pairsamtools Documentation

Release 0.0.1

Mirny Lab

Aug 20, 2022

CONTENTS

1 Quickstart **3**
- 1.1 Installation . . . 3

2 Parsing alignments into pairs **5**
- 2.1 Overview . . . 5
- 2.2 Terminology . . . 5
- 2.3 Unmapped/multimapped reads . . . 6
- 2.4 Multiple ligations (walks) . . . 6
- 2.5 Interpreting gaps between alignments . . . 6
- 2.6 Rescuing single ligations . . . 7
- 2.7 Pair flipping . . . 7
- 2.8 Other reporting options . . . 7

3 .pairsam format **9**
- 3.1 specification . . . 9
- 3.2 pair types . . . 10

4 Command-line tools **11**
- 4.1 pairsamtools . . . 11
  - 4.1.1 dedup . . . 11
  - 4.1.2 filterbycov . . . 13
  - 4.1.3 markasdup . . . 15
  - 4.1.4 merge . . . 16
  - 4.1.5 parse . . . 17
  - 4.1.6 phase . . . 19
  - 4.1.7 restrict . . . 20
  - 4.1.8 select . . . 21
  - 4.1.9 sort . . . 23
  - 4.1.10 split . . . 24
  - 4.1.11 stats . . . 25

5 Indices and tables **27**

Index **29**

pairsamtools is a set of simple and fast command-line tools to process sequencing data from Hi-C experiments.

pairsamtools operate on sequence alignments and perform the following operations:

- detect and classify ligation sites (a.k.a. *Hi-C pairs*) produced in Hi-C experiments
- sort Hi-C pairs for downstream analyses
- detect, tag and remove PCR/optical duplicates
- generate extensive statistics of Hi-C datasets
- select Hi-C pairs given flexibly defined criteria
- restore and tag .sam files for selected subsets of Hi-C pairs

pairsamtools produce .pairs files compliant with the 4DN standards.

Contents:

CHAPTER

ONE

QUICKSTART

1.1 Installation

Requirements:

- python 3.4 and higher
- unix sort
- bgzip
- pbgzip (optional)
- samtools >= 1.4
- Python packages `Cython`, `numpy`, `click`, `nose`

We highly recommend using the conda package manager to install scientific packages and their dependencies. To get `conda`, you can download either the full Anaconda Python distribution which comes with lots of data science software or the minimal Miniconda distribution which is just the standalone package manager plus Python. In the latter case, you can install the packages as follows:

```
$ conda install samtools sort bgzip pbgzip Cython numpy pip
$ pip install click nose
```

Install the latest version of pairsamtools using pip.

```
$ pip install git+https://github.com/mirnylab/pairsamtools
```

CHAPTER

TWO

PARSING ALIGNMENTS INTO PAIRS

2.1 Overview

Hi-C experiments aim to measure the frequencies of contacts between all pairs of loci in the genome. In these experiments, the spacial structure of chromosomes if first fixed with formaldehyde crosslinks, after which DNA is partially digested with restriction enzymes and then re-ligated back. Then, DNA is shredded into smaller pieces, released from nucleus, sequenced and aligned to the reference genome. The resulting sequence alignments reveal if DNA molecules were formed through ligations between DNA from different locations in the genome. These ligation events imply that ligated loci were close to each other when the ligation enzyme was active, i.e. they formed "a contact".

`pairsamtools parse` detects ligation events in the aligned sequences of DNA molecules formed in Hi-C experiments and reports them in the .pairs/.pairsam format.

2.2 Terminology

Throughout this document we will be using the same visual language to describe how DNA sequences (in the .fastq format) are transformed into sequence alignments (.sam/.bam) and into ligation events (.pairs).

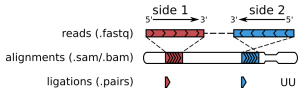

Fig. 1: DNA sequences (reads) are aligned to the reference genome and converted into ligation events

Short-read sequencing determines the sequences of the both ends (or, **sides**) of DNA molecules (typically 50-300 bp), producing **read pairs** in .fastq format (shown in the first row on the figure above). In such reads, base pairs are reported from the tips inwards, which is also defined as the **5'->3'** direction (in accordance of the 5'->3' direction of the DNA strand that sequence of the corresponding side of the read).

Alignment software maps both reads of a pair to the reference genome, producing **alignments**, i.e. segments of the reference genome with matching sequences. Typically, there will be only two alignments per read pair, one on each side. But, sometimes, the parts of one or both sides may map to different locations on the genome, producing more than two alignments per DNA molecule (see *Multiple ligations (walks)*).

`pairsamtools parse` converts alignments into **ligation events** (aka **Hi-C pairs** aka **pairs**). In the simplest case, when each side has only one unique alignment (i.e. the whole side maps to a single unique segment of the genome), for each side, we report the chromosome, the genomic position of the outer-most (5') aligned base pair and the strand of the reference genome that the read aligns to. `pairsamtools parse` assigns to such pairs the type `UU` (unique-unique).

2.3 Unmapped/multimapped reads

Sometimes one side or both sides of a read pair may not align to the reference genome:

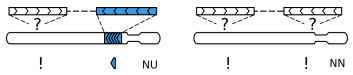

Fig. 2: A read pair missing an alignment on one or both sides

In this case, `pairsamtools parse` fills in the chromosome of the corresponding side of Hi-C pair with `!`, the position with `0` and the strand with `-`. Such pairs are reported as type `NU` (null-unique, when the other side has a unique alignment) or `NN` (null-null, when both sides lack any alignment).

Similarly, when one or both sides map to many genome locations equally well (i.e. have non-unique, or, multi-mapping alignments), `pairsamtools parse` reports the corresponding sides as (chromosome= `!`, position= `0`, strand= `-`) and type `MU` (multi-unique) or `MM` (multi-multi) or `NM` (null-multi), depending on the type of the alignment on the other side.

