

Mapping

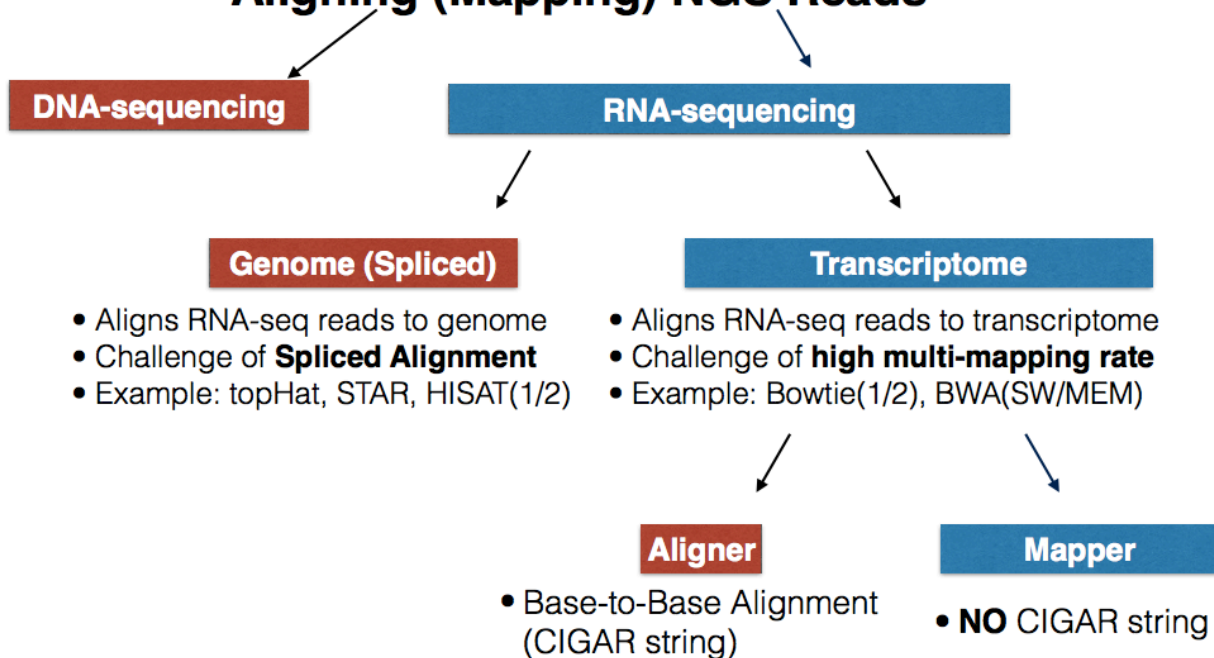
Lecture 12
Sept 28, 2016

ANNOUNCEMENTS

- Article for Friday group

WHICH MAPPER

Aligning (Mapping) NGS Reads



MAPPING

Format specification: <http://samtools.github.io/hts-specs/SAMv1.pdf>

MAPPING - BWT

Alignment Fields

Col1

Col2

Col3

Col4

Col5

Col6

Col7

Col8

Col9

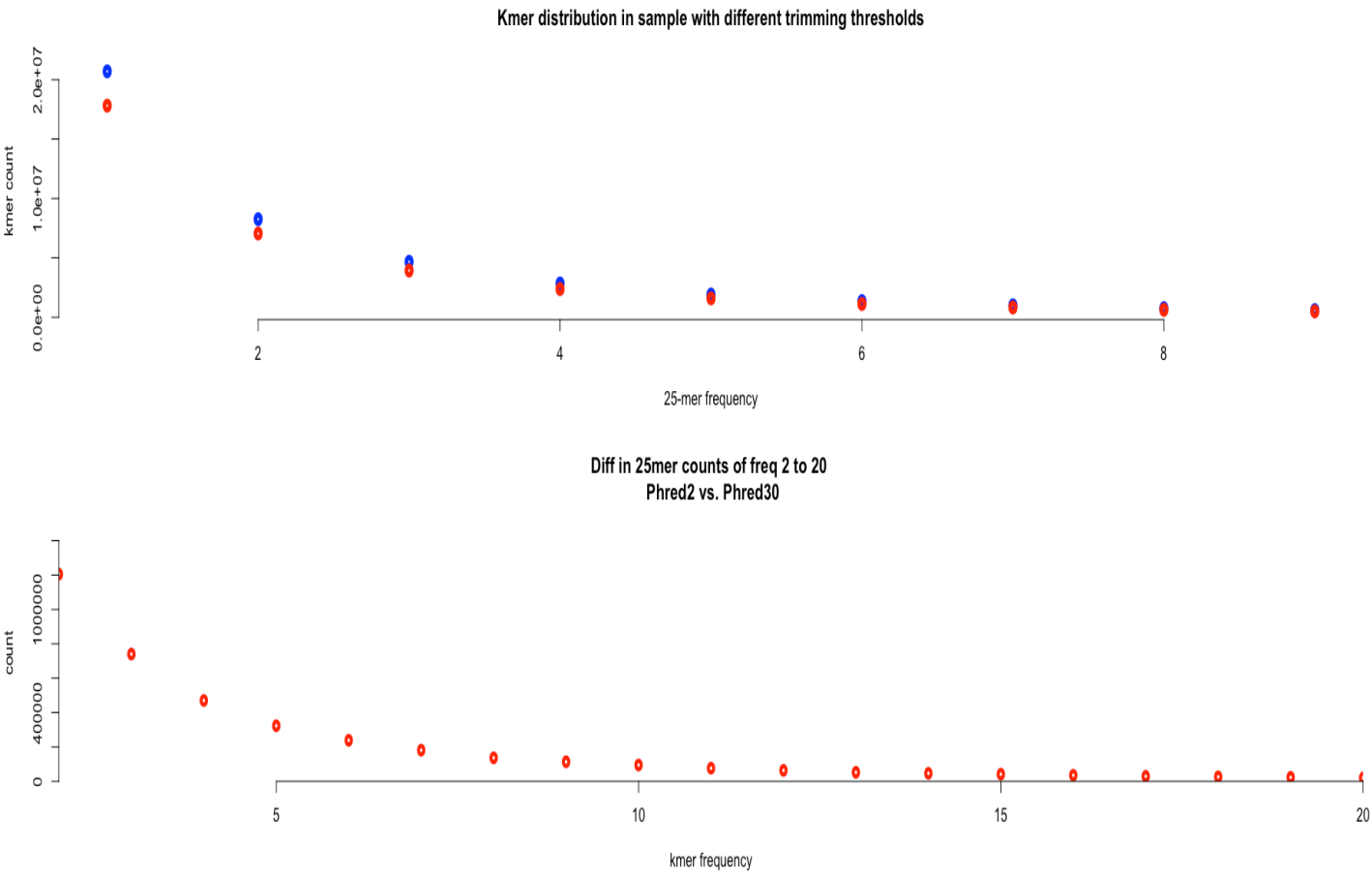
Col10

Col11

MAPPING - BWT

<http://broadinstitute.github.io/picard/explain-flags.html>

BRIEF INTRO TO KMERS



BRIEF INTRO TO KMERS

Read Error Correction

ERROR CORRECTION



ERROR CORRECTION

TATACAATTTGTTTTATGAAAACCTCTAAAAGCAAACATATTTACCAACAATCCTTGCATACGAAATAACCGATTCTATTTAAGCATTGCTCCTATTTTATACAATTTGTTTTATGAAAACCTCTAAAAGCAAACATATTTACCAACAATCCTTGCATACGAAATAACCGATTCTATTTAAGCATTG



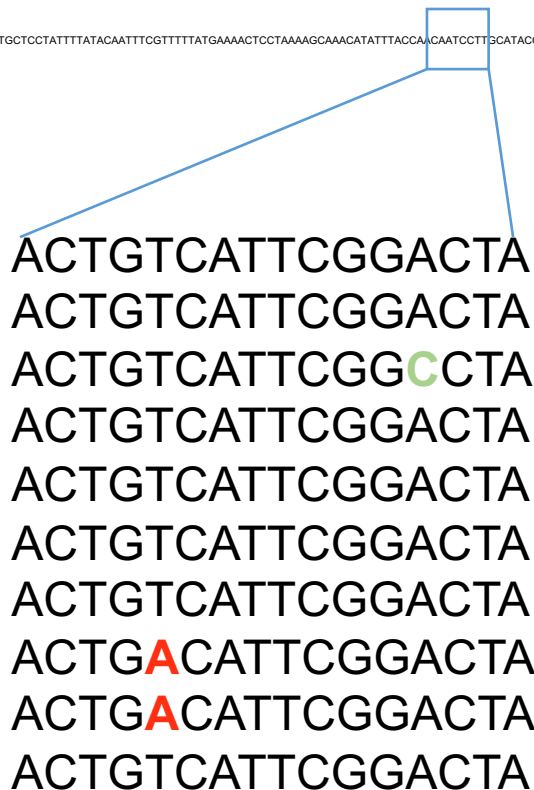
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGCCTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGACATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA

The diagram illustrates a sequence alignment process. A blue box highlights a specific region in the top sequence, and a blue line connects it to a list of ten sequences below. In the third sequence of the list, the letter 'C' is highlighted in green. In the eighth sequence, the letter 'A' is highlighted in red. The sequences are all variations of the consensus sequence 'ACTGTCATTCGGACTA'.

