merge nearby sites into contiguous regions and extract specific sites from JACUSA2 output with `http://bedtools.readthedocs.org/en/latest/bedtools` "intersect":

merge sites `bedtools merge -d 500 singular_sites.bed > \`
`contigous_regions.bed`

run JACUSA2 `java -jar JACUSA2.jar call-2 -b contigous_regions.bed -r JACUSA2.out mapping_1.sorted.bam mapping_2.sorted.bam`

extract sites `bedtools intersect -wa -a JACUSA2.out -b singular_sites.bed`

3.3 Output

JACUSA2 writes its output to a user defined file. When using multiple threads, JACUSA2 will create a temporary file for each allocated thread in the temporary directory that is provided by the operating system. Chosen command line parameters and current genomic position are printed to the command prompt and serve as a status guard. Furthermore, depending on the provided command line parameters, JACUSA2 will generate a file with sites that have been identified as potential artefacts when "-s" is provided. Currently, JACUSA2 supports the following output formats, controlled by "-f":

michael
user should be able to change this

michael
add progress bar

- Default (JACUSA2 output — varies between JACUSA2 methods)
- Variant Call Format (VCF)[3]

The default output format is based on BED6[4] with additional JACUSA2 methods specific columns. The actual number of columns depends on the JACUSA2 method and the number of provided BAM files.

Table 2: JACUSA2 default output format — core elements

Column:	1	2	3	4	5	6	. . .	N-1	N
	1	100	101	variant	8.07 . . .	-	JACUSA2 method specific	*	*
			. . .				. . .		

(1, 2, 3) contig + start + end 0-based, genomic coordinates of potential variant site

(4) name Currently, constant string: "variant". This dummy field is to ensure BED6 compatibility

(5) score Test-statistic $z \in \mathbb{R}$ that indicates the likelihood that this is a true variant. Higher number indicates a higher likelihood for a variant

(6) strand Possible values are: ".", "+", and "-" which correspond to "unstranded", "positive strand", and "negative strand" respectively. If strand is != ".", then the following base columns will be indicating base counts according to the strand - inverted base count if on the "negative strand"

(7-N-2) method specific The number of base columns depends on the JACUSA2 method — check method specific explanation.

(N-1) info Additional info for this specific site. Currently, details about the parameter estimation of the underlying distribution can be shown, and additional method specific data. If nothing provided, the empty field is equal to "*"

[3] `http://samtools.github.io/hts-specs/VCFv4.1.pdf]`VCF file format
[4] `http://genome.ucsc.edu/FAQ/FAQformat.html#format1`BED file format

(N) filter_info Relevant, if feature filter(s) X have been provided with “-a X” on the command line. The column will contain a comma-separated list of feature filters that predict this site to be a potential artefact. Possible values depend on the utilized JACUSA2-method:

4 Feature/Artefact filter

michael

add filter figure from paper

Value	Description of potential artefact
D	Variant call in the vicinity of Read Start/End, Intron, and/or INDEL position
B	Variant call in the vicinity of Read Start/End
I	Variant call in the vicinity of INDEL position
S	Variant call in the vicinity of Splice Site
Y	Variant call in the vicinity of homopolymer
M	Max allowed alleles exceeded
H	“Control” sample contains non-homozygous pileup

5 Variant detection

5.1 Identification of RNA editing sites

In order to identify RNA editing sites by comparing gDNA and *stranded* RNA-Seq (single or paired end) use:

first strand sequenced “-P1 UNSTRANDED -P2 RF-FIRSTSTRAND”

second strand sequenced “-P2 UNSTRANDED -P2 FR-SECONDSTRAND”

. When your RNA-Seq is unstranded use: “-P1 UNSTRANDED -P2 UNSTRANDED” and infer the correct orientation from annnotation.

Use the following command line to identify RNA-DNA differences in BAM files that might give rise to RNA editing sites:

```
java -jar call-2 -r JACUSA.out -s -a H:1 gDNA.bam cDNA.bam
```

Option “-a H:1” ensures that potential polymorphisms in gDNA will be eliminated as artefacts. The number $x \in \{1, 2\}$ determines which sample has to be homomorph - in this case: gDNA.bam.

Use the following command line to identify RNA-DNA differences:

```
java -jar call-2 -r JACUSA2.out -s cDNA1.bam cDNA2.bam
```

WARNING: If you want to identify RNA-RNA differences make sure NOT to use the filter “-a H:x”! Otherwise, potential valid variants will be filtered out.

6 Reverse transcriptase arrest events

michael

add text

7 Usage

Calling JACUSA2 without any arguments will print the available tools which currently are:

```
java -jar JACUSA2.jar
  METHOD        DESCRIPTION
  call-1        Call variants - 1 condition
  call-2        Call variants - 2 conditions
  pileup        SAMtools like mpileup (2 conditions)
  rt-arrest     Reverse Transcription Arrest - 2 conditions
  lrt-arrest    Linkage arrest to base substitution - 2 conditions
Version:  [...]
Libraries:
```

7.1 call-1

Single sample (call-1) allows to call variants against a reference. Internally, an *in silico* sample is created from information that is provided by the "MD" field in BAM files.

The number of base columns depends on the number of BAM files. In basesIJ: I corresponds to sample and J to the respective replicate. Numbers indicate the base count of the following base vector: (A, C, G, T)

Sites that have a $>$ alleles are considered candidate variant sites and for this sites a test-score will be computed.

7.2 call-2

michael

nicely combine call-1, call2

7.3 pileup

See "Call variant - two samples" for details.

