Reference-based PCR Duplicate Removal

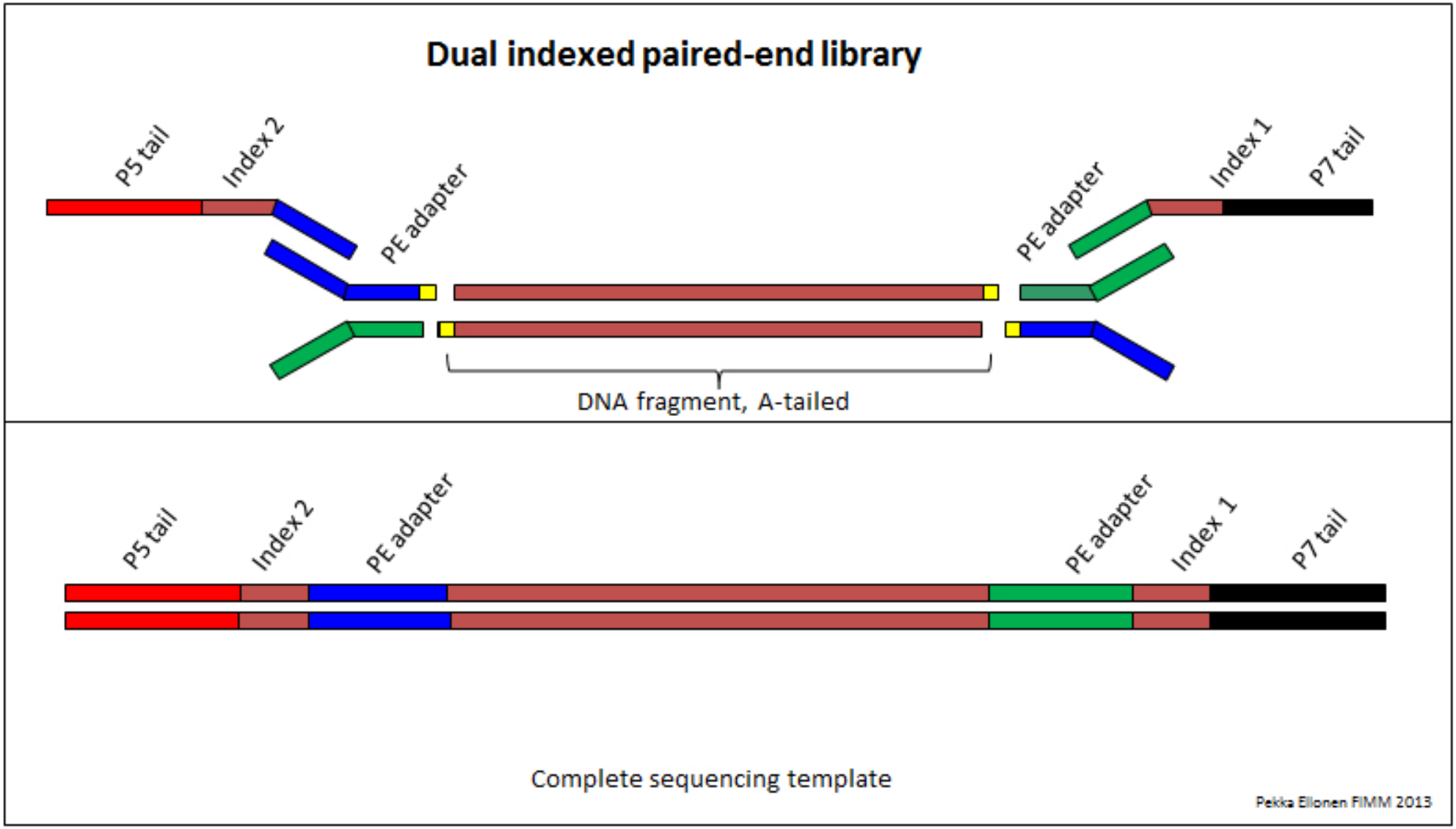
October 8, 2020

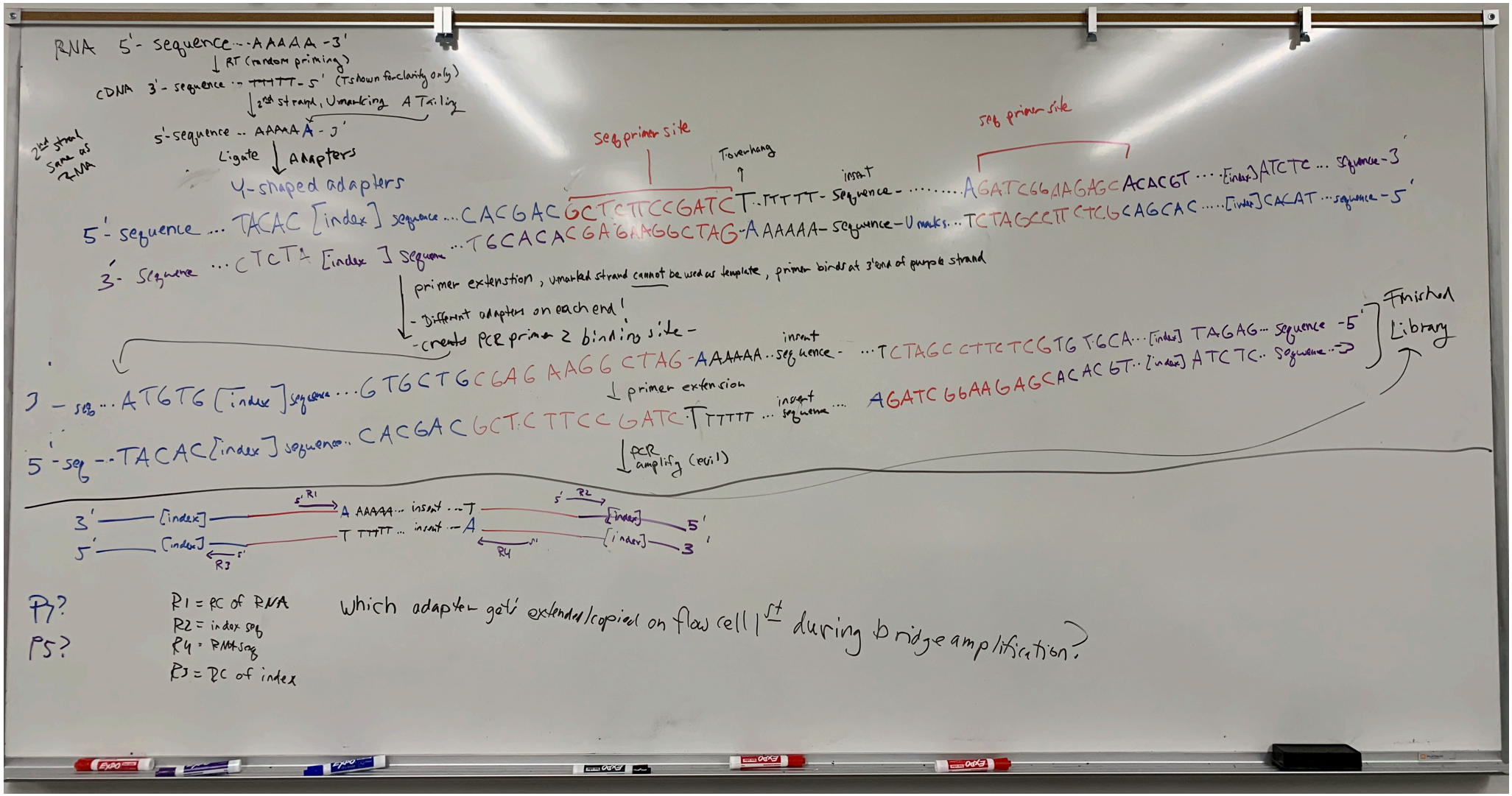
PCR duplicates

- What is a PCR duplicate?
- Where do they come from?
- Why do we care about removing them?
- What are some strategies to identify them?