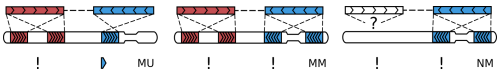

Fig. 3: A read pair with a non-unique (multi-) alignment on one side

`pairsamtools parse` calls an alignment to be multi-mapping when its MAPQ score (which depends on the scoring gap between the two best candidate alignments for a segment) is equal or greater than the value specied with the `--min-mapq` flag (by default, 1).

2.4 Multiple ligations (walks)

Finally, a read pair may contain more than two alignments:

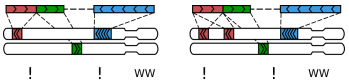

Fig. 4: A sequenced Hi-C molecule that was formed via multiple ligations

Molecules like these typically form via multiple ligation events and we call them walks[1]. Currently, `pairsamtools parse` does not process such molecules and tags them with type `WW`. Note that, each of the alignments

2.5 Interpreting gaps between alignments

Reads that are only partially aligned to the genome can be interpreted it in two different ways. One possibility is to assume that this molecule was formed via at least two ligations (i.e. it's a *walk*) but the non-aligned part (a **gap**) was missing from the reference genome for one reason or another. Another possibility is to simply ignore this gap (for example, because it could be an insertion or a technical artifact), thus assuming that our molecule was formed via a single ligation and has to be reported:

Both options have their merits, depending on a dataset, quality of the reference genome and sequencing. *pairsamtools parse* ignores shorter *gaps* and keeps longer ones as "null" alignments. The maximal size of ignored *gaps* is set by the `--max-inter-align-gap` flag and, by default, equals 20bp.

[1] Following the lead of C-walks

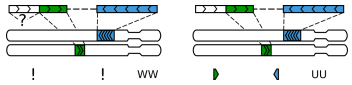

Fig. 5: A gap between alignments can interpeted as a legitimate segment without an alignment or simply ignored

2.6 Rescuing single ligations

Importantly, some of DNA molecules containing only one ligation junction may still end up with three alignments:

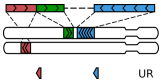

Fig. 6: Not all read pairs with three alignments come from "walks"

A molecule formed via a single ligation gets three alignments when one of the two ligated DNA pieces is shorter than the read length, such that that read on the corresponding side sequences through the ligation site and into the other piece[2]. The fraction of such molecules depends on the type of the restriction enzyme, the typical size of DNA molecules in the Hi-C library and the read length, and sometimes can be considerable.

`pairsamtools parse` detects such molecules and **rescues** them (i.e. changes their type from a *walk* to a single-ligation molecule). It tests walks with three aligments using three criteria:

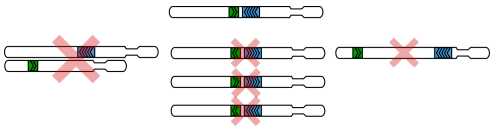

Fig. 7: The three criteria used for "rescue"

1. On the side with two alignments, the "inner" one must be on the same chromosome as the alignment on the other side.
2. The "inner" alignment and the alignment on the other side must point toward each other.
3. These two alignments must be within the distance specified with the `--max-molecule-size` flag (by default, 2000bp).

Sometimes, the "inner" alignment is non-unique or "null" (i.e. when the unmapped segment is longer than `--max-inter-align-gap`, as described in *Interpreting gaps between alignments*). *pairsamtools parse* rescues such *walks* as well.

2.7 Pair flipping

2.8 Other reporting options

[2] This procedure was first introduced in HiC-Pro and the in Juicer .

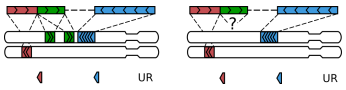

Fig. 8: A walk with three alignments get rescued, when the middle alignment is multi- or null.

CHAPTER

THREE

.PAIRSAM FORMAT

3.1 specification

pairsamtools define .pairsam, a simple tabular format to store the information on ligation junctions detected in sequences of DNA molecules generated by Hi-C experiments.

.pairsam is a valid extension of the .pairs format and is fully compliant with its specification, defined by the 4DN Consortium.

A pairsam starts with an arbitrary number of header lines, each starting with a "#" character. .pairsam headers contain all information mandated by the .pairs format. Additionally, .pairsam format stored the header of the .sam file that it was generated from. When multiple .pairsam files get merge, the stored .sam headers get checked for consistency and merged. Each pairsamtool applied to a .pairsam file adds a brief record to the .sam header.

The body of a pairsam contains a table with a variable number of fields separated by a "t" character (a horizontal tab):

index	name	description
1	read_id	the ID of the read as defined in fastq files
2	chrom1	the chromosome of the alignment on side 1
3	pos1	the 1-based genomic position of the outer-most (5') mapped bp on side 1
4	chrom2	the chromosome of the alignment on side 2
5	pos2	the 1-based genomic position of the outer-most (5') mapped bp on side 2
6	strand1	the strand of the alignment on side 1
7	strand2	the strand of the alignment on side 2
8	pair_type	the type of a Hi-C pair
9	sam1	the sam alignment(s) on side 1; separate supplemental alignments by NEXT_SAM
10	sam2	the sam alignment(s) on side 2; separate supplemental alignments by NEXT_SAM

The sides 1 and 2 as defined in pairsam file do not correspond to side1 and side2 in sequencing data! Instead, side1 is defined as the side with the alignment with a lower sorting index (using the lexographic order for chromosome names, followed by the numeric order for positions and the lexicographic order for pair types). This procedure is defined as upper-triangular flipping, or triu-flipping.

The rows of the table are block-sorted: i.e. first lexicographically by chrom1 and chrom2, then numerically by pos1 and pos2, then lexicographically by pair_type.

Null/ambiguous/chimeric alignments are stored as chrom='!', pos=0, strand='-'.

The columns of the sam records in lines 9 and 10 are separated by a UNIT SEPARATOR character (031) instead of the horizontal tab character, such that it does not affect the columns of the pairsam file.

