SNPs

alignments

Niklaus Zemp
23 June 2020

Genetic Diversity Centre (GDC)
Bioinformatics
ETH Zurich

Sequence alignment

http://hurlab.med.und.edu/hurlab/cgi-bin/WebTools/SimpleSeqAlign/SimpleSeqAlign.cgi

AATTTCCC
AATATCCC

AATTTCCC
AATTCCC

AATTTCCC
AATTCCCAAT

Sequence alignment

http://hurlab.med.und.edu/hurlab/cgi-bin/WebTools/SimpleSeqAlign/SimpleSeqAlign.cgi

```
AATTTCCC     AATTTCCC     AATTTCCC
AATATCCC     AATT CCC     AATT CCCAAT
```

Sequence alignment - global

http://hurlab.med.und.edu/hurlab/cgi-bin/WebTools/SimpleSeqAlign/SimpleSeqAlign.cgi

AATTTCCC
AATATCCC

```
seq1: 1 AATTTCCC
        |||*||||
seq2: 1 AATATCCC
```

AATTTCCC
AATTCCC

```
seq1: 1 AATTTCCC
        |||| |||
seq2: 1 AATT-CCC
```

```
seq1: 1 AATTTCCC
        ||| ||||
seq2: 1 AAT-TCCC
```

```
seq1: 1 AATTTCCC
        || |||||
seq2: 1 AA-TTCCC
```

AATTTCCC
AATTCCCAAT

```
seq1:   1 AATTTCCC--
          ||||*||*
seq2:   1 AATTCCCAAT
```

```
seq1:   1 AATTTCC-C-
          ||||*|| *
seq2:   1 AATTCCCAAT
```

```
seq1:   1 AATTTCC--C
          ||||*||  *
seq2:   1 AATTCCCAAT
```

Sequence alignment - local

AATTTCCC

AATTCCCAAT

```
seq1: 1 AATTTCCC
        |||| |||
seq2: 1 AATT-CCC

seq1: 1 AATTTCC
        ||||*||
seq2: 1 AATTCCC
```

Alignments

https://www.ndsu.edu/pubweb/~mcclean/plsc411/Blast-explanation-lecture-and-overhead.pdf

Local

Smith-Waterman (algorithm)

Uses a dynamic programming approach

Fast because only small part to work on but works only locally

Global

Needleman-Wunsch (algorithm)

slow because large sequences to align, therefore CPU-”expensive”

By Stefan Zoller

BLAST-local alignment

TruSeqUniversalAdapter

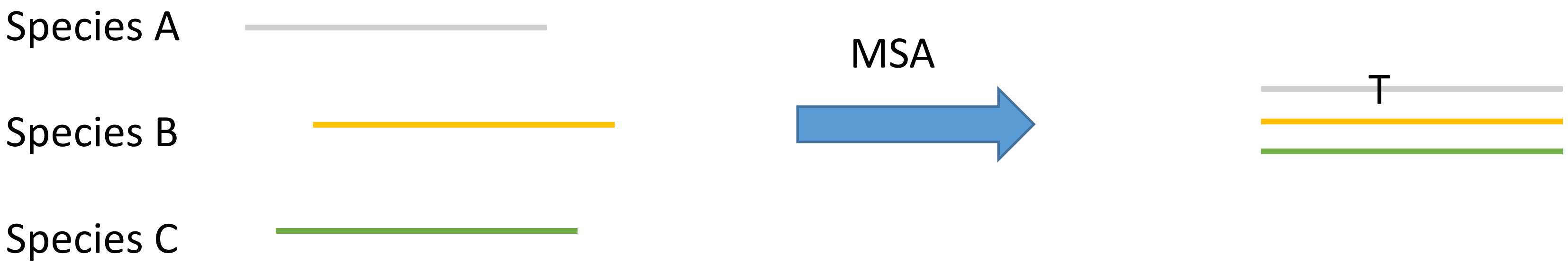
5’