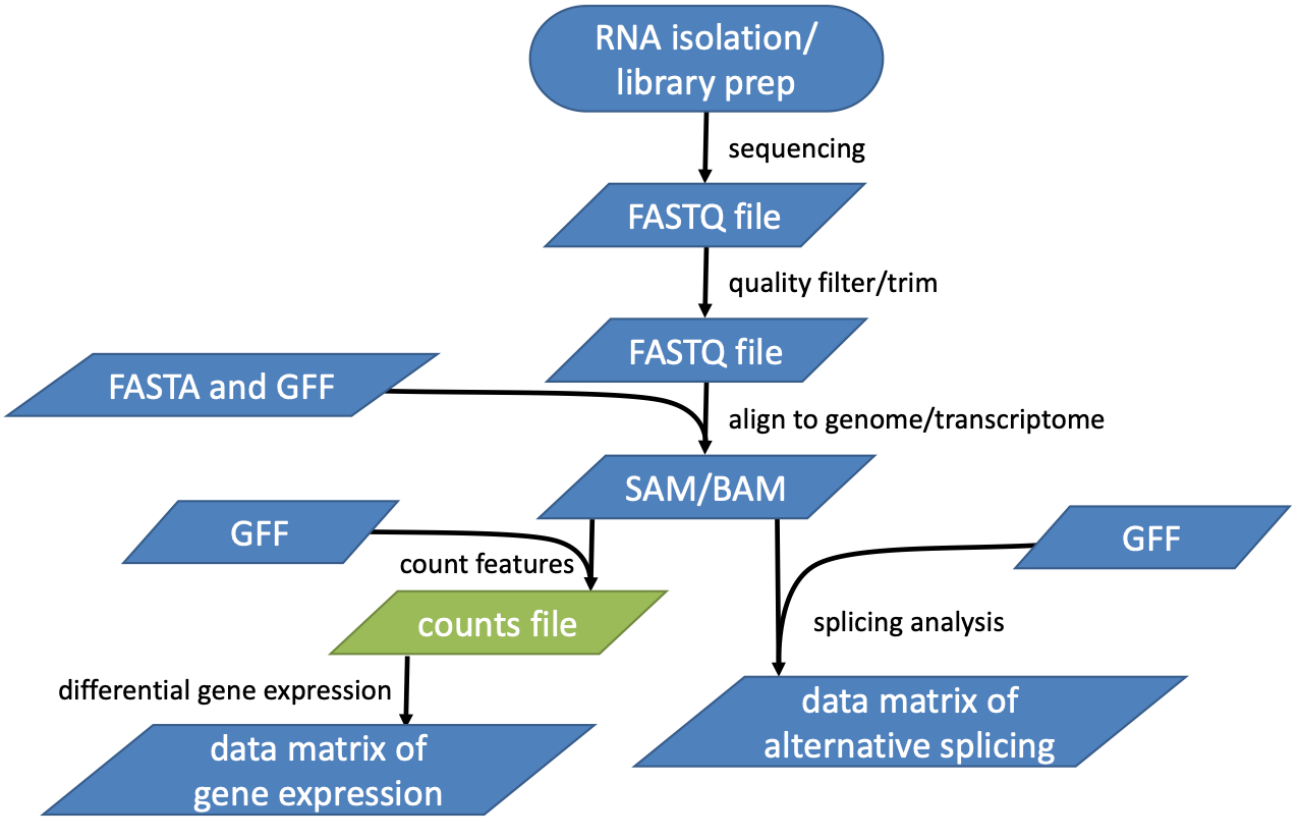
RNA-seq experimental workflow

- Experimental design
- Isolate RNA from biological sample
- Check RNA quality
- Enrichment
- Fragmentation of RNA
- Synthesize cDNA
- Add adapters
- Indexing (barcoding)
- Amplification
- Purification