Notes of the motivation behind some of the technical decisions in the definition of pairsam: - while the information in columns 1-8 may appear redundant to sam alignments in the columns 9+, extracting this information is non-trivial and thus is better done only once with results stored. - storing sam entries together with pairs drastically speeds up

and simplifies several operations like filtering and tagging of unmapped/ambiguous/duplicated Hi-C molecules. - pair flipping and sorting is essential for the processing steps like PCR duplicate removal and aggregation. - the exclamation mark "!" is used as a character for unmapped chromosomes because it has a lexicographic sorting order lower than that of "0", good interpretability and no other reserved technical roles.

3.2 pair types

pairsamtools uses a simple two-character notation to define all possible pair types by the quality of alignment. For each pair, its type can be defined unambiguously using the table below. To use this table, identify which side has an alignment of a "poorer" quality (unmapped < multimapped < unique alignment) and which side has a "better" alignment and find the corresponding row in the table.

.	Less informative alignment		More informative alignment		.	.	.
>2 alignments	Mapped	Unique	Mapped	Unique	Pair type	Code	Sidedness
✓					chimeric-chimeric	CC	0[1]
					null	NN	0
			✓		null-multi	NM	0
			✓	✓	null-unique	NU	1
✓			✓	✓	null-rescued-chimeric	NR	1[2]
	✓		✓		multi-multi	MM	0
	✓		✓	✓	multi-unique	MU	1
✓	✓		✓	✓	multi-rescued-chimeric	MR	2[2]
	✓	✓	✓	✓	unique-unique	UU	2
✓	✓	✓	✓	✓	rescured-chimeric	UR or RU	2[2]
	✓	✓	✓	✓	duplicate	DD	2[3]

[1] chimeric reads represent Hi-C molecules formed via multiple ligation events and thus cannot be reported as a single pair.

[2] some chimeric reads correspond to valid Hi-C molecules formed via a single ligation event, with the ligation junction sequenced through on one side. Following the procedure introduced in [HiC-Pro](https://github.com/nservant/HiC-Pro) [Juicer](https://github.com/theaidenlab/juicer), pairsamtools rescue such molecules, report their outer-most mapped positions and tag them as "UR" or "RU" pair type. Such molecules can and should be used in downstream analysis.

[3] pairsamtools detect molecules that could be formed via PCR duplication and tags them as "DD" pair type. These pairs should be excluded from downstream analyses.

CHAPTER

FOUR

COMMAND-LINE TOOLS

4.1 pairsamtools

```
pairsamtools [OPTIONS] COMMAND [ARGS]...
```

Options

--post-mortem

Post mortem debugging

--output-profile <output_profile>

Profile performance with Python cProfile and dump the statistics into a binary file

--version

Show the version and exit.

4.1.1 dedup

find and remove PCR duplicates.

Find PCR duplicates in an upper-triangular flipped sorted pairs/pairsam file. Allow for a +/-N bp mismatch at each side of duplicated molecules.

PAIRSAM_PATH : input triu-flipped sorted .pairs or .pairsam file. If the path ends with .gz/.lz4, the input is decompressed by pbgzip/lz4c. By default, the input is read from stdin.

```
pairsamtools dedup [OPTIONS] [PAIRSAM_PATH]
```

Options

-o, --output <output>

output file for pairs after duplicate removal. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. By default, the output is printed into stdout.

--output-dups <output_dups>

output file for duplicated pairs. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. If the path is the same as in –output or -, output duplicates together with deduped pairs. By default, duplicates are dropped.

--output-unmapped <output_unmapped>

output file for unmapped pairs. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. If the path is the same as in –output or -, output unmapped pairs together with deduped pairs. If the path is the same as –output-dups, output unmapped reads together with dups. By default, unmapped pairs are dropped.

--output-stats <output_stats>

output file for duplicate statistics. If file exists, it will be open in the append mode. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. By default, statistics are not printed.

--max-mismatch <max_mismatch>

Pairs with both sides mapped within this distance (bp) from each other are considered duplicates.

--method <method>

define the mismatch as either the max or the sum of the mismatches ofthe genomic locations of the both sides of the two compared molecules

Default
max

Options
max | sum

--sep <sep>

Separator (t, v, etc. characters are supported, pass them in quotes)

--comment-char <comment_char>

The first character of comment lines

--send-header-to <send_header_to>

Which of the outputs should receive header and comment lines

Options
dups | dedup | both | none

--c1 <c1>

Chrom 1 column; default 1

--c2 <c2>

Chrom 2 column; default 3

--p1 <p1>

Position 1 column; default 2

--p2 <p2>

Position 2 column; default 4

--s1 <s1>

Strand 1 column; default 5

--s2 <s2>

Strand 2 column; default 6

--unmapped-chrom <unmapped_chrom>

Placeholder for a chromosome on an unmapped side; default !

--mark-dups

If specified, duplicate pairs are marked as DD in “pair_type” and as a duplicate in the sam entries.

--extra-col-pair <extra_col_pair>

Extra columns that also must match for two pairs to be marked as duplicates. Can be either provided as 0-based column indices or as column names (requires the "#columns" header field). The option can be provided multiple times if multiple column pairs must match. Example: –extra-col-pair "phase1" "phase2"

--nproc-in <nproc_in>

Number of processes used by the auto-guessed input decompressing command.

Default
3

--nproc-out <nproc_out>

Number of processes used by the auto-guessed output compressing command.

Default
8

--cmd-in <cmd_in>

A command to decompress the input file. If provided, fully overrides the auto-guessed command. Does not work with stdin. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -dc -n 3

--cmd-out <cmd_out>

A command to compress the output file. If provided, fully overrides the auto-guessed command. Does not work with stdout. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -c -n 8

Arguments

PAIRSAM_PATH

Optional argument

4.1.2 filterbycov

filter out pairs from locations with suspiciously high coverage. Useful for single-cell Hi-C experiments, where coverage is naturally limited by the chromosome copy number.

Find and remove pairs with >(MAX_COV-1) neighbouring pairs within a +/- MAX_DIST bp window around either side.

