INTRODUCTION TO READ MAPPING

Dhivya Arasappan
(With several slides borrowed from Dr. Jeff Barrick)
What does an alignment look like?

Ref=TAGATCAGATTTCGATACCAGACCATGATCATAACGATCC

Read=AGACCATG

Found at offset 18!
TAGATCAGATTTCGATACCAGACCATGATCATAACGATCC
What does an alignment look like?

https://vlsci.github.io/slcc_docs/tutorials/rna_seq_dge_basic/rna_seq_basic_tutorial/
What is an alignment look like?

Type in a genomic interval and press Go to view the region.

gatkforums.broadinstitute.org
Why is alignment a difficult problem?

- 100’s of millions of reads
- Billions of bases to search through
- Approximate matching
- Looking for a tiny pattern (~100-120 bp read) in a large, often redundant sequence.
Basic steps of mapping reads

- Pre-mapping QC
- Build a reference sequence index.
- Map sequencing reads to the reference index.
- Convert results to SAM/BAM format and obtain mapping statistics.
- Post-mapping analysis.
What to know about your data before mapping?

KNOW YOUR DATA!
• Paired end? Single end?
• Traditional RNA-Seq? 3’ tag?
• Insert size estimate?

PREPROCESSING
• Adaptor sequences trimmed?
• Primer sequences/barcodes removed?
• Poor quality regions trimmed?
What will your reads look like?

FASTQ FORMAT

@HWI-EAS216_91209:1:2:454:192#0/1
GTTGATGAATTTCTCCAGCGCGAATTTGTGGGCT
+HWI-EAS216_91209:1:2:454:192#0/1
B@BBBBBB@BBBBAAAA>@AABA?BBBAAB??>A?

Line 1: @read name
Line 2: called base sequence
Line 3: +read name (optional after +)
Line 4: base quality scores
What does the reference look like?

- **Genome**: All the DNA of an individual, organized by chromosome, containing non-coding and coding regions.

- **Transcriptome**: All the gene isoforms. No non-coding sequences.
What to know about your reference before mapping?

- Mapping to genome vs transcriptome?
- Is your reference the right version?
- Does your annotation match your reference?
What will your reference look like?

• FASTA Format

>gi|254160123|ref|NC_012967.1| Escherichia coli B str. REL606
agcttttctattcgtgactgcaacgggcaatatgtctctgtgtggattaaaaaaaaagagtgct
tgatagcagctttctgaactgggtacctgcgggtgagtaaattaatattttattgactttagg
tcactaaatatattaaccaaatataggcatagcgacagcacagataaaaaattacagagtac
acaacatatcatgaaacgcattagcaccaccattaccaccaccatcaccattaccaccaggt ....

• Using complex reference sequence names is a common problem during analysis. Might rename:

>REL606
agcttttctattcgtgactgcaacgggcaatatgtctctgtgtggattaaaaaaaaagagtgct
tgatagcagctttctgaactgggtacctgcgggtgagtaaattaatattttattgactttagg
What will your annotation look like?

• GFF3 Format
  – seqname - The name of the sequence.
  – source - The program that generated this feature.
  – feature - Examples: "CDS", "start_codon", "stop_codon", and "exon".
  – start - The starting position of the feature in the sequence.
  – end - The ending position of the feature (inclusive).
  – score - A score between 0 and 1000.
  – strand - Valid entries include '+', '-', or '. (for don't know/don't care).
  – Frame - reading frame
  – group - ID and other information about the entry

Example:
  Rel606 refseq cds 1450 1540 500 + . Gene_id=« test_gene »

• Make sure the GFF3 file matches your reference fasta file.
Where to get your references?

- Ensembl ftp
- UCSC
- Gencode
- Organism specific databases/websites.
First Step: Reference Indexing

- **Indexing**: Think of the index of the book.

- Break reference into substrings of K length (Kmers) and index all locations of the Kmer.

