Mapping Next Generation Sequence Reads

Bingbing Yuan

Dec. 2, 2010
What happen if reads are not mapped properly?

• Some data won’t be used, thus fewer reads would be aligned.

• Reads are mapped to the wrong location, creating false positives and false negatives
Our pipeline outline

- Part 1
  - Quality Control
    - Mapping
      - (Chip-seq + RNA-seq)
    - Visualization
- Part 2
- Part 3
Illumina data format

- Fastq format: (QualityScore/s_1_sequence.txt)
  
  ```
  @ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
  GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
  +ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
  hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
  ```

  - `@seq identifier`
  - `seq`
  - `+any description`
  - `seq quality values`
Check read quality

• Overall read distribution, read quality
• Per-cycle base call, quality scores
• May need to
  – remove reads with lower quality
  – Trim the read seq
  – Remove adapter/linker seq
Freely Available Tools for QC

• Galaxy:
  – http://main.g2.bx.psu.edu/
  – Many functions
  – Long time for uploading files since it is on remote server

• Fastx toolkit:
  – http://hannonlab.cshl.edu/fastx_toolkit/
  – galaxy integration, Linux(Tak), MacOSX

• FastQC (picard):
  – http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc
  – Linux(Tak), Window, MacOSX

• Shortread:
  – R package, Linux (Tak), Window, Mac
## General information about reads

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filename</td>
<td>SRR015149.fastq</td>
</tr>
<tr>
<td>File type</td>
<td>Conventional base calls</td>
</tr>
<tr>
<td>Total Sequences</td>
<td>8923918</td>
</tr>
<tr>
<td>Sequence length</td>
<td>26</td>
</tr>
<tr>
<td>%GC</td>
<td>43</td>
</tr>
</tbody>
</table>

Created with FastQC
Overall read quality

Quality Score distribution over all sequences

- Mean Sequence Quality (Phred Score)
- Ave. quality per read

FastQC from Babraham Bioinformatics

Galaxy: Filter FASTQ, filter by quality
FASTX toolkit: fastq_quality_filter
Phred quality scores are logarithmically linked to error probabilities

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.90%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10000</td>
<td>99.99%</td>
</tr>
</tbody>
</table>
Most abundant Reads

provide clues to the source of over-represented sequences. Some of these reads are filtered by the alignment algorithms; other duplicate reads might point to sample preparation issues.

<table>
<thead>
<tr>
<th>sequence</th>
<th>count</th>
<th>lane</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATCGGAAAGAGCTCGTATGCGGTCTTT</td>
<td>231002</td>
<td>character</td>
</tr>
<tr>
<td>GNNNNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
<td>8626</td>
<td>character</td>
</tr>
<tr>
<td>ANNNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
<td>7405</td>
<td>character</td>
</tr>
<tr>
<td>AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA</td>
<td>5539</td>
<td>character</td>
</tr>
<tr>
<td>TNNNNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
<td>4502</td>
<td>character</td>
</tr>
<tr>
<td>CNNNNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
<td>4334</td>
<td>character</td>
</tr>
<tr>
<td>GATCGGAAAGAGCTCGTATGCGGTCTTT</td>
<td>3809</td>
<td>character</td>
</tr>
<tr>
<td>AAAATCATGGAAAAATGATTCTAGATC</td>
<td>3171</td>
<td>character</td>
</tr>
<tr>
<td>GATCGGAAAGAGCTCGTATGCGGTCTTT</td>
<td>2988</td>
<td>character</td>
</tr>
<tr>
<td>GATCGGAAAGAGCTCGTATGCGGTCTTT</td>
<td>2143</td>
<td>character</td>
</tr>
<tr>
<td>GATNNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
<td>2054</td>
<td>character</td>
</tr>
<tr>
<td>GATNNNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
<td>2012</td>
<td>character</td>
</tr>
<tr>
<td>GATCGGAAAGAGCTCGTATGCGGTCTTT</td>
<td>1999</td>
<td>character</td>
</tr>
<tr>
<td>GAATATGGCAAGAAACTGAAAATCA</td>
<td>1956</td>
<td>character</td>
</tr>
<tr>
<td>GATCGGAAAGAGCTCGTATGCGGTATT</td>
<td>1905</td>
<td>character</td>
</tr>
<tr>
<td>GATCGGAAAGAGCTCGTATGCGGTATT</td>
<td>1774</td>
<td>character</td>
</tr>
<tr>
<td>GATCGGAAAGAGCTCGTATGCGGTATT</td>
<td>1698</td>
<td>character</td>
</tr>
<tr>
<td>GATCGGAAAGAGCTCGTATGCGGTATT</td>
<td>1659</td>
<td>character</td>
</tr>
<tr>
<td>GATCGGAAAGAGCTCGTATGCGGTATT</td>
<td>1603</td>
<td>character</td>
</tr>
<tr>
<td>GAANNNNNNNNNNNNNNNNNNNNNNNNNNN</td>
<td>1502</td>
<td>character</td>
</tr>
</tbody>
</table>
## Most abundant Reads