PAIRSAM_PATH : input triu-flipped sorted .pairs or .pairsam file. If the path ends with .gz/.lz4, the input is decompressed by pbgzip/lz4c. By default, the input is read from stdin.

```
pairsamtools filterbycov [OPTIONS] [PAIRSAM_PATH]
```

Options

-o, --output <output>

output file for pairs from low coverage regions. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. By default, the output is printed into stdout.

--output-highcov <output_highcov>

output file for pairs from high coverage regions. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. If the path is the same as in –output or -, output duplicates together with deduped pairs. By default, duplicates are dropped.

--output-unmapped <output_unmapped>

output file for unmapped pairs. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. If the path is the same as in –output or -, output unmapped pairs together with deduped pairs. If the path is the same as –output-highcov, output unmapped reads together. By default, unmapped pairs are dropped.

--output-stats <output_stats>

output file for statistics of multiple interactors. If file exists, it will be open in the append mode. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. By default, statistics are not printed.

--max-cov <max_cov>

The maximum allowed coverage per region.

--max-dist <max_dist>

The resolution for calculating coverage. For each pair, the local coverage around each end is calculated as (1 + the number of neighbouring pairs within +/- max_dist bp)

--method <method>

calculate the number of neighbouring pairs as either the sum or the max of the number of neighbours on the two sides

Default
max

Options
max | sum

--sep <sep>

Separator (t, v, etc. characters are supported, pass them in quotes)

--comment-char <comment_char>

The first character of comment lines

--send-header-to <send_header_to>

Which of the outputs should receive header and comment lines

Options
lowcov | highcov | both | none

--c1 <c1>

Chrom 1 column; default 1

--c2 <c2>

Chrom 2 column; default 3

--p1 <p1>

Position 1 column; default 2

--p2 <p2>

Position 2 column; default 4

--s1 <s1>

Strand 1 column; default 5

--s2 <s2>

Strand 2 column; default 6

--unmapped-chrom <unmapped_chrom>

Placeholder for a chromosome on an unmapped side; default !

--mark-multi

If specified, duplicate pairs are marked as FF in "pair_type" and as a duplicate in the sam entries.

--nproc-in <nproc_in>

Number of processes used by the auto-guessed input decompressing command.

Default
3

--nproc-out <nproc_out>

Number of processes used by the auto-guessed output compressing command.

Default
8

--cmd-in <cmd_in>

A command to decompress the input file. If provided, fully overrides the auto-guessed command. Does not work with stdin. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -dc -n 3

--cmd-out <cmd_out>

A command to compress the output file. If provided, fully overrides the auto-guessed command. Does not work with stdout. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -c -n 8

Arguments

PAIRSAM_PATH

Optional argument

4.1.3 markasdup

tag all pairsam entries with a duplicate tag.

PAIRSAM_PATH : input .pairsam file. If the path ends with .gz, the input is gzip-decompressed. By default, the input is read from stdin.

```
pairsamtools markasdup [OPTIONS] [PAIRSAM_PATH]
```

Options

-o, --output <output>

output .pairsam file. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. By default, the output is printed into stdout.

--nproc-in <nproc_in>

Number of processes used by the auto-guessed input decompressing command.

Default
3

--nproc-out <nproc_out>

Number of processes used by the auto-guessed output compressing command.

Default
8

--cmd-in <cmd_in>

A command to decompress the input file. If provided, fully overrides the auto-guessed command. Does not work with stdin. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -dc -n 3

--cmd-out <cmd_out>

A command to compress the output file. If provided, fully overrides the auto-guessed command. Does not work with stdout. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -c -n 8

Arguments

PAIRSAM_PATH

Optional argument

4.1.4 merge

merge sorted pairs/pairsam files.

Merge triu-flipped sorted pairs/pairsam files. If present, the @SQ records of the SAM header must be identical; the sorting order of these lines is taken from the first file in the list. The ID fields of the @PG records of the SAM header are modified with a numeric suffix to produce unique records. The other unique SAM and non-SAM header lines are copied into the output header.

PAIRSAM_PATH : upper-triangular flipped sorted pairs/pairsam files to merge or a group/groups of .pairsam files specified by a wildcard. For paths ending in .gz/.lz4, the files are decompressed by pbgzip/lz4c.

```
pairsamtools merge [OPTIONS] [PAIRSAM_PATH]...
```

Options

-o, --output <output>

output file. If the path ends with .gz/.lz4, the output is compressed by pbgzip/lz4c. By default, the output is printed into stdout.

--max-nmerge <max_nmerge>

The maximal number of inputs merged at once. For more, store merged intermediates in temporary files.

Default
8

--tmpdir <tmpdir>

Custom temporary folder for merged intermediates.

--memory <memory>

The amount of memory used by default.

Default
2G

--compress-program <compress_program>

A binary to compress temporary merged chunks. Must decompress input when the flag -d is provided. Suggested alternatives: lz4c, gzip, lzop, snzip. NOTE: fails silently if the command syntax is wrong.

Default

--nproc <nproc>

Number of threads for merging.

Default
8

--nproc-in <nproc_in>

Number of processes used by the auto-guessed input decompressing command.

Default
1

--nproc-out <nproc_out>

Number of processes used by the auto-guessed output compressing command.

Default
8

--cmd-in <cmd_in>

A command to decompress the input. If provided, fully overrides the auto-guessed command. Does not work with stdin. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -dc -n 3

--cmd-out <cmd_out>

A command to compress the output. If provided, fully overrides the auto-guessed command. Does not work with stdout. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -c -n 8

Arguments

PAIRSAM_PATH

Optional argument(s)

4.1.5 parse

parse .sam and make .pairsam.

SAM_PATH : input .sam file. If the path ends with .bam, the input is decompressed from bam. By default, the input is read from stdin.

```
pairsamtools parse [OPTIONS] [SAM_PATH]
```

Options

-c, --chroms-path <chroms_path>

Required Chromosome order used to flip interchromosomal mates: path to a chromosomes file (e.g. UCSC chrom.sizes or similar) whose first column lists scaffold names. Any scaffolds not listed will be ordered lexicographically following the names provided.