<table>
<thead>
<tr>
<th>Reference: TTACTTTTACG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACG</td>
</tr>
<tr>
<td>ACT</td>
</tr>
<tr>
<td>CTT</td>
</tr>
<tr>
<td>TAC</td>
</tr>
<tr>
<td>TTA</td>
</tr>
<tr>
<td>TTT</td>
</tr>
</tbody>
</table>

**Read:** CTT

Find the prefix and extend the alignment
Indexing

- Many different ways to represent reference indexes

**hash table**

<table>
<thead>
<tr>
<th>keys</th>
<th>hash function</th>
<th>buckets</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lisa Smith</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sandra Dee</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**suffix array**

| 1 | attcatg$  |
| 2 | ttcatg$   |
| 3 | tcag$     |
| 4 | catg$     |
| 5 | atg$      |
| 6 | tg$       |
| 7 | g$        |
| 8 | $         |

sort the suffixes alphabetically

the indices just "come along for the ride"

**suffix tree**

- image from wikipedia
- image from discuss.codechef.com
- image from wikipedia
# Types of Mappers

<table>
<thead>
<tr>
<th>Class</th>
<th>Category</th>
<th>Package</th>
<th>Notes</th>
<th>Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read mapping</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unspliced aligners*</td>
<td>Seed methods</td>
<td>Short-read mapping package</td>
<td>Smith-Waterman extension</td>
<td>Aligning reads to a reference transcriptome</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(SHRiMP)\textsuperscript{41}</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stampy\textsuperscript{39}</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bowtie\textsuperscript{43}</td>
<td>Probabilistic model</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BWA\textsuperscript{44}</td>
<td>Incorporates quality scores</td>
<td></td>
</tr>
<tr>
<td>Burrows-Wheeler</td>
<td>transform methods</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spliced aligners</td>
<td>Exon-first methods</td>
<td>MapSplice\textsuperscript{52}</td>
<td>Works with multiple unspliced aligners</td>
<td>Aligning reads to a reference genome. Allows for the identification of novel splice junctions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SpliceMap\textsuperscript{50}</td>
<td>Uses Bowtie alignments</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TopHat\textsuperscript{51}</td>
<td>Can use SNP databases</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Seed-extend methods</td>
<td>GSNAP\textsuperscript{53}</td>
<td>Smith-Waterman for large gaps</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>QPALMA\textsuperscript{54}</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Garber et al, Nature Methods, 2011*
## Unspliced Mapping

![Diagram of Unspliced Mapping](image)

### Mapping to genome

<table>
<thead>
<tr>
<th>Class</th>
<th>Category</th>
<th>Package</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read mapping</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unspliced aligners(^a)</td>
<td>Seed methods</td>
<td>Short-read mapping package</td>
<td>Smith-Waterman extension (SHRiMP)(^{41})</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stampy(^{39})</td>
<td>Probabilistic model</td>
</tr>
<tr>
<td>Burrows-Wheeler</td>
<td>Transform methods</td>
<td>Bowtie(^{43})</td>
<td>Incorporates quality scores</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BWA(^{44})</td>
<td></td>
</tr>
</tbody>
</table>

---

Figure: Garber et al, Nature Methods, 2011
**Burrows-Wheeler transform** compresses sequence.

<table>
<thead>
<tr>
<th>Input</th>
<th>SIX.MIXED.PIXIES.SIFT.SIXTY.PIXIE.DUST.BOXES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>TExyDST.E.IIXIXXSSMPPS.B..E.S.EUSFXDII0IIIT</td>
</tr>
</tbody>
</table>

**Suffix tree** enables fast lookup of subsequences.


Exact matches at all positions below a node.

Spliced mapping

• Needed for quantifying and identifying splice variants from RNA Seq data.

• Tools:
  – HISAT2
  – STAR
  – Tophat
  – SpliceMap
  – MapSplice
  – RUM

Spliced mapping

Figure:
Garber et al, Nature Methods, 2011
Paired end mapping

- Pairs map with expected insert size.
- One part of the pair, after mapping, is the anchor for the next read’s mapping.

Paired end read mapping is very helpful in RNA seq!!

http://genomebiology.com/content/12/1/R6
Mapping Quality

• Mapping quality is the probability that a read is aligned to the wrong place.
  \[ p = 10^{2 \cdot (-q/10)} \]

• BWA mapping quality calculated by considering:
  • Repeat structure of reference
  • Read alignment quality (mismatches etc)
  • Number of mappings
  • BWA will assign a mapping quality of 0 to reads that mapped equally well to multiple places
Mappers comparisons

Accuracy Performance of Aligners

- bwa
- gsnap
- rum
- star
- tophat

Fraction of total reads

- Correctly Mapped
- Ambiguously Mapped
- Incorrectly Mapped
- Unmapped

New benchmarking analysis performed by Raghav Shroff
Mappers comparison

New benchmarking analysis performed by Raghav Shroff
SAM file format

- Alignment results generated in Sequence Alignment/Map format
- Tab delimited, with fixed columns followed by user-extendable key:data values.
- Most mappers also output unmapped reads in SAM file.
- SAMTOOLS - toolkit to manipulate, parse SAM files.
SAM File Format

