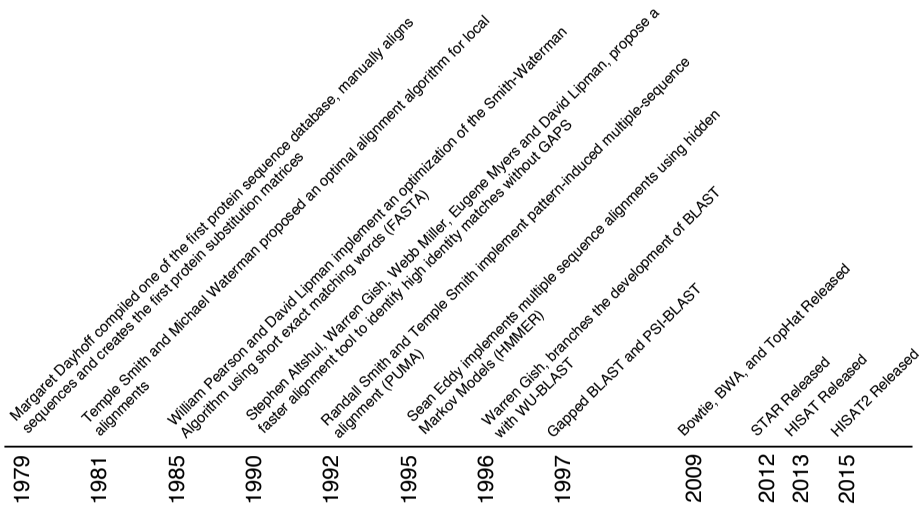
Short Read Alignment

Mapping Reads to a Reference

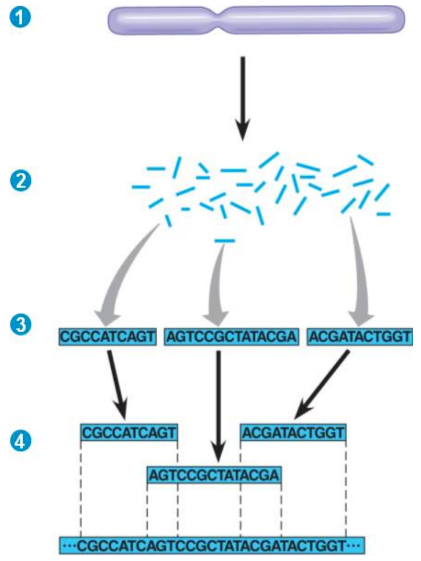
Brandi Cantarel, Ph.D. & Daehwan Kim, Ph.D.
BICF
05/2019

- Introduction to Mapping
- Short Read Aligners
- DNA vs RNA
- Alignment Quality
- Pitfalls and Improvements