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Count</th>
<th>Percentage</th>
<th>Possible Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTGTAGGCACCATCAATTGTGATGCTGCTTTCTGCT</td>
<td>1175220</td>
<td>5.90</td>
<td>No Hit</td>
</tr>
<tr>
<td>AAGAGGTGCAACCAATCGACCGATCTGACTGTAGGCA</td>
<td>359160</td>
<td>1.80</td>
<td>No Hit</td>
</tr>
<tr>
<td>TACACGGAGTCGACGCACCGCAACGCGACTGTAGGCACC</td>
<td>77161</td>
<td>0.38</td>
<td>No Hit</td>
</tr>
<tr>
<td>TGTAGGCACCATCAATTGTGACGCTGCTTTCTGCTT</td>
<td>70591</td>
<td>0.34</td>
<td>Illumina Single End Apapter 2 (95% over 21bp)</td>
</tr>
<tr>
<td>ACGCGAAACTCAGGTGCTGCAATCTCTGTAGGCACC</td>
<td>67674</td>
<td>0.34</td>
<td>No Hit</td>
</tr>
<tr>
<td>TCGAAGAGTCGAGTTTTGGAATGCTGCTGACC</td>
<td>66160</td>
<td>0.33</td>
<td>No Hit</td>
</tr>
</tbody>
</table>

Created with FastQC
Per-cycle quality score

Quality Scores across all bases (Illumina >v1.3 encoding)

Position in read (bp)

From FastQC in Babraham Bioinformatics
Per-cycle base call

• Trim reads:
  FASTX toolkit: \textit{fastx\_trimmer}
  Galaxy: FASTQ Trimmer
  Trim sequences

\textit{from Fastx in}
\textit{http://hannonlab.cshl.edu/fastx\_toolkit}
Remove adapter/linker

Remove adapter/linker

From fastx: **fastq_clipper** in http://hannonlab.cshl.edu/fastx_toolkit
Sequence length distribution after clipping

Sequence Lengths Distribution (After clipping)

Created with FASTX toolkit: fastx_quality_stats, fastx_nucleotide_distribution_graph.sh
Challenges of mapping short reads

- Large genome
- Billions of reads
- Speed
- Repeat regions
- Sequencing errors, reference genome variations

Mapping Techniques:
- Index genome:
  - Burrows-Wheeler Transform: order the genome
  - FM index: index genome
Free Mapping software for Chip-seq

• Bowtie (tak):
  • Langmead et al. (2009) Genome Biology, 10:R25
  • http://bowtie-bio.sourceforge.net/index.shtml
  • One of the fastest alignment software for short reads
  • Not gapped-alignment
  • Base quality can be used for evaluating alignments
  • Mismatch: 0-3
  • Flexible reporting mode including SAM format

• BWA (Burrows-Wheeler Alignment Tool ):
  • http://bio-bwa.sourceforge.net/bwa.shtml
  • Short reads up to 200bp
  • Gapped alignment
  • Base quality not used for evaluating alignments
  • Allow >3 mismatches
  • Need to run samse/sampe to get SAM format
Reporting the alignments by bowtie

