next-gen sequencing informatics primer

Adrian Heilbut <amh@bu.edu>

BU Bioinfo Program Retreat
May 22 2010
1. Technologies & Costs
2. Applications
   Library Design
3. Informatics
   Inputs: FASTQ
   Alignment Algorithms
   Outputs: SAM/BAM
   QC issues
   Analysis: SVs, SNPs
growth in sequencing DBs

Nature, 1 April 2010
Technologies

454
ABI Solid
Illumina (Solexa)
Helicos

PacBio
Complete Genomics
Ion Torrent
**454 (Roche)**

- first 2nd generation technology - 2005
- can do long reads (400bp)
- suitable for de novo genome sequencing of moderately sized genomes
Adaptors
Beads
PCR
Ligase

- 60 gigabases per run
- “emulsion PCR”
- sequencing by ligation
- colorspace output
The SOLiD sequencing process involves several steps:

1. **Prime and Ligate**
   - The process begins with the prime and ligation of the template sequence.
   - The primer contains a PAL (POH) handle for ligation.

2. **Image**
   - This step involves exciting the fluorophore and observing the fluorescence.
   - Each cycle produces a new fluorophore-labeled read.

3. **Cap Unextended Strands**
   - After each cycle, unextended strands are capped to prevent further extension.
   - A phosphatase enzyme is used to dephosphorylate the ends.

4. **Cleave off Fluor**
   - The fluorophore is cleaved off using a cleavage agent.
   - This step is crucial for distinguishing the next cycles.

5. **Repeat steps 1-4 to Extend Sequence**
   - The process is repeated for multiple cycles to extend the sequence.
   - Each cycle produces a new read, allowing for the accumulation of information.

6. **Primer Reset**
   - Once all cycles are completed, the primer is reset to the original state.
   - A new primer is added to initiate the next cycle.

The SOLiD system enables massively parallel sequencing, providing both sequence and structural information. The diagram illustrates the key steps of the process, emphasizing the cycle of primer, ligation, imaging, and fluorophore cleavage.
7. Repeat steps 1-5 with new primer

8. Repeat Reset with n-2, n-3, n-4 primers

| Read Position | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 |
|---------------|---|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1 Universal seq primer (n) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 2 Universal seq primer (n-1) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 3 Universal seq primer (n-2) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4 Universal seq primer (n-3) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 5 Universal seq primer (n-4) |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

- Indicates positions of interrogation

Ligation Cycle: 1 2 3 4 5 5 7
Figure 2. SOLiD Color Space Code — Four dyes encode for sixteen potential two base combinations
20–200 gigabases per run (depending on instrument)
1. **PREPARE GENOMIC DNA SAMPLE**

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. **ATTACH DNA TO SURFACE**

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. **BRIDGE AMPLIFICATION**

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
1. PREPARE GENOMIC DNA

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGING AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE-STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.
7. **DETERMINE FIRST BASE**

The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

8. **IMAGE FIRST BASE**

After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

9. **DETERMINE SECOND BASE**

The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.
10. IMAGE SECOND CHEMISTRY CYCLE

After laser excitation, the image is captured as before, and the identity of the second base is recorded.

11. SEQUENCING OVER MULTIPLE CHEMISTRY CYCLES

The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

12. ALIGN DATA

The data are aligned and compared to a reference, and sequencing differences are identified.
Current Costs (eg. Illumina)

~$1k per lane (54 cycle, single end)

~$2k per lane (74 cycle, paired end)

→ 6 gigabases of sequence

→ 1–2x human genome coverage
• De Novo sequencing
• Resequencing
  • SNP discovery
  • Structural variation
• RNA-seq
  • gene expression
  • small RNAs
Library Preparation

- ~ 3µg source DNA needed
- whole genome amplification
  - every PCR step adds potential bias and artifacts
- Paired-ends, Jumping libraries
  - increase effective coverage
Hybrid Capture / Enrichment

- Pre-designed exon capture arrays
- Design-your-own capture arrays
- may require optimization / iteration

Figure 2  SureSelect Target Enrichment System Capture Process
Figure 9  Origin and Alignment of Inward and Outward-Facing Reads

Source: Illumina Mate Pair Library V2 Sample Prep Guide
Informatics

Read files & quality scores
Primary QC
Quality recalibration
Alignment algorithms
Alignment output formats
Open problems?
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;7;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;7;;;;;;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGCGTGCTGGGTGATGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;9;7;.7;393333
PHRED quality encoding

\[ Q = -10 \log_{10} P \]

\[ P = 10^{\frac{-Q}{10}} \]

Encoded in in FASTQ / SAM by quality string of

ASCII value – 33

scales differ slightly between PHRED and older versions of Illumina outputs
Quality vs. Cycle; Quality vs. Tile

Often bias towards errors at 3’ end

Fastx_tools, PIQA helpful tools for visualizing quality
Aligner Performance Varies Widely

The speed varies... by Bala et al.