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

Transcript

TTGTTAAAAAATTTTTTTTAAGTTTTTTTCTCTTTTTTTCATTTAAATATATTTTATAAATTTCTATGAA
ATAGTTAACATTGAATAAGCGAATTTAAAAAAAAATGTTCATGATCTTAGATAGACTAATAACGACCTGAT
TATATTCGAGCTGTAGTATTTTTATATTTCACTATTATGTATGAAATTTTTAACATCACAGCCAAGTTAA
TATAACCTCGCTCCAAACCTGAACATTCAAACACTAACTATACTTAAAACGCTAGTTTTGTTAAGTCTAT
CTAAGACCATGATGTAGTTGTATAGCTCGGATCATTTTGAAAATAATAATTGGACTAAACTATAAAAAAA
AAAACATTGGAACATTGTATTATGTAAGTTCATCCAGTTAACTTGGAAAAATTAACTTGGAATGGAAACG
TAAGCTGAACTAAACTTTTCATTCACTTCAAAGCATCCGTATATTCTTGTCGGTGTATGGACTTGTTATG
TAGGATAATTCCATGTTGTGGATTGTTGATTGCGGACAATTGTCGTTTGTTTTAACATGACAATGTTTAT
GACATTTTATTAAACAATCTCTGCATTCGTAACCTTGTTTTCCTAATCTTCGAGCTATGCTTTTACTACA
AACTTGGCACACTGTTCCACCATTTAAGTGCTTGGCAATAAATGTATGATCATTAAAAATGTGCAATTTT
GTGCCTTTTTTACGCCATCTTGATTTTTGTGCTAATGATAATGGTACCAAATAATGTTTTTTAATACCAT
TTTCAAGTGTTTCAAGTACTAATGTGCTTGCTTCATTTAAGTATGTTCGAGTTGAAGCACCGCTTATACC

Multiple sequence alignments-global alignment



Species A

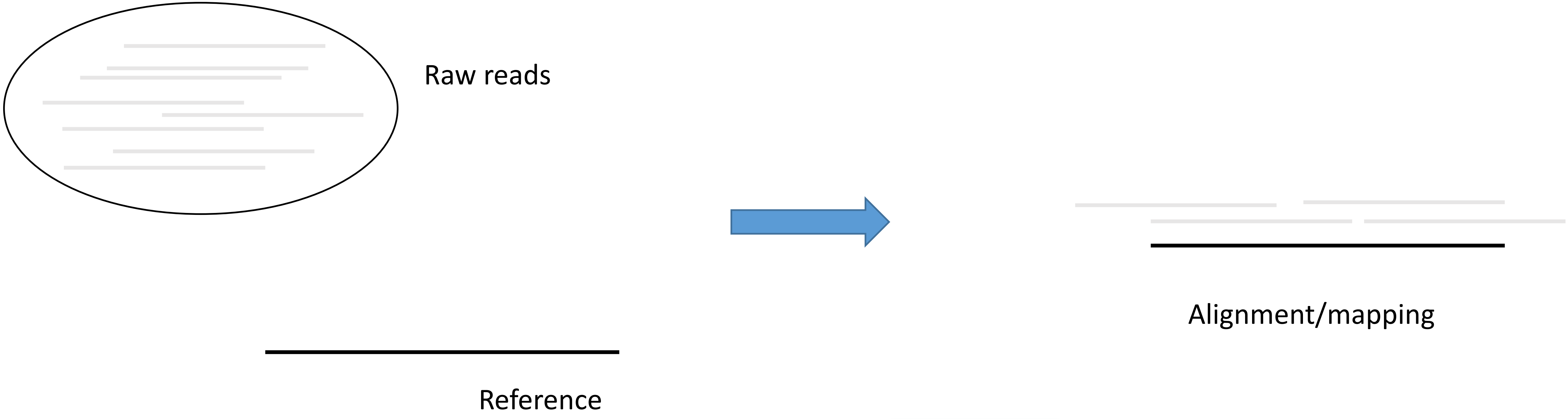
Species B

Species C

MSA

T

Mapping



Raw reads

Reference

Alignment/mapping

Integrative Genomics Viewer

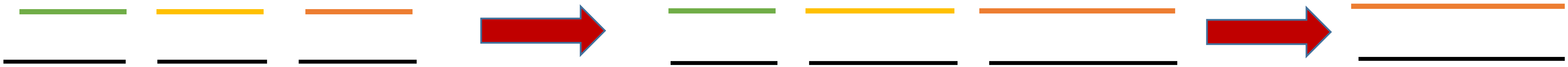
Mappers

Problem

The fast and exact algorithms for local alignments do not scale to large genomes. Do not handle high sequence errors well.

New approaches needed Solution

First apply very fast algorithms that match short local regions exactly. Then extend the short regions to larger regions.



By Stefan Zoller

Global mappers

k-mer based alignment -> RNA-seq

can be fast and quite accurate AGCTTTAGAC ->3-mers: AGC, GCT, CTT, TTT, TTA, TAG, AGA, GAC

when k-mers are redundant, i.e. appear often in sequences/genome

suffix-tree

a tree-like structure that contains all suffixes of the sequences (genome).

Subsequences (reads) can be looked-up very quickly.

needs a lot of memory

compressed suffix-tree

a compressed form, e.g. Burrows-Wheeler **transform** very fast, very memory efficient.