Consensus= ACTGTCATTCGGACTA

ERROR CORRECTION

TATACAATTTCGTTTTATGAAAACCTCCTAAAAAGCAACATATTTACCAACAATCCTTGCATACGAAATAACCGATTCTATTTAAGCATTGCTCCTATTTTATACAATTTCGTTTTATGAAAACCTCCTAAAAAGCAACATATTTACCAACAATCCTTGCATACGAAATAACCGATTCTATTTAAGCATTGCT



Consensus= ACTG{A,T}CATTCTGGACTA

ERROR CORRECTION

TATACAATTTGTTTTATGAAACTCCTAAAGCAAACATATTTACCAACAATCCTTGCATACGAAATAACCGATTCTATTTAAGCATTGCTCCTATTTTATACAATTTGTTTTATGAAACTCCTAAAGCAAACATATTTACCAACAATCCTTGCATACGAAATAACCGATTCTATTTAAGCATTG



ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGCCTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
GCTGATAACCGGACTA
ACTGACATTCGGACTA
ACTGTCATTCGGACTA

The diagram illustrates an error correction process. A blue box highlights a specific region in the top sequence. A blue line connects this box to a corresponding region in the sequence alignment below. In the alignment, the 10th sequence, 'GCTGATAACCGGACTA', contains a red 'G' at the start and a red 'A' at the 8th position, indicating an error. The 9th sequence, 'ACTGACATTCGGACTA', shows the correction where the 'G' is replaced by 'A' at the 10th position.

Consensus= ACTGTCATTCGGACTA

ERROR CORRECTION

Hamming Distance:

http://en.wikipedia.org/wiki/Hamming_distance

ERROR CORRECTION

3 different strategies

ERROR CORRECTION

Kmer-spectra based

ERROR CORRECTION

MSA based

ERROR CORRECTION

Evaluation of Correction