-o, --output <output>

output file. If the path ends with .gz or .lz4, the output is pbgzip-/lz4-compressed.By default, the output is printed into stdout.

--assembly <assembly>

Name of genome assembly (e.g. hg19, mm10) to store in the pairs header.

--min-mapq <min_mapq>

The minimal MAPQ score to consider a read as uniquely mapped

Default
1

--max-molecule-size <max_molecule_size>

The maximal size of a Hi-C molecule; used to rescue single ligationsfrom molecules with three alignments.

Default
2000

--drop-readid

If specified, do not add read ids to the output

--drop-seq

If specified, remove sequences and PHREDs from the sam fields

--drop-sam

If specified, do not add sams to the output

--add-columns <add_columns>

Report extra columns describing alignments Possible values (can take multiple values as a comma-separated list): a SAM tag (any pair of uppercase letters) or mapq, pos5, pos3, cigar, read_len, matched_bp, algn_ref_span, algn_read_span, dist_to_5, dist_to_3, seq.

--output-parsed-alignments <output_parsed_alignments>

output file for all parsed alignments, including walks. Useful for debugging and rnalysis of walks. If file exists, it will be open in the append mode. If the path ends with .gz or .lz4, the output is pbgzip-/lz4-compressed. By default, not used.

--output-stats <output_stats>

output file for various statistics of pairsam file. By default, statistics is not generated.

--report-alignment-end <report_alignment_end>

specifies whether the 5' or 3' end of the alignment is reported as the position of the Hi-C read.

Options
5 | 3

--max-inter-align-gap <max_inter_align_gap>

read segments that are not covered by any alignment and longer than the specified value are treated as "null" alignments. These null alignments convert otherwise linear alignments into walks, and affect how they get reported as a Hi-C pair (see –walks-policy).

Default
20

--walks-policy <walks_policy>

the policy for reporting unrescuable walks (reads containing more than one alignment on one or both sides, that can not be explained by a single ligation between two mappable DNA fragments). "mask" - mask walks (chrom="!", pos=0, strand="-"); "all" - report all pairs of consecutive alignments [NOT IMPLEMENTED]; "5any" - report the 5'-most alignment on each side; "5unique" - report the 5'-most unique alignment on each side, if present; "3any" - report the 3'-most alignment on each side; "3unique" - report the 3'-most unique alignment on each side, if present.

Default
mask

Options
mask | all | 5any | 5unique | 3any | 3unique

--no-flip
If specified, do not flip pairs in genomic order and instead preserve the order in which they were sequenced.

--nproc-in <nproc_in>
Number of processes used by the auto-guessed input decompressing command.

Default
3

--nproc-out <nproc_out>
Number of processes used by the auto-guessed output compressing command.

Default
8

--cmd-in <cmd_in>
A command to decompress the input file. If provided, fully overrides the auto-guessed command. Does not work with stdin. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -dc -n 3

--cmd-out <cmd_out>
A command to compress the output file. If provided, fully overrides the auto-guessed command. Does not work with stdout. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -c -n 8

Arguments

SAM_PATH
Optional argument

4.1.6 phase

phase a pairsam file mapped to a diploid genome.

PAIRSAM_PATH : input .pairsam file. If the path ends with .gz or .lz4, the input is decompressed by pbgzip/lz4c. By default, the input is read from stdin.

```
pairsamtools phase [OPTIONS] [PAIRSAM_PATH]
```

Options

-o, --output <output>
output file. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. By default, the output is printed into stdout.

--phase-suffixes <phase_suffixes>
phase suffixes.

--clean-output
drop all columns besides the standard ones and phase1/2

--nproc-in <nproc_in>

Number of processes used by the auto-guessed input decompressing command.

Default
3

--nproc-out <nproc_out>

Number of processes used by the auto-guessed output compressing command.

Default
8

--cmd-in <cmd_in>

A command to decompress the input file. If provided, fully overrides the auto-guessed command. Does not work with stdin. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -dc -n 3

--cmd-out <cmd_out>

A command to compress the output file. If provided, fully overrides the auto-guessed command. Does not work with stdout. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -c -n 8

Arguments

PAIRSAM_PATH

Optional argument

4.1.7 restrict

identify the restriction fragments that got ligated into a Hi-C molecule.

PAIRSAM_PATH : input .pairsam file. If the path ends with .gz/.lz4, the input is decompressed by pbgzip/lz4c. By default, the input is read from stdin.

```
pairsamtools restrict [OPTIONS] [PAIRSAM_PATH]
```

Options

-f, --frags <frags>

Required a tab-separated BED file with the positions of restriction fragments (chrom, start, end). Can be generated using cooler digest.

-o, --output <output>

output pairsam file. If the path ends with .gz/.lz4, the output is compressed by pbgzip/lz4c. By default, the output is printed into stdout.

--nproc-in <nproc_in>

Number of processes used by the auto-guessed input decompressing command.

Default
3

--nproc-out <nproc_out>

Number of processes used by the auto-guessed output compressing command.

Default
8

--cmd-in <cmd_in>

A command to decompress the input file. If provided, fully overrides the auto-guessed command. Does not work with stdin. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -dc -n 3

--cmd-out <cmd_out>

A command to compress the output file. If provided, fully overrides the auto-guessed command. Does not work with stdout. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -c -n 8

Arguments

PAIRSAM_PATH

Optional argument

4.1.8 select

select pairsam entries.

CONDITION : A Python expression; if it returns True, select the read pair. Any column declared in the #columns line of the pairs header can be accessed by its name. If the header lacks the #columns line, the columns are assumed to follow the pairs/pairsam standard (readID, chrom1, chrom2, pos1, pos2, strand1, strand2, pair_type). Finally, CONDITION has access to COLS list which contains the string values of columns. In Bash, quote CONDITION with single quotes, and use double quotes for string variables inside CONDITION.

