Short Read Alignment (NGS)

Matthias Haimel
Fastq → Quality control → Fastq → Alignment → BAM → Visualization → BAM → Variant calling → VCF → Variant annotation → VCF → Variant association
Short Read Alignment (NGS)

Data → Tool → Output
Short Read Alignment (NGS)

- Reads (FASTQ)
Short Read Alignment (NGS)

- Reads (FASTQ)

```
@ERR001268.10000142/1
CTATCCTAATCCCAAGAACACATGAAGATGTGACCT
+
IIIIIIICIIIIDIB@5=30;?0?F/37257029.*F
@ERR001268.10000296/1
TCGGAGAATTCACTGGAGAGAAACCTTACAAATG
+
8F@BII.IICII:4<?EB44670:9)1712.5*000
@ERR001268.10000504/1
CCTTGCTTTTCAGTTGTCGCCACCTTTCCAGACCAAAC
+
=IIIIIIIIIIIIIIF?73809?946*/+03,+/6
```
Short Read Alignment (NGS)

- Reads (FASTQ)

Data

Fragments

Reads

Single-end

Paired-end
Short Read Alignment (NGS)

- Reference

Data

Single-end

Paired-end

Reference

Reads
Short Read Alignment (NGS)

- Human Reference
- GRCh37 <-> GRCh38

Data

Reference

Reads

Single-end

Paired-end
Reference Genome

  - New coordinate system
  - Alternate sequence representation
  - Centromeres and heterochromatin models added
  - > 8000 sequencing errors corrected
- **Primary assembly**
  - Chromosomes + unlocalised & unplaced contigs + MT
- **Full assembly**
  - Primary assembly + alternate loci
- **Decoy contigs**
  - 2,385 contigs identified by Heng Li
Short Read Alignment (NGS)
Short Read Alignment (NGS)

- Global alignment
  - Attempt to align every base
  - Good for equal size sequence

Global

CT-----AT-TTACT-----AT
CTGGCTATGTTACTATGCAT

Local

-----CTAT-TTACTAT-----

- Local alignment
  - Contain a region of target sequence
  - Good for unequal sequence length
Short Read Alignment (NGS)

- Local alignment
- Fast
- Sensitive
- Specific

Reference

Reads  _____  _____  _____

University of Cambridge
Short Read Alignment (NGS)

- Align each read to the reference sequence
- Reads are not perfect!!!

Reference:

```
TTTCCCTGAGTTACACTGAAGATGGTCTAATTTCAAA
```

Reads:

```
CCCTGAGTTACACTGAAGATG**ATCT**
AGTTACACTGAAGATGGTCTAA
GTTAG**ACTGAAGATGGTCTAATTT**
CACTGAAGATG**ATCTAATTT**CA
```
WGS example calculation

- Human genome:
  - $3.2 \times 10^9$

- Phred Quality Score:
  - 40 (1 in 10,000 probably incorrect)

- Probably incorrect bases:
  - $3.2 \times 10^5 = 320,000$

- 40x sequence coverage (average):
  - 12,800,000 probably incorrect bases
Short Read Alignment (NGS)

- Reference missing?
  - *de novo* assembly
- Mismatches
  - Species polymorphism
  - Sequencing error

Tools

```
TTTCCCTGAGTTACACTGAAGATGGTCTAATTTCAAA
- Transitions
- Transversions
```

**Ti/Tv ratio**
- random: 0.5
- whole genome: ~2
- whole exome: ~3

Reference Reads

```
TTTCCCTGAGTTACACTGAAGATGGTCTAATTTCAAA
```

```
CCCTGAGTTACACTGAAGATGATCT
```

```
AGTTACACTGAAGATGGTCTAA
```

```
GTTAGACTGAAGATGGTCTAATT
```

```
TGATCTAATTTCAATG
```

Short Read Alignment (NGS)

- Reference missing?
  - *de novo* assembly
- Mismatches
  - Species polymorphism
  - Sequencing error
- Type of data
  - Whole Genome / Exome
  - RNA-Seq
  - ChIP-Seq
  - ...
Short Read Alignment (NGS)

Data → Tool → Output
Short Read Alignment (NGS)

- **SAM** *(Sequence Alignment / Map format)*
- **BAM** *(Binary version of SAM)*
- Well defined specification
- Large tool collection for reading / writing
  - samtools: basic utility tool
  - Picard: more comprehensive utility tool
SAM format

- Header section
- Alignment section

```
@HD VN:1.5 SQ:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+5S6M,30,1;
r001 83 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```
Sequence Alignment / Map format

- Header stores extra information about the alignment
  - E.g. Reference, Program, Read groups

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
```

- Start with ‘@’ followed by a two-letter code
### Sequence Alignment / Map format

#### Alignment

```
1001  83  ref 37 30 9M  =  7 -39  CAGCGGCGAT  *  NM:i:1
```

<table>
<thead>
<tr>
<th>Col</th>
<th>Field</th>
<th>Type</th>
<th>Regexp/Range</th>
<th>Brief description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>QNAME</td>
<td>String</td>
<td>[!-?A-^-]{1,255}</td>
<td>Query template NAME</td>
</tr>
<tr>
<td>2</td>
<td>FLAG</td>
<td>Int</td>
<td>[0, 2^16-1]</td>
<td>bitwise FLAG</td>
</tr>
<tr>
<td>3</td>
<td>RNAME</td>
<td>String</td>
<td>*[!-()]+-&lt;--&gt;^ ![~]*</td>
<td>Reference sequence NAME</td>
</tr>
<tr>
<td>4</td>
<td>POS</td>
<td>Int</td>
<td>[0, 2^31-1]</td>
<td>1-based leftmost mapping POSition</td>
</tr>
<tr>
<td>5</td>
<td>MAPQ</td>
<td>Int</td>
<td>[0, 2^8-1]</td>
<td>MAPping Quality</td>
</tr>
<tr>
<td>6</td>
<td>CIGAR</td>
<td>String</td>
<td>**((0-9)[0-9]+[MIDNSHPX=]+)</td>
<td>CIGAR string</td>
</tr>
<tr>
<td>7</td>
<td>RNEXT</td>
<td>String</td>
<td>**=~-[!-()]+-&lt;--&gt;^ ![~]*</td>
<td>Ref. name of the mate/next read</td>
</tr>
<tr>
<td>8</td>
<td>PNEXT</td>
<td>Int</td>
<td>[0, 2^31-1]</td>
<td>Position of the mate/next read</td>
</tr>
<tr>
<td>9</td>
<td>TLEN</td>
<td>Int</td>
<td>[-2^31+1, 2^31-1]</td>
<td>observed Template LENgth</td>
</tr>
<tr>
<td>10</td>
<td>SEQ</td>
<td>String</td>
<td>**([A-Za-z=.]+)</td>
<td>segment SEQuence</td>
</tr>
<tr>
<td>11</td>
<td>QUAL</td>
<td>String</td>
<td>![~]+</td>
<td>ASCII of Phred-scaled base QUALity+33</td>
</tr>
</tbody>
</table>
Short Read Alignment (NGS)
Hands-on