History of Sequence Similarity



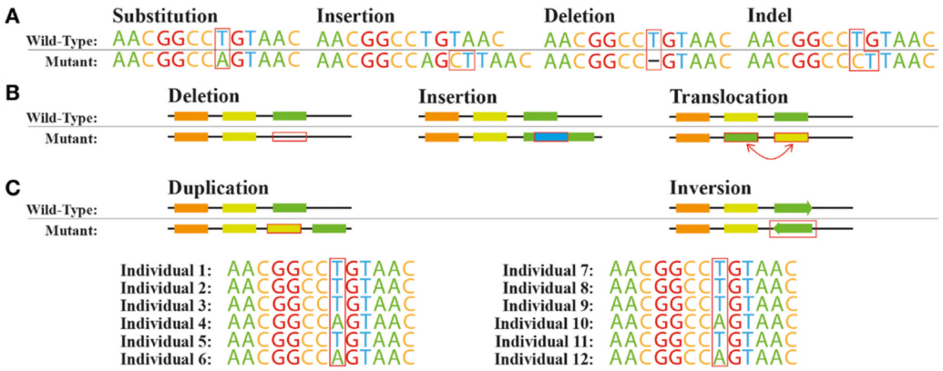
Whole Genome Shotgun



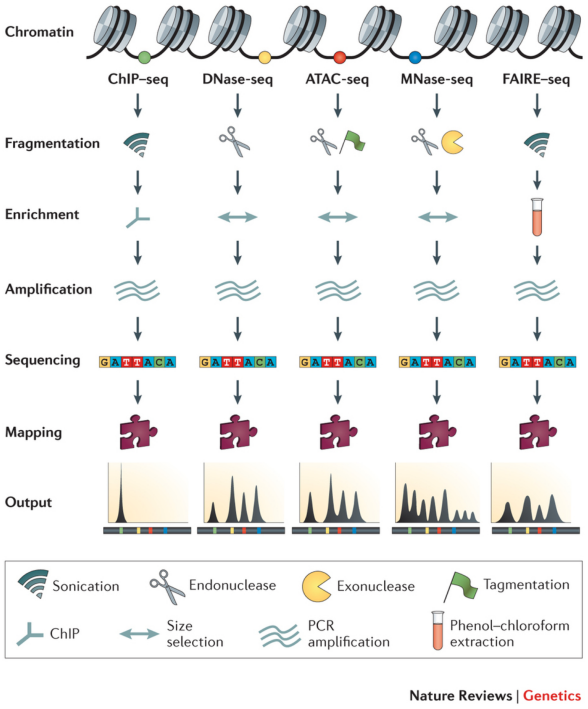
Utility of Mapping

- DNASeq
  - Identify Variation
- RNASeq
  - Estimate the abundance of transcripts and genes
- ChromatinSeq
  - Determine the structure of DNA (open, close, bound to proteins, etc)

Genetic Variation



Cardoso JG, Andersen MR, Herrgård MJ, Sonnenschein N. Analysis of genetic variation and potential applications in genome-scale metabolic modeling. Front Bioeng Biotechnol. 2015 Feb 16;3:13. doi: 10.3389/fbioe.2015.00013. eCollection 2015. Review. PubMed PMID: 25763369; PubMed Central PMCID: PMC4329917.



Nature Reviews | Genetics

Chromatin Seq

How to Make an Alignment

```
PMILGYWNVRGL      PMILGYWNVRGL      PM-ILGYWNVRGL
:  .              :  .  :. :::      :  :.  :. :::
PPYTIVYFPVRG      PPYTIVYFPVRG      PPYTIV-YFPVRG
```

```
  PMILGYWNVRGL
P:
P:
Y  .  :.
T        ..
I  :.     .  .
V  ...    :  .
Y  .   :.
F  ...   ..    .
P:
V  ...    :  .
R            :
G             :
```

Local Alignment

```
AAPMILGYWNVRGLBB
  :  .  :. :::
DDPPYTIVYFPVRGCC
```

Global Alignments

Global Alignment

```
-PMILGYWNVRGL
 :  .  :: ::
PPYTIVYFPVRG-
```

Basis:

$F_{0j} = d * j$

$F_{i0} = d * i$

Recursion, based on the principle of optimality:

$$F_{ij} = \max(F_{i-1,j-1} + S(A_i, B_j),\ F_{i,j-1} + d,\ F_{i-1,j} + d)$$

The pseudo-code for the algorithm to compute the F matrix therefore looks like this:

```
for i=0 to length(A)
  F(i,0) ← d*i
for j=0 to length(B)
  F(0,j) ← d*j
for i=1 to length(A)
  for j=1 to length(B)
  {
    Match ← F(i-1,j-1) + S(A_i, B_j)
    Delete ← F(i-1, j) + d
    Insert ← F(i, j-1) + d
    F(i,j) ← max(Match, Insert, Delete)
  }
```

Local Alignments

Local Alignment

```
 AAPMILGYWNVRGLBB
   :  .  :: ::
DDPPYTIVYFPVRGCC
```

A matrix H is built as follows:

$H(i,0) = 0,\ 0 \le i \le m$

$H(0,j) = 0,\ 0 \le j \le n$

if $a_i = b_j$ then $w(a_i,b_j) = w(\text{match})$ or if $a_i != b_j$ then $w(a_i,b_j) = w(\text{mismatch})$

$$H(i,j) = \max \left\{ \begin{array}{cc} 0 & \\ H(i-1,j-1) + w(a_i,b_j) & \text{Match/Mismatch} \\ H(i-1,j) + w(a_i,-) & \text{Deletion} \\ H(i,j-1) + w(-,b_j) & \text{Insertion} \end{array} \right\},\ 1 \le i \le m, 1 \le j \le n$$