7.4 rt-arrest - 2 conditions

In this method base call counts of arrest and read through reads are modelled by a Beta-Binomial distribution and differences between conditions are to be identified by means of a likelihood-ratio test. Subsequent approximiation with χ^2 distribution to compute a pvalue.

Sites are considered candidate arrest sites, if in all BAM files there is at least one read through AND one read arrest event. Furthermore, coverage filter and minBASQ of Base Call apply that will affect the output.

7.5 lrt-arrest - 2 conditions

michael

make fluent text

lrt-arrest allows to link pileups to their arrest position. Output consists of read arrest and read through counts and a references to the associated arrest positions. There are cases, where currently an arrest position cannot be defined, e.g.: non properly paired reads. Output consits of at least one line. Each separate arrest position adds an additional row is The first row contains the unstratified data or total, the "arrest_pos" column is set to "*". Any following sites with identical coordinates (contig, start, end, strand) will have a different arrest position reference in the "arrest_pos" column.

This method supports partial artefact filtering. Currently, filters only apply to the unstratified data — sites with "*" in in "arrest_pos". Furthermore, coverage filter and minBASQ of Base Call apply that will affect the output.

8 Description of command line options

8.1 Input / Output

8.1.1 Input BAM files

8.1.2 Output file

-r RESULT-FILE	results are written to RESULT-FILE	call-1 call-2 pileup rt-arrest lrt-arrest

8.2 Filtering artefacts

8.2.1 Configure feature filter

-a FEATURE-FILTER	[...] Use -h to see extended help	call-1 call-2 pileup rt-arrest lrt-arrest

8.2.2 Output artefacts to separate file

-s	Store feature-filtered results in another file (= RESULT-FILE.filtered)	call-1 call-2 pileup rt-arrest lrt-arrest

8.3 Input BED file

-b BED	BED file to scan for variants	call-1 call-2 pileup rt-arrest lrt-arrest

8.4 Reference fasta file

-R REF-FASTA	use reference FASTA file (must be indexed)	call-1 call-2 pileup rt-arrest lrt-arrest

8.5 Library type

-P LIB-TYPE	multirow3 multirow2 lrt	call-1 call-2 pileup rt-arrest lrt-arrest

8.6 Read base changes

-B READ-SUB	Count non-reference base substitution per read and stratify. Requires stranded library type. (Format for T to C mismatch: T2C; use ',' to separate substitutions) Default: none	call-1 call-2 pileup rt-arrest

8.7 Show deletion score

-D	Show deletion score	call-1 call-2 pileup rt-arrest

8.8 General filtering

8.8.1 Filter by mapping quality

-m1 MIN-MAPQ1	filter positions with MAPQ < MIN-MAPQ1 for condition 1 default: 20	call-2 pileup rt-arrest lrt-arrest

8.8.2 Filter by base call quality

-q1 MIN-BASQ1	filter positions with base quality < MIN-BASQ1 for condition 1 default: 20	call-2 pileup rt-arrest lrt-arrest

8.8.3 Filter by minimal coverage

-c1 MIN-COVERAGE1	filter positions with coverage < MIN-COVERAGE1 for condition 1 default: 5	call-2 pileup rt-arrest lrt-arrest

8.8.4 Limit maximal depth

-d1 MAX-DEPTH1	max read depth for condition 1 default: -1	call-2 pileup lrt-arrest

8.9 Specific filtering

8.9.1 Filter by flag(s)

8.9.2 Retain by flag(s)

-F FLAG	filter reads with flags FLAG default: 0	call-1 call-2
	filter reads with flags FLAG for all conditions default: 0	pileup rt-arrest
	filter reads with flags FLAG for all conditions default: 0	lrt-arrest

8.9.3 Filter by number of hits

-filterNH_1 NH	Max NH-VALUE for SAM tag NH for condition 1	call-2 pileup rt-arrest lrt-arrest

8.9.4 Filter by number of mismatches

-filterNM_1 NM	Max NM-VALUE for SAM tag NM for condition 1	call-2 pileup rt-arrest lrt-arrest

8.10 Thread related

8.10.1 Number of parallel threads

-p THREADS	use # THREADS default: 1	call-1 call-2 pileup rt-arrest lrt-arrest

8.10.2 Actual thread window size

-w WINDOW	size of the window used for caching. Make sure this is greater than the read size default: 10000	call-1 call-2 pileup
	size of the window used for caching. Make sure this is greater than the read size default: 10000	rt-arrest
	size of the window used for caching. Make sure this is greater than the read size default: 5000	lrt-arrest

8.10.3 Reserved thread window size

-W THREAD-WINDOW	size of the window used per thread default: 100000	call-1 call-2 pileup rt-arrest lrt-arrest

8.11 Test-statistic options

-T THRESHOLD	Filter positions based on test-statistic THRESHOLD default: DO NOT FILTER	call-1 call-2 rt-arrest lrt-arrest
-u MODE	[...] Use -h to see extended help	call-1 call-2 rt-arrest lrt-arrest

8.12 Filtering by Test-statistic threshold

8.13 Misc

-h	Print extended usage information	call-1 call-2 pileup rt-arrest lrt-arrest
-x	turn on Debug modus	call-1 call-2 pileup rt-arrest lrt-arrest

9 Used libraries

Libray	**Version**	**Source**
htsjdk	2.12.0	`https://github.com/samtools/htsjdk`
Apache commons-cli	1.4	`https://commons.apache.org/proper/commons-cli`
Apache commons-math3	3.6.1	`http://commons.apache.org/proper/commons-math`

References