• Unique alignment only
• Reporting ambiguous hits
  – Randomly report one
  – Report all alignments above cutoff parameters
  – Report all alignments with best alignment scores
Bowtie options

- Index genome once: `/nfs/genome/`
  - `mm9.1.ebwt`  `mm9.2.ebwt`  `mm9.3.ebwt`  `mm9.4.ebwt`
  - `mm9.rev.1.ebwt`  `mm9.rev.2.ebwt`

- Alignment:
  - Seed: The first L bases are called the “seed” (-l)
  - Max mismatches in seed (-n)

- Reporting:
  - Report number of alignment per read (-k)
  - Suppress all alignments (-m)
  - Report best hits (--best, --strata)

- Output:
  - unalign reads (--un)
  - reads over –m cutoff (--max)
  - Sam format (-S)
Bowtie examples

- bowtie solexa1.3-quals -n 1 -l 36
  /nfs/genomes/mouse_gp_jul_07_no_random/bowtie/mm9
  s3_sequence.txt

- bowtie solexa1.3-quals -n 1 -l 36 –m 10 –k 10 –max mapOver10.fq
  /nfs/genomes/mouse_gp_jul_07_no_random/bowtie/mm9
  s3_sequence.txt

- bowtie solexa1.3-quals -n 1 -l 36 –m 10 –k 2 ---best --strata
  /nfs/genomes/mouse_gp_jul_07_no_random/bowtie/mm9
  s3_sequence.txt
RNA-Seq

Mapping RNA-seq

• Map to genome:
  – Computationally expensive
  – Potential novel transcripts
  – reads across exon-exon junction will not be aligned

• Map to transcripts:
  – Computationally inexpensive
  – Limited by the annotation files

• De novo assembly
  – No reference genome
  – With reference genome: allows detection of chimeric transcripts
    • Assemble into contigs and align (BLAT) against genome
Mapping RNA-seq with Tophat

• [http://tophat.cbcb.umd.edu/](http://tophat.cbcb.umd.edu/)
• Trapnell et al, 2009. PMID: 19289445
• built on the ultrafast short read mapping program Bowtie
• find splice junctions without a reference annotation.
  1. first mapping RNA-Seq reads to the genome
  2. builds a database of possible splice junctions
     a. distinct regions of piled up reads in the initial mapping
     b. evidence for a splice junction: such as alignments across "GT-AG" introns
     c. paired end reads
• Linux (Tak)
SAM/BAM

• Examples:
  – read_28833_29006_6945 99 chr20 28833 20 10M1D25M = 28993 195 \ AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<<<<<<\NM:i:1 RG:Z:L1
  – read_28701_28881_323b 147 chr20 28834 30 35M = 28701 -168 \ ACCTATATCTTGGCCCTGATGCCTTTGCA <<<<<<;<<<<7;<<<<6;<<<<<<<<<<<<7<<<< \MF:i:18 RG:Z:L2
  – SRR015149.61819 16 chr3 29065583 255 26M * 0 0 ACNAATTGCNATGCAGACACTTCACC """".6,""""=1!l+IIIIIBII)CII XA:i:2 MD:Z:2G6G16 NM:i:2

• Files for browsers:
  – Convert SAM to BAM, then sort, index bam files with samtools (tak) (http://samtools.sourceforge.net/)
  – Sort, index SAM with IGV tools (http://www.broadinstitute.org/igv/)
  – UCSC genomes browser:
    • BAM format needs to be on http/ftp server
    • convert BAM to bedGraph with genomeCoverageBed from bedtools (tak) (http://bioinformatics.oxfordjournals.org/content/26/6/841.full)
IGV
UCSC Genome Browser
Our pipeline

Part 1: QC
FastQC, FASTX-toolkit, shortread

Part 2: mapping
Bowtie, TopHat

Part 3: view
IGV

Chip-Seq RNA-Seq

fastq
References

- FastQC to check the quality of high throughput sequence
  http://www.youtube.com/watch?v=bz93ReOv87Y