Where:

- a,b = Strings over the Alphabet Σ
- m = length(a)
- n = length(b)
- $H(i,j)$ - is the maximum Similarity-Score between a suffix of a[1...i] and a suffix of b[1...j]
- $w(c,d),\ c,d \in \Sigma \cup \{'-'\}$, '-' is the gap-scoring scheme

Short Read Aligners

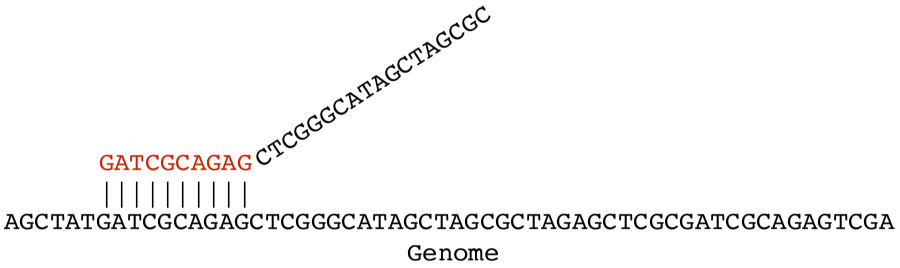
- Introduction to Mapping
- Short Read Aligners
- DNA vs RNA
- Alignment Quality
- Pitfalls and Improvements

- Short read aligners assume that the read came from “intact” from the reference
- So the alignment is “global” from the read perspective and “local” from the reference perspective

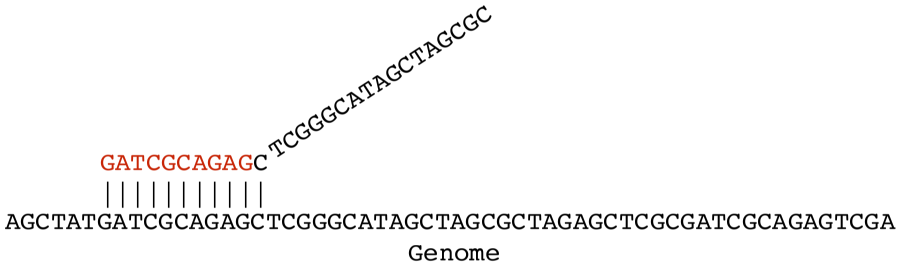
Short Read Aligners

Short Read Aligners

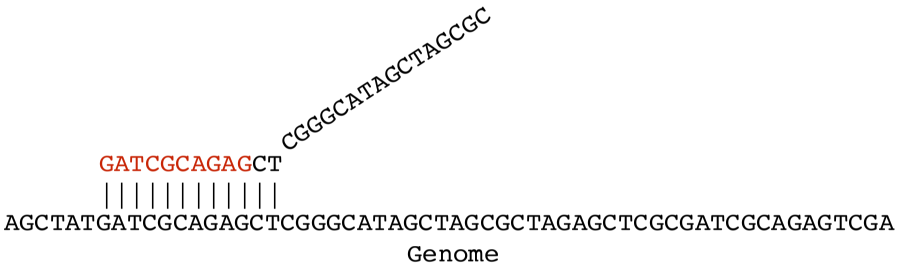
Short Read Aligners



Short Read Aligners



Short Read Aligners

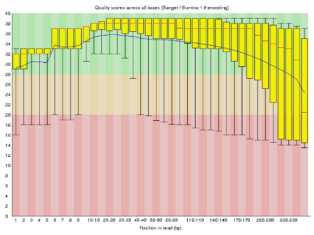


SRA Features: Seeding

- Seeding represents the first few tens of base pairs of a read. The seed part of a read is expected to contain less erroneous characters due to the specifics of the NGS technologies. Therefore, the seeding property is mostly used to maximize performance and accuracy. The alignments are then extended from the seed.

SRA Features: Base Quality

- Base quality scores provide a measure on correctness of each base in the read. The base quality score is assigned by a phred-like algorithm. The score Q is equal to -10 log10(e), where e is the probability that the base is wrong. Some tools use the quality scores to decide mismatch locations. Others accept or reject the read based on the sum of the quality scores at mismatch positions.