Dual indexed paired-end library
P5 tail
Index 2
PE adapter
PE adapter
Index 1
P7 tail
DNA fragment, A-tailed
P5 tail
Index 2
PE adapter
PE adapter
Index 1
P7 tail
Complete sequencing template
Pekka Ellonen FIMM 2013


RNA 5'- sequence ---AAAAA -3'
RT (random priming)
cDNA 3'- sequence ---TTTTT- 5' (T shown for clarity only)
2nd strand, Unmarking, A Tailing
5'-sequence -- AAAAA A - 3'
2nd strand same as RNA
Ligate Adapters
Y-shaped adapters
Seq primer site
T overhang
insert
Seq primer site
5'- sequence ... TACAC [index] sequence ... CACGAC GCTCTTCCGATCT ..TTTTT- sequence - ..... AGATCGGAAGAGCACACGT ... [index] ATCTC ... sequence -3'
3'- sequence ... CTCTA [index] sequence ... TGCACACGAGAAGGCTAG-AAAAAA- sequence - U marks ... TCTAGCCTTCTCGTGCAGCAC ..... [index] CACAT ... sequence -5'
primer extension, unmarked strand cannot be used as template, primer binds at 3' end of purple strand
- Different adapters on each end!
- creates PCR primer binding site -
Finished Library
3' - seq ... ATGTG [index] sequence ... GTGCTGCGAG AAGG CTAG - AAAAAA .. seq sequence - insert ... TCTAGC CTTCTCGTG TGCA ... [index] TAGAG ... sequence -5'
primer extension
5' - seq -- TACAC [index] sequence .. CACGAC GCTCTTCCGATC T TTTTT ... insert sequence .... AGATC GGAAGAGCACAC GT .. [index] ATCTC .. sequence ...
PCR amplify (cycl)
3'
5'
[index]
5' R1
A AAAAA ... insert ... T
T TTTT ... insert ... A
R3
5' R2
R4
[index]
5'
3'
P7?
P5?
R1 = RC of RNA
R2 = index seq
R4 = RNA seq
R3 = RC of index
Which adapter gets extended/copied on flow cell 1st during bridge amplification?

RNA-seq workflow – where should we deduplicate?



What does a PCR-duplicate look like?

- Same alignment position
  - Chromosome
  - Position
  - Strand (strand specific?)
- How do we determine these?

SAM format - refresher

Chromosome?
Position?
Strand?

- 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
```

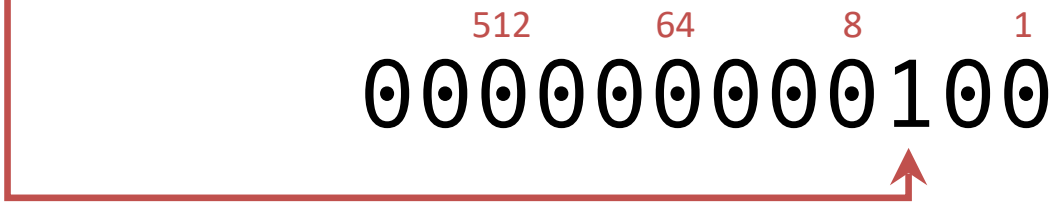
Col	Field	Type	Description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	Reference sequence NAME
4	POS	Int	1-based leftmost mapping POSition
5	MAPQ	Int	MAPping quality
6	CIGAR	String	CIGAR string
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next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

SAM: parsing the bitwise FLAG

https://broadinstitute.github.io/picard/explain-flags.html