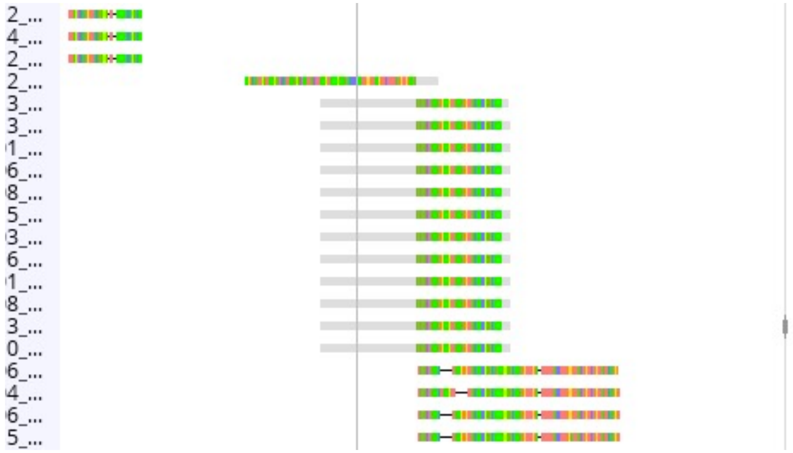
gets rather slow and inaccurate with high sequence error rates or long reads

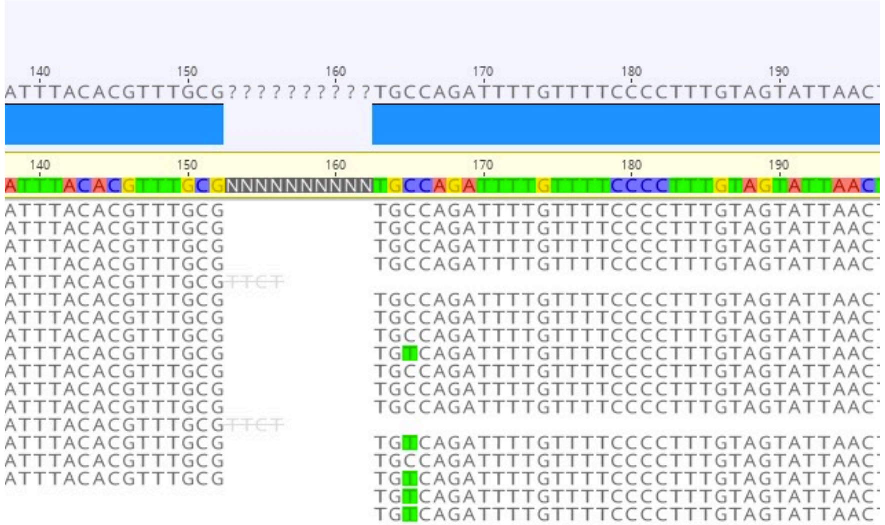
MEM-mapping

maximal exact match

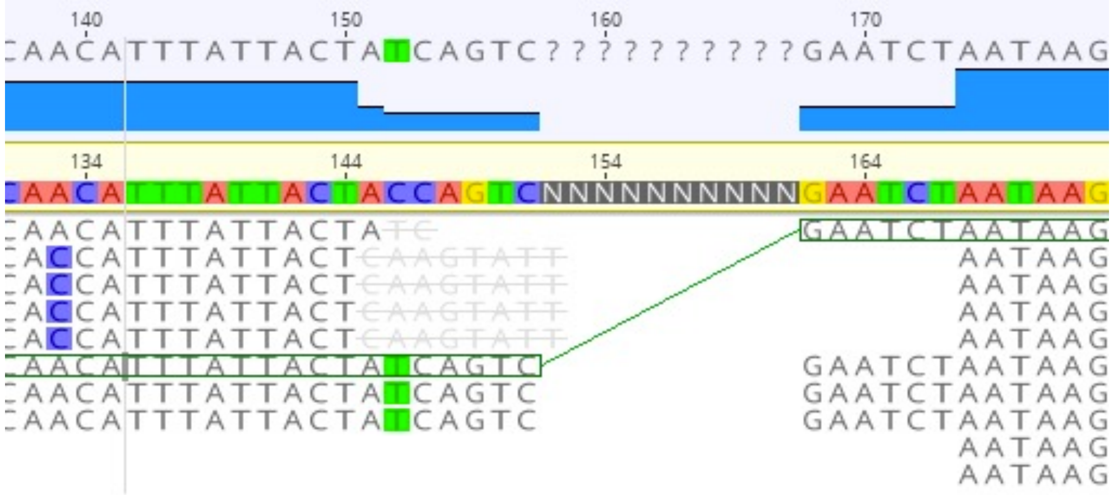
cannot be extended

Soft clipping during read mapping (bwa)





140 150 160 170 180 190
ATTTACACGTTTGCG?????????TGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCGNNNNNNNNNNTGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG TGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG TGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG TGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG TGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG~~TTCT~~
ATTTACACGTTTGCG TGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG TGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG TGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG TGTCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG TGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG TGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG TGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG~~TTCT~~
ATTTACACGTTTGCG TGTCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG TGCCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
ATTTACACGTTTGCG TGTCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
TGTCAGATTTTGTTTTCCCCTTTGTAGTATTAAC
TGTCAGATTTTGTTTTCCCCTTTGTAGTATTAAC



