Reference-based PCR Duplicate Removal

October 9, 2017

Adapted from Peter Batzel

PCR duplicate

- What is a PCR duplicate?
- Why do we care about removing them?
- What are some stratagies to identify them?

What does a PCR-duplicate look like?

- SAM format: https://samtools.github.io/hts-specs/SAMv1.pdf

```
NS500451:154:HWKTMBGXX:1:11101:16635:1076-GAGAANAG^GAAGACCA;0^0
83   11    67245662    36    71M    =    67245443   -293
AGGTGTACAACTCCGTGGGTGCCCTGGCCAAGTCCATGTATGAGAAGATGTTCCTATGGATGGTCACCC
GCEEEAEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
EE66MD:Z:71   NH:i:1    HI:i:1     NM:i:0     SM:i:36 XQ:i:40    X2:i:0
XO:Z:CU    XG:Z:A
```

Col	Field	Type	Description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	Reference sequence NAME
4	POS	Int	1-based leftmost mapping POSition
5	MAPQ	Int	MAPping quality
6	CIGAR	String	CIGAR string
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

What does a PCR-duplicate look like?

- Same alignment position
  - Chromosome
  - Position
  - Strand (strand specific?)
- Soft Clipping

```
if((flag & 4) != 4):
  mapped = True
```

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate
0x800	supplementary alignment

Soft clipping

- What is it?
- What does it look like?
- Why would something be soft clipped?
  - Sequence error/heterozygosity
  - Over-penalizing indels
  - Splicing with just a few nucleotides in an exon
  - Novel splicing
- Where in the alignment could soft clipping occur?

The CIGAR string

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

The CIGAR string

- Example 1:

  Reference: `...CTTCTATTATCCTT...`

  Read: `CTTCTATTATCCTT`

- CIGAR string: 14M

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

http://zenfractal.com/2013/06/19/playing-with-matches/

The CIGAR string

- Example 2:

  Reference: . . . CTTCTATTATCCTT . . .

  Read: CTTATATTATCCTT

- CIGAR string: ?

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

http://zenfractal.com/2013/06/19/playing-with-matches/

The CIGAR string

- Example 2:

  Reference: . . . CTTCTATTATCCTT . . .

  Read: CTTATATTATCCTT

- CIGAR string: 14M
  - Lookup MG tag in SAM specs

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

http://zenfractal.com/2013/06/19/playing-with-matches/

The CIGAR string

- Example 3:

  Reference: . . . CTTCTATTATCCTT . . .

  Read: **AG**TCTATTATCCTT

- CIGAR string: 2S12M
- Where does the alignment start?

What does a PCR-duplicate look like?

- Same alignment position
  - Chromosome
  - Position
  - Strand (strand specific?)
- Soft Clipping
- Same Unique Molecular Index (UMI or “randomer”)

PCR Duplicate Removal Tools: samtools

- Samtools rmdup
  - Same alignment position
    - Chromosome
    - Position
    - Strand (if strand specific)
  - Soft clipping?
  - http://www.htslib.org/doc/samtools.html

PCR Duplicate Removal Tools: Picard

- Picard MarkDuplicates
  - Same alignment position
    - Chromosome
    - Position
    - Strand (if strand specific)
  - Accounts for soft clipping
  - https://broadinstitute.github.io/picard/commandline-overview.html#MarkDuplicates

PCR Duplicate Removal Tools: UMI-tools

- UMI-tools
  - Same alignment position
    - Chromosome
    - Position
    - Strand (if strand specific)
  - Adjusts 5’ alignment for “simple soft clipping”
  - Accounts for UMIs
  - https://github.com/CGATOxford/UMI-tools

  - What if the 5’ with soft clipping is misleading/not simple?
  - What if 5’ ends are low quality?

PCR Duplicate Removal Tools: SuperDeDuper

- SuperDeDuper – Jason!
  - Same alignment position
    - Chromosome
    - Position
    - Strand (if strand specific)
  - Accounts for UMIs
  - Accounts for complex soft clipping
    - Splicing, indels, SNPs, quality trimmed nucleotides, etc

What does a PCR-duplicate look like?

- Same alignment position
  - Chromosome
  - Position
  - Strand (strand specific?)
- Soft Clipping
- Same Unique Molecular Index (UMI or “randomer”)
- Single-end vs Paired-end?

Reference-free Duplicate Removal

- FastUniq
  - Alphanumeric sort, then remove (adjacent) duplicates
  - Slow, memory intensive, no mismatches
- Clone_filter (Stacks)
  - Buckets sequences by UMIs
    - pairs of UMIs if paired-end
  - Removes duplicates
  - Slow, memory intensive, no mismatches

Reference-free Duplicate Removal

- Problem is sequencing error
  - Errors in UMIs → False negatives
  - Errors in reads → False negatives
- Error correction of UMIs?
  - Maybe…
- Error correction of reads?
  - Not a great idea
    - Computationally cost prohibitive
    - Can cause more errors (i.e. misalignment)

Your algorithm!

- Given a SAM file of uniquely mapped reads, remove all PCR duplicates (retain only a single copy of each read)

- Samtools sort
- Adjust for soft clipping
- Start with single-end, then expand to paired-end
- Start with known UMIs, then expand to randomers
- Millions of reads! Develop a strategy to avoid just loading everything into memory!

DeDuper Part 1

- Your assignment is to develop a strategy to tackle this problem
- Do NOT write any code for Part 1 of this assignment!
- Define the problem, write examples
- Develop your algorithm using pseudocode
- Determine high level functions
  - Description
  - Function header
  - Test examples
  - Return statement