PAIRSAM_PATH : input .pairsam file. If the path ends with .gz or .lz4, the input is decompressed by pbgzip/lz4c. By default, the input is read from stdin.

The following functions can be used in CONDITION besides the standard Python functions:

- csv_match(x, csv) - True if variable x is contained in a list of

comma-separated values, e.g. csv_match(chrom1, 'chr1,chr2')

- wildcard_match(x, wildcard) - True if variable x matches a wildcard,

e.g. wildcard_match(pair_type, 'C*')

- regex_match(x, regex) - True if variable x matches a Python-flavor regex,

e.g. regex_match(chrom1, 'chrd')

Examples:
pairsamtools select '(pair_type=="UU") or (pair_type=="UR") or (pair_type=="RU")'
pairsamtools select 'chrom1==chrom2'
pairsamtools select 'COLS[1]==COLS[3]'
pairsamtools select '(chrom1==chrom2) and (abs(pos1 - pos2) < 1e6)'
pairsamtools select '(chrom1=="!") and (chrom2!="!")'
pairsamtools select 'regex_match(chrom1, "chrd+") and regex_match(chrom2, "chrd+")'

pairsamtools select 'True' –chr-subset mm9.reduced.chromsizes

```
pairsamtools select [OPTIONS] CONDITION [PAIRSAM_PATH]
```

Options

-o, --output <output>

output file. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. By default, the output is printed into stdout.

--output-rest <output_rest>

output file for pairs of other types. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. By default, such pairs are dropped.

--send-comments-to <send_comments_to>

Which of the outputs should receive header and comment lines

Default
both

Options
selected | rest | both | none

--chrom-subset <chrom_subset>

A path to a chromosomes file (tab-separated, 1st column contains chromosome names) containing a chromosome subset of interest. If provided, additionally filter pairs with both sides originating from the provided subset of chromosomes. This operation modifies the #chromosomes: and #chromsize: header fields accordingly.

--nproc-in <nproc_in>

Number of processes used by the auto-guessed input decompressing command.

Default
3

--nproc-out <nproc_out>

Number of processes used by the auto-guessed output compressing command.

Default
8

--cmd-in <cmd_in>

A command to decompress the input file. If provided, fully overrides the auto-guessed command. Does not work with stdin. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -dc -n 3

--cmd-out <cmd_out>

A command to compress the output file. If provided, fully overrides the auto-guessed command. Does not work with stdout. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -c -n 8

Arguments

CONDITION

Required argument

PAIRSAM_PATH

Optional argument

4.1.9 sort

sort a pairs/pairsam file.

The resulting order is lexicographic along chrom1 and chrom2, numeric along pos1 and pos2 and lexicographic along pair_type.

PAIRSAM_PATH : input .pairsam file. If the path ends with .gz or .lz4, the input is decompressed by pbgzip or lz4c, correspondingly. By default, the input is read as text from stdin.

```
pairsamtools sort [OPTIONS] [PAIRSAM_PATH]
```

Options

-o, --output <output>

output pairsam file. If the path ends with .gz or .lz4, the output is compressed by pbgzip or lz4, correspondingly. By default, the output is printed into stdout.

--nproc <nproc>

Number of processes to split the sorting work between.

Default
8

--tmpdir <tmpdir>

Custom temporary folder for sorting intermediates.

--memory <memory>

The amount of memory used by default.

Default
2G

--compress-program <compress_program>

A binary to compress temporary sorted chunks. Must decompress input when the flag -d is provided. Suggested alternatives: gzip, lzop, lz4c, snzip.

Default

--nproc-in <nproc_in>

Number of processes used by the auto-guessed input decompressing command.

Default
3

--nproc-out <nproc_out>

Number of processes used by the auto-guessed output compressing command.

Default
8

--cmd-in <cmd_in>

A command to decompress the input file. If provided, fully overrides the auto-guessed command. Does not work with stdin. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -dc -n 3

--cmd-out <cmd_out>

A command to compress the output file. If provided, fully overrides the auto-guessed command. Does not work with stdout. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -c -n 8

Arguments

PAIRSAM_PATH

Optional argument

4.1.10 split

split a .pairsam file into pairs and sam.

PAIRSAM_PATH : input .pairsam file. If the path ends with .gz or .lz4, the input is decompressed by pbgzip or lz4c. By default, the input is read from stdin.

```
pairsamtools split [OPTIONS] [PAIRSAM_PATH]
```

Options

--output-pairs <output_pairs>

output pairs file. If the path ends with .gz or .lz4, the output is pbgzip-/lz4c-compressed. If -, pairs are printed to stdout. If not specified, pairs are dropped.

--output-sam <output_sam>

output sam file. If the path ends with .bam, the output is compressed into a bam file. If -, sam entries are printed to stdout. If not specified, sam entries are dropped.

--nproc-in <nproc_in>

Number of processes used by the auto-guessed input decompressing command.

Default

3

--nproc-out <nproc_out>

Number of processes used by the auto-guessed output compressing command.

Default

8

--cmd-in <cmd_in>

A command to decompress the input file. If provided, fully overrides the auto-guessed command. Does not work with stdin. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -dc -n 3

--cmd-out <cmd_out>

A command to compress the output file. If provided, fully overrides the auto-guessed command. Does not work with stdout. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -c -n 8

Arguments

PAIRSAM_PATH

Optional argument

4.1.11 stats

calculate various statistics of a pairs/pairsam file.

INPUT_PATH : by default, a .pairsam file to calculate statistics. If not provided, the input is read from stdin. If –merge is specified, then INPUT_PATH is interpreted as an arbitrary number of stats files to merge.

The files with paths ending with .gz/.lz4 are decompressed by pbgzip/lz4c.

```
pairsamtools stats [OPTIONS] [INPUT_PATH]...
```

Options

-o, --output <output>

output stats tsv file.