140 150 160 170
CAACATTTATTACTATCAGTC??????????GAATCTAATAAG
134 144 154 164
CAACATTTATTACTACCAGTCNNNNNNNNNNGAATCTAATAAG
CAACATTTATTACTA~~TC~~ GAATCTAATAAG
CACCATTTATTACT~~CAAGTATT~~ AATAAG
CACCATTTATTACT~~CAAGTATT~~ AATAAG
CACCATTTATTACT~~CAAGTATT~~ AATAAG
CACCATTTATTACT~~CAAGTATT~~ AATAAG
CAACATTTATTACTATCAGTC GAATCTAATAAG
CAACATTTATTACTATCAGTC GAATCTAATAAG
CAACATTTATTACTATCAGTC GAATCTAATAAG
AATAAG
AATAAG

Sam/bam format

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

```
@HD            VN:1.3          SO:coordinate
@SQ            SN:scaffold3_8_ref0000002_ref0000084            LN:96419
@RG            ID:ERR3418961   SM:ERR3418961  PL:Illumina
@PG            ID:bwa          PN:bwa          VN:0.7.12-r1039  CL:bwa mem -R @RG\tID:ERR3418961\tSM:ERR3418961\tPL:Illumina Ref_C_bombi.fasta ERR3418961_1.fq.gz ERR3418961_2.fq.gz
ERR3418961.5587914     163             scaffold3_8_ref0000002_ref0000084       1238        60          151M        =           1565        478
               GTGGGGAGATAAAACCGAACTGGTGTGTTAACATCAACGTCTGTCCCTTCACGGTGGGGACCGCGTGCACGACGCGCATCATCCGTCTGCGTGGTGGGGGGTGGTAGTGTAGAATTTCGTTGGTGCTCCTGCCGTCGCAGGACC
GCCTCCA        AA7FFFAJFFJJJJJJFJJJJFJFJJJAJFJFJJJJJFJJJF7FJJJFJFJJJJJJJJJ<FJ7F<FJFJJJJJJ<JJJFAJFJJJJJFJJJJJJJJJJJFF-AJJJJJJJJJJJ-FFJJJJFFJFJJJFFJJ<-A<7FF-<AAF<<7<FF)            NM:i:0      MD:Z:151
               AS:i:151        XS:i:0          RG:Z:ERR3418961

ERR3418961.5587914     83              scaffold3_8_ref0000002_ref0000084       1565        60          151M        =           1238        -478
               CCGACGTCCAGGCGGACGCTCGAGCTCGTGACGGGGAGGTAATCGATGTAGGGCACGACAGTGGTGCCGGAGAAGCCGACGTCTACGACGATCGCAGTCCCACCCGCCTCACACGTCAATTCTCTGTCAATAACATTAGTACCA
ACACTAG        JJJFJJJFJFJJJJJJJJFFJJAFJFAJFFFFJFFJJJJJJJJJJJJJFJJJJFJJJJAFJJJJJJJJJJJJFFJJJFJJJJJJFJJFJJFJJJJJJFFFJJJJJJFFFJJJFJFJFJJFFJF<FJJJJJJJFJJJJJJJJJJJJJFFFAA      NM:i:0      MD:Z:151    AS:i:151    XS:i:0
               RG:Z:ERR3418961

ERR3418961.19317       185             scaffold3_8_ref0000002_ref0000084       4042        16          101S36M14S  =           4042
               0GCCGTGCATACGCGCAGCGCTCCACGTGACTGCAATTCGTCACATGCTCACTAGTACGTATTATCTCGCTGAACTGCGCTGGCGCTATATGTATATATATATATATATATATACATATATATATATATATATATGCAAAGTGCAC
GCGT           )JJJJFJF7<FJJJJJJJJJJJJJJJJFAFFJAJJJF7JJJJJJJJFAFFJFFAJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA      NM:i:1      MD:Z:12T23  AS:i:31
               XS:i:27         RG:Z:ERR3418961                 XA:Z:scaffold3_8_ref0000002_ref0000084,-56115,97S27M27S,0;scaffold3_8_ref0000002_ref0000084,+51496,34S26M91S,0;
```

Read columns: 1) read name; 2) flag(binary); 3) contig/chromosome; 4) position; 5) mapping quality; 6) CIGAR string;7) RNEXT; 8) Position of next read in alignment (pair); 9) observed template length; 10) sequence; 11) quality per base; 12) optional information

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

Take home massage

- (global) alignments are computational intensive
- Mappers are faster but are less precise
- Mappings can be full of noise