SAM fixed fields:

<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>[0,2^29-1]</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]+[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 segment</td>
</tr>
<tr>
<td>8</td>
<td>PNEXT</td>
<td>Int</td>
<td>[0,2^29-1]</td>
<td>Position of the mate/next segment</td>
</tr>
<tr>
<td>9</td>
<td>TLEN</td>
<td>Int</td>
<td>[-2^29+1,2^29-1]</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>

SRR030257.264529   99   NC_012967   1521   29   34M2S   =   1564
79   CTGGCCATTATCTCGGTGGTAGGACATGGCATGCC
AAAAAAA;AA;AAAAAAA??A%.;?&'3735',()0*,
# CIGAR score

Ref: CTGGCCATTATCTC--GGTGGTAGGACATGGCATGCCC!

Read: aaATGTCGCGGTG.TAGGAggatcc!

![Cigar](image.jpg)

<table>
<thead>
<tr>
<th>Op</th>
<th>BAM</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>0</td>
<td>alignment match (can be a sequence match or mismatch)</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>insertion to the reference</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>deletion from the reference</td>
</tr>
<tr>
<td>*N</td>
<td>3</td>
<td>skipped region from the reference</td>
</tr>
<tr>
<td>S</td>
<td>4</td>
<td>soft clipping (clipped sequences present in SEQ)</td>
</tr>
<tr>
<td>*H</td>
<td>5</td>
<td>hard clipping (clipped sequences NOT present in SEQ)</td>
</tr>
<tr>
<td>*P</td>
<td>6</td>
<td>padding (silent deletion from padded reference)</td>
</tr>
<tr>
<td>*</td>
<td>7</td>
<td>sequence match</td>
</tr>
<tr>
<td>*X</td>
<td>8</td>
<td>sequence mismatch</td>
</tr>
</tbody>
</table>

CIGAR = "Concise Idiosyncratic Gapped Alignment Report"

2S5M2I4M1D4M6S
BAM format

• SAM files are converted to BAM format through SAMTOOLS command:
  
  • samtools view -b -S samfile > bamfile

• BAM file is binary format.

• BAM file is compressed.

• BAM files are usually what you need for post mapping analysis and visualization.
TAKEAWAYS

• **Unspliced mapper** -
  – Most suited for mapping to transcriptome
  – Example: BWA

• **Spliced mapper** -
  – Most suited for mapping to genome
  – Example: Hisat2, Star

• **Mapping output**
  – SAM File: tab-delimited file
  – Filter SAM file, Assess mapping stats
The QNAME is the query name. For the FLAG of 163 we transform this into a binary string: 10100011. So accordingly to the flag table:

<table>
<thead>
<tr>
<th>Flag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x0001</td>
<td>the read is paired in sequencing, no matter whether it is mapped in a pair</td>
</tr>
<tr>
<td>0x0002</td>
<td>the read is mapped in a proper pair (depends on the protocol, normally inferred during alignment) (^1)</td>
</tr>
<tr>
<td>0x0004</td>
<td>the query sequence itself is unmapped</td>
</tr>
<tr>
<td>0x0008</td>
<td>the mate is unmapped (^1)</td>
</tr>
<tr>
<td>0x0010</td>
<td>strand of the query (0 for forward; 1 for reverse strand)</td>
</tr>
<tr>
<td>0x0020</td>
<td>strand of the mate (^1)</td>
</tr>
<tr>
<td>0x0040</td>
<td>the read is the first read in a pair (^1,2)</td>
</tr>
<tr>
<td>0x0080</td>
<td>the read is the second read in a pair (^1,2)</td>
</tr>
<tr>
<td>0x0100</td>
<td>the alignment is not primary (a read having split hits may have multiple primary alignment records)</td>
</tr>
<tr>
<td>0x0200</td>
<td>the read fails platform/vendor quality checks</td>
</tr>
<tr>
<td>0x0400</td>
<td>the read is either a PCR duplicate or an optical duplicate</td>
</tr>
</tbody>
</table>

\(^1\) Most modern aligners infer strand use.

\(^2\) The strand of a pair is determined using the strand of the read with the lowest 1000 bases.

---

1  the read is paired in sequencing, no matter whether it is mapped in a pair
0  mate is not unmapped
1  the read is mapped in a proper pair
0  forward strand
1  mate strand is negative
0  the read is not the first read in a pair
1  the read is the second read in a pair
0  not unmapped