--merge

If specified, merge multiple input stats files instead of calculating statistics of a pairsam file. Merging is performed via summation of all overlapping statistics. Non-overlapping statistics are appended to the end of the file.

--nproc-in <nproc_in>

Number of processes used by the auto-guessed input decompressing command.

Default
3

--nproc-out <nproc_out>

Number of processes used by the auto-guessed output compressing command.

Default
8

--cmd-in <cmd_in>

A command to decompress the input file. If provided, fully overrides the auto-guessed command. Does not work with stdin. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -dc -n 3

--cmd-out <cmd_out>

A command to compress the output file. If provided, fully overrides the auto-guessed command. Does not work with stdout. Must read input from stdin and print output into stdout. EXAMPLE: pbgzip -c -n 8

Arguments

INPUT_PATH

Optional argument(s)

CHAPTER

FIVE

INDICES AND TABLES

- genindex
- modindex
- search

INDEX

Symbols

--add-columns
    pairsamtools-parse command line option, 18
--assembly
    pairsamtools-parse command line option, 17
--c1
    pairsamtools-dedup command line option, 12
    pairsamtools-filterbycov command line option, 14
--c2
    pairsamtools-dedup command line option, 12
    pairsamtools-filterbycov command line option, 14
--chrom-subset
    pairsamtools-select command line option, 22
--chroms-path
    pairsamtools-parse command line option, 17
--clean-output
    pairsamtools-phase command line option, 19
--cmd-in
    pairsamtools-dedup command line option, 13
    pairsamtools-filterbycov command line option, 15
    pairsamtools-markasdup command line option, 15
    pairsamtools-merge command line option, 17
    pairsamtools-parse command line option, 19
    pairsamtools-phase command line option, 20
    pairsamtools-restrict command line option, 20
    pairsamtools-select command line option, 22
    pairsamtools-sort command line option, 23
    pairsamtools-split command line option, 24
    pairsamtools-stats command line option, 25
--cmd-out
    pairsamtools-dedup command line option, 13
    pairsamtools-filterbycov command line option, 15
    pairsamtools-markasdup command line option, 16
    pairsamtools-merge command line option, 17
    pairsamtools-parse command line option, 19
    pairsamtools-phase command line option, 20
    pairsamtools-restrict command line option, 21
    pairsamtools-select command line option, 22
    pairsamtools-sort command line option, 23
    pairsamtools-split command line option, 24
    pairsamtools-stats command line option, 25
--comment-char
    pairsamtools-dedup command line option, 12
    pairsamtools-filterbycov command line option, 14
--compress-program
    pairsamtools-merge command line option, 16
    pairsamtools-sort command line option, 23
--drop-readid
    pairsamtools-parse command line option, 18
--drop-sam
    pairsamtools-parse command line option,

18
--drop-seq
pairsamtools-parse command line option, 18
--extra-col-pair
pairsamtools-dedup command line option, 12
--frags
pairsamtools-restrict command line option, 20
--mark-dups
pairsamtools-dedup command line option, 12
--mark-multi
pairsamtools-filterbycov command line option, 14
--max-cov
pairsamtools-filterbycov command line option, 14
--max-dist
pairsamtools-filterbycov command line option, 14
--max-inter-align-gap
pairsamtools-parse command line option, 18
--max-mismatch
pairsamtools-dedup command line option, 12
--max-molecule-size
pairsamtools-parse command line option, 18
--max-nmerge
pairsamtools-merge command line option, 16
--memory
pairsamtools-merge command line option, 16
pairsamtools-sort command line option, 23
--merge
pairsamtools-stats command line option, 25
--method
pairsamtools-dedup command line option, 12
pairsamtools-filterbycov command line option, 14
--min-mapq
pairsamtools-parse command line option, 17
--no-flip
pairsamtools-parse command line option, 19
--nproc
pairsamtools-merge command line option, 16
pairsamtools-sort command line option, 23
--nproc-in
pairsamtools-dedup command line option, 13
pairsamtools-filterbycov command line option, 15
pairsamtools-markasdup command line option, 15
pairsamtools-merge command line option, 17
pairsamtools-parse command line option, 19
pairsamtools-phase command line option, 19
pairsamtools-restrict command line option, 20
pairsamtools-select command line option, 22
pairsamtools-sort command line option, 23
pairsamtools-split command line option, 24
pairsamtools-stats command line option, 25
--nproc-out
pairsamtools-dedup command line option, 13
pairsamtools-filterbycov command line option, 15
pairsamtools-markasdup command line option, 15
pairsamtools-merge command line option, 17
pairsamtools-parse command line option, 19
pairsamtools-phase command line option, 20
pairsamtools-restrict command line option, 20
pairsamtools-select command line option, 22
pairsamtools-sort command line option, 23
pairsamtools-split command line option, 24
pairsamtools-stats command line option, 25
--output
pairsamtools-dedup command line option, 11
pairsamtools-filterbycov command line option, 13
pairsamtools-markasdup command line option, 15
pairsamtools-merge command line option,