Heng Li (Broad Institute)

Bala et al. via Heng Li
Hashing / Spaced Seed Aligners

Hash Queries:
- Eland (Cox – illumina)
- MAQ (Li, Durbin)

Hash Genome:
- Mosaik (Stromberg, Marth)
Burrows–Wheeler Alignment

BWA (Li, Durbin)

Bowtie (UMD)

SOAP2 (BGI)

Use FM-index data structure to effectively maintain a suffix array index of the genome in < 3Gb RAM

Fast exact matching

Mismatches searched by varying query sequence
**Sequence Alignment / Map Format**

standardized alignment output format

supported by most modern alignment programs

simple, tab-delimited text-file with an optional bzip-compressed binary encoding (BAM)
The alignment section consists of multiple TAB-delimited lines with each line describing an alignment. Each line is:

- **QNAME**: Query pair NAME if paired; or Query NAME if unpaired
- **FLAG**: bitwise FLAG (Section 2.2.2)
- **RNAME**: Reference sequence NAME
- **POS**: 1-based leftmost POSition/coordinate of the clipped sequence
- **MAPQ**: MAPping Quality (phred-scaled posterior probability that the mapping position of this read is incorrect)
- **CIGAR**: extended CIGAR string
- **MRNM**: Mate Reference sequence NaMe; “=” if the same as <RNAME>
- **MPOS**: 1-based leftmost Mate POSition of the clipped sequence
- **ISIZE**: inferred Insert SIZE
- **SEQ**: query SEQuence; “=” for a match to the reference; n/N/ for ambiguity; cases are not maintained
- **QUAL**: query QUALity; ASCII-33 gives the Phred base quality
- **TAG**: TAG
- **VTYPE**: Value TYPE
- **VALUE**: match <VTYPE> (space allowed)

### Table of Field Descriptions

<table>
<thead>
<tr>
<th>Field</th>
<th>Regular expression</th>
<th>Range</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>QNAME</td>
<td>^ \t\n\r]+</td>
<td></td>
<td>Query pair NAME if paired; or Query NAME if unpaired</td>
</tr>
<tr>
<td>FLAG</td>
<td>[0-9]+</td>
<td>[0,2^{16}-1]</td>
<td>bitwise FLAG (Section 2.2.2)</td>
</tr>
<tr>
<td>RNAME</td>
<td>^ \t\n\r@=]+</td>
<td></td>
<td>Reference sequence NAME</td>
</tr>
<tr>
<td>POS</td>
<td>[0-9]+</td>
<td>[0,2^{29}-1]</td>
<td>1-based leftmost POSition/coordinate of the clipped sequence</td>
</tr>
<tr>
<td>MAPQ</td>
<td>[0-9]+</td>
<td>[0,2^{8}-1]</td>
<td>MAPping Quality (phred-scaled posterior probability that the mapping position of this read is incorrect)</td>
</tr>
<tr>
<td>CIGAR</td>
<td>([0-9]+[MIDNSHP])+</td>
<td>*</td>
<td>extended CIGAR string</td>
</tr>
<tr>
<td>MRNM</td>
<td>^ \t\n\r@]+</td>
<td></td>
<td>Mate Reference sequence NaMe; “=” if the same as &lt;RNAME&gt;</td>
</tr>
<tr>
<td>MPOS</td>
<td>[0-9]+</td>
<td>[0,2^{29}-1]</td>
<td>1-based leftmost Mate POSition of the clipped sequence</td>
</tr>
<tr>
<td>ISIZE</td>
<td>-?[0-9]+</td>
<td>[-2^{29},2^{29}]</td>
<td>inferred Insert SIZE</td>
</tr>
<tr>
<td>SEQ</td>
<td>[acgtnACGTN.=]+</td>
<td>*</td>
<td>query SEQuence; “=” for a match to the reference; n/N/ for ambiguity; cases are not maintained</td>
</tr>
<tr>
<td>QUAL</td>
<td>[!--]+</td>
<td>[0,93]</td>
<td>query QUALity; ASCII-33 gives the Phred base quality</td>
</tr>
<tr>
<td>TAG</td>
<td>[A-Z][A-Z0-9]</td>
<td></td>
<td>TAG</td>
</tr>
<tr>
<td>VTYPE</td>
<td>[AifZH]</td>
<td></td>
<td>Value TYPE</td>
</tr>
<tr>
<td>VALUE</td>
<td>^ \t\n\r]+</td>
<td></td>
<td>match &lt;VTYPE&gt; (space allowed)</td>
</tr>
</tbody>
</table>
### SAM Bitwise Flags

<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>

- flags added together to produce value in FLAG field
- to test a flag: (value && testFlag)==testFlag
C API and common tools for manipulating SAM & BAM files

GATK / Picard does similar things in Java
Alignment Viewers

IGV

via Jim Robinson
• Many current sequencing projects are looking at SNPs -> lots of infrastructure for pileups and SNP-calling

• No standardized tools for analyzing structural variation from next-gen data

• GATLING is a set of lightweight utilities (a la samtools) for analyzing SAM/BAM output for SV and visualizing results

• still under development...
open problems

alignment speed / error tolerance
sorting speed
storage
???
Resources

SeqAnswers.com
i.seqanswers.com
biostar.stackexchange.com

samtools-devel
bio-bwa-help
questions / discussion