```
if ((flag & 4) == 4):
    mapped = False
```

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

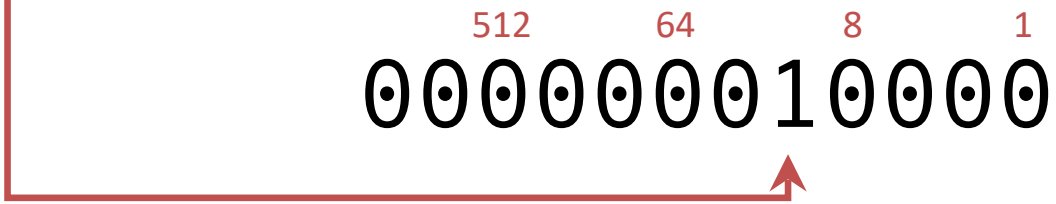


SAM: parsing the bitwise FLAG–which strand?

https://broadinstitute.github.io/picard/explain-flags.html

```
if ((flag & 16) == 16):
    rev_comp = True
```

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment



What does a PCR-duplicate look like?

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

RNAME (SAM col 3)

POS (SAM col 4)

FLAG (SAM col 2)

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?

How do you know if your sequence was soft clipped? 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

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:

```
                  112      115             120             125
Reference: . . . CTTCTATTATCCTT . . .
Read:            AGTCTATTATCCTT
```

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

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: `...ATTGTAGT...GCCCATT...`

  Read: `ATTGT CCATT`

- CIGAR string: 5M1024N5M

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

Example CIGAR strings from real data

1S27M113N38M	13M1185N53M	23M686N43M	40M1I25M
1S36M284N29M	13M83N53M	25M246N41M	43M32325N23M
1S45M301N20M	13M85N53M	25M470N41M	44M5064N22M
1S61M4S	14M903N52M	27M2645N39M	46M2001N20M
1S64M1S	15M470N51M	28M138N38M	49M211N17M
2S31M745N33M	16M1540N50M	28M1794N38M	51M97N15M
2S62M2S	17M1097N49M	28M3349N38M	52M1198N14M
2S63M1S	18M12872N48M	29M2D37M	52M1904N14M
2S64M	19M20545N47M	30M1244N36M	54M12S
3S18M104N45M	20M2979N46M	31M1043N35M	54M407N12M
3S63M	20M456N46M	31M271N35M	55M954N11M
4S62M	20M631N46M	35M113N31M	56M10S
5S61M	21M982N45M	35M128N31M	57M5055N9M
6S42M284N18M	22M103N44M	35M289N31M	58M8S
6S60M	22M138N44M	35M5284N31M	62M4S
7S59M	22M4529N44M	36M103N30M	63M3S