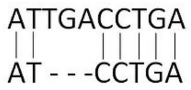


Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

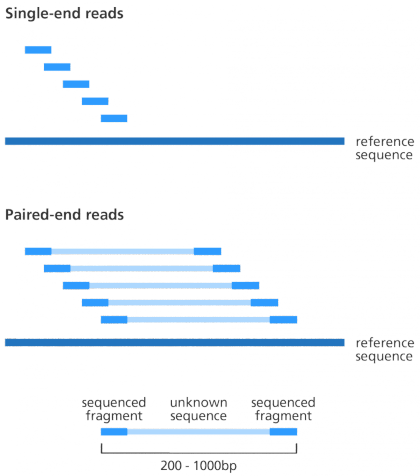
SRA Features: Gaps

- Existence of indels necessitates inserting or deleting nucleotides while mapping a sequence to a reference genome (gaps). The complexity of choosing a gap location increases with the read length. Therefore, some tools do not allow any gaps while others limit their locations and numbers.



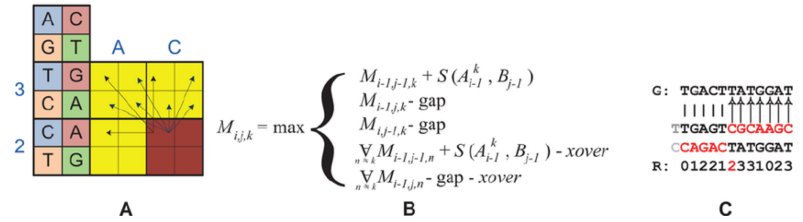
SRA Features: Paired-End

- Paired-end reads result from sequencing both ends of a DNA molecule. Mapping paired-end reads increases the confidence in the mapping locations due to having an estimation of the distance between the two ends.



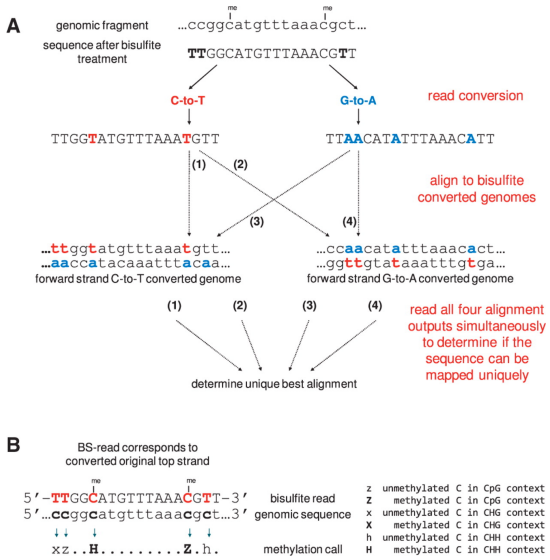
SRA Features: Color-Space

- Color space read is a read type generated by SOLiD sequencers. In this technology, overlapping pairs of letters are read and given a number (color) out of four numbers [17]. The reads can be converted into bases, however, performing the mapping in the color space has advantages in terms of error detection.



SRA Features: Bisulphite

- Bisulphite treatment is a method used for the study of the methylation state of the DNA [3]. In bisulphite treated reads, each unmethylated cytosine is converted to uracil. Therefore, they require special handling in order not to misalign the reads.



Short Read Aligners

Table 1 Features supported by the tools

	Bowtie	Bowtie2	BWA	SOAP2	MAQ	RMAP	GSNAP	FANGS	Novoalign	mrFAST	mrsFAST
Seed mm.	Up to 3		Any	Up to 2	Any	Any					
Non-seed mm.	QS	AS	Count	Count	QS	Count	Count	Count	QS	Count	Count
Var. seed len.	> 5		Any	> 28							
Mapping qual.		Yes	Yes		Yes				Yes		
Gapped align.		Yes	Yes	PE	PE		Yes	Yes	Yes	Yes	
Colorspace	Yes		Yes		Yes				Yes		
Splicing							Yes				
SNP tolerance							Yes				
Bisulphite reads						Yes	Yes		Yes	Yes	

PE: paired-end only, mm.: mismatches, QS: base quality score, count: total count of mismatches in the read, AS: alignment score, and empty cells mean not supported.