16
pairsamtools-parse command line option, 17
pairsamtools-phase command line option, 19
pairsamtools-restrict command line option, 20
pairsamtools-select command line option, 22
pairsamtools-sort command line option, 23
pairsamtools-stats command line option, 25
--output-dups
pairsamtools-dedup command line option, 11
--output-highcov
pairsamtools-filterbycov command line option, 13
--output-pairs
pairsamtools-split command line option, 24
--output-parsed-alignments
pairsamtools-parse command line option, 18
--output-profile
pairsamtools command line option, 11
--output-rest
pairsamtools-select command line option, 22
--output-sam
pairsamtools-split command line option, 24
--output-stats
pairsamtools-dedup command line option, 12
pairsamtools-filterbycov command line option, 14
pairsamtools-parse command line option, 18
--output-unmapped
pairsamtools-dedup command line option, 11
pairsamtools-filterbycov command line option, 13
--p1
pairsamtools-dedup command line option, 12
pairsamtools-filterbycov command line option, 14
--p2
pairsamtools-dedup command line option, 12
pairsamtools-filterbycov command line option, 14
--phase-suffixes
pairsamtools-phase command line option, 19
--post-mortem
pairsamtools command line option, 11
--report-alignment-end
pairsamtools-parse command line option, 18
--s1
pairsamtools-dedup command line option, 12
pairsamtools-filterbycov command line option, 14
--s2
pairsamtools-dedup command line option, 12
pairsamtools-filterbycov command line option, 14
--send-comments-to
pairsamtools-select command line option, 22
--send-header-to
pairsamtools-dedup command line option, 12
pairsamtools-filterbycov command line option, 14
--sep
pairsamtools-dedup command line option, 12
pairsamtools-filterbycov command line option, 14
--tmpdir
pairsamtools-merge command line option, 16
pairsamtools-sort command line option, 23
--unmapped-chrom
pairsamtools-dedup command line option, 12
pairsamtools-filterbycov command line option, 14
--version
pairsamtools command line option, 11
--walks-policy
pairsamtools-parse command line option, 18
-c
pairsamtools-parse command line option, 17
-f
pairsamtools-restrict command line option, 20
-o
pairsamtools-dedup command line option, 11

pairsamtools-filterbycov command line option, 13
pairsamtools-markasdup command line option, 15
pairsamtools-merge command line option, 16
pairsamtools-parse command line option, 17
pairsamtools-phase command line option, 19
pairsamtools-restrict command line option, 20
pairsamtools-select command line option, 22
pairsamtools-sort command line option, 23
pairsamtools-stats command line option, 25

C

CONDITION
pairsamtools-select command line option, 22

I

INPUT_PATH
pairsamtools-stats command line option, 25

P

PAIRSAM_PATH
pairsamtools-dedup command line option, 13
pairsamtools-filterbycov command line option, 15
pairsamtools-markasdup command line option, 16
pairsamtools-merge command line option, 17
pairsamtools-phase command line option, 20
pairsamtools-restrict command line option, 21
pairsamtools-select command line option, 22
pairsamtools-sort command line option, 24
pairsamtools-split command line option, 24
pairsamtools command line option
--output-profile, 11
--post-mortem, 11
--version, 11
pairsamtools-dedup command line option
--c1, 12
--c2, 12
--cmd-in, 13
--cmd-out, 13
--comment-char, 12
--extra-col-pair, 12
--mark-dups, 12
--max-mismatch, 12
--method, 12
--nproc-in, 13
--nproc-out, 13
--output, 11
--output-dups, 11
--output-stats, 12
--output-unmapped, 11
--p1, 12
--p2, 12
--s1, 12
--s2, 12
--send-header-to, 12
--sep, 12
--unmapped-chrom, 12
-o, 11
PAIRSAM_PATH, 13
pairsamtools-filterbycov command line option
--c1, 14
--c2, 14
--cmd-in, 15
--cmd-out, 15
--comment-char, 14
--mark-multi, 14
--max-cov, 14
--max-dist, 14
--method, 14
--nproc-in, 15
--nproc-out, 15
--output, 13
--output-highcov, 13
--output-stats, 14
--output-unmapped, 13
--p1, 14
--p2, 14
--s1, 14
--s2, 14
--send-header-to, 14
--sep, 14
--unmapped-chrom, 14
-o, 13
PAIRSAM_PATH, 15
pairsamtools-markasdup command line option
--cmd-in, 15
--cmd-out, 16
--nproc-in, 15
--nproc-out, 15
--output, 15

-o, 15
PAIRSAM_PATH, 16
pairsamtools-merge command line option
--cmd-in, 17
--cmd-out, 17
--compress-program, 16
--max-nmerge, 16
--memory, 16
--nproc, 16
--nproc-in, 17
--nproc-out, 17
--output, 16
--tmpdir, 16
-o, 16
PAIRSAM_PATH, 17
pairsamtools-parse command line option
--add-columns, 18
--assembly, 17
--chroms-path, 17
--cmd-in, 19
--cmd-out, 19
--drop-readid, 18
--drop-sam, 18
--drop-seq, 18
--max-inter-align-gap, 18
--max-molecule-size, 18
--min-mapq, 17
--no-flip, 19
--nproc-in, 19
--nproc-out, 19
--output, 17
--output-parsed-alignments, 18
--output-stats, 18
--report-alignment-end, 18
--walks-policy, 18
-c, 17
-o, 17
SAM_PATH, 19
pairsamtools-phase command line option
--clean-output, 19
--cmd-in, 20
--cmd-out, 20
--nproc-in, 19
--nproc-out, 20
--output, 19
--phase-suffixes, 19
-o, 19
PAIRSAM_PATH, 20
pairsamtools-restrict command line option
--cmd-in, 20
--cmd-out, 21
--frags, 20
--nproc-in, 20
--nproc-out, 20
--output, 20
-f, 20
-o, 20
PAIRSAM_PATH, 21
pairsamtools-select command line option
--chrom-subset, 22
--cmd-in, 22
--cmd-out, 22
--nproc-in, 22
--nproc-out, 22
--output, 22
--output-rest, 22
--send-comments-to, 22
-o, 22
CONDITION, 22
PAIRSAM_PATH, 22
pairsamtools-sort command line option
--cmd-in, 23
--cmd-out, 23
--compress-program, 23
--memory, 23
--nproc, 23
--nproc-in, 23
--nproc-out, 23
--output, 23
--tmpdir, 23
-o, 23
PAIRSAM_PATH, 24
pairsamtools-split command line option
--cmd-in, 24
--cmd-out, 24
--nproc-in, 24
--nproc-out, 24
--output-pairs, 24
--output-sam, 24
PAIRSAM_PATH, 24
pairsamtools-stats command line option
--cmd-in, 25
--cmd-out, 25
--merge, 25
--nproc-in, 25
--nproc-out, 25
--output, 25
-o, 25
INPUT_PATH, 25

S

SAM_PATH
pairsamtools-parse command line option, 19