11M4797N55M	23M1290N25M2D18M	36M12071N29M1S	65M1S
11M592N55M	23M15018N43M	36M1496N30M	66M

Why does soft-clipping matter for deduplicating?

- Example 3:

  Reference: . . . CTTCTATTATCCTT . . .

  Read: AGTCTATTATCCTT

- CIGAR string: 2S12M
- Pos: 114

How do we know if these two reads originated from the same molecule of DNA/RNA isolated from our sample?

  Reference: . . . CTTCTATTATCCTT . . .

  Read: CTTCTATTATCCTT

- CIGAR string: 14M
- Pos: 112

What does a PCR-duplicate look like?

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

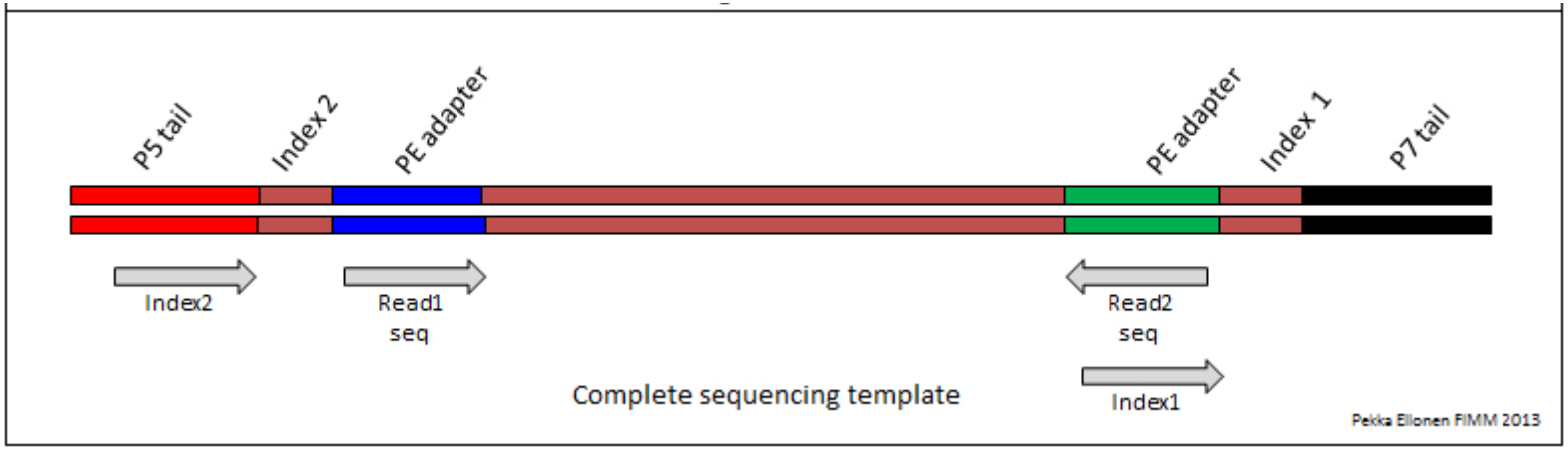
RNAME (SAM col 3)

POS (SAM col 4)

FLAG (SAM col 2)

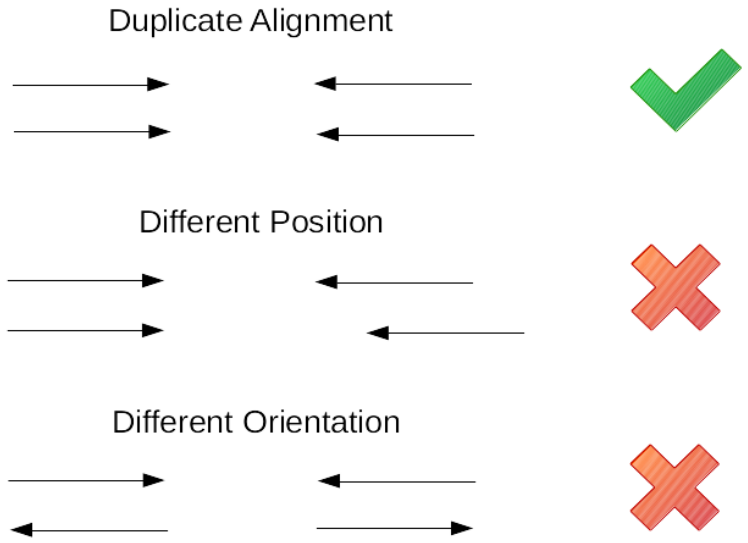
CIGAR (SAM col 6)

QNAME (SAM col 1)



PCR Duplicate Removal Tools: samtools

- Samtools markdup (rmdup deprecated)
  - Same alignment position
    - Chromosome
    - Position
    - Strand (if strand specific)
  - Soft clipping? – “Coordinates are based on the unclipped position of the read against the reference.”



http://www.htslib.org/algorithms/duplicate.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/command-line-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: Dupligänger

- Dupligänger – 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

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)

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?

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
- Single-end reads
- Known UMIs
- Considerations:
  - Millions of reads – avoid loading everything into memory!
  - Be sure to utilize functions appropriately
  - Appropriately commend code and include doc strings
- CHALLENGE: Include options for
  - Single-end vs paired-end
  - Known UMIs vs randomers
  - Choice of duplicate written to file

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

https://github.com/Leslie-C/Deduper