Next Generation Sequencing Technologies

Rob Mitra
2/1/2021
Outline

• Overview of next-generation sequencing
  – How does it work?
  – What technologies are being used?
  – How would one use it in practice?

• Math basics for Genomics
  – Poisson Distribution
Forward Genetics

Mutation in APC Gene

Two groups:
1. Develop Colorectal cancer At Young Age
2. Do not

Genotype

Hypothesis
APC is a Tumor Suppressor Gene

Phenotype

Test Hypothesis By Genetic Manipulation

Delete APC in Mouse Control: Isogenic APC+
The Cycle of Forward Genetics

Observation → Phenotype

Genotype

?Sequencing?

Hypothesis

In 2005
$9 million/genome
Not feasible

Test Hypothesis
By Genetic Manipulation

Gene Deletion/Replacement
Recombinant Technology

Thinking
The Problem with Forward Genetics

- **Sequencing**
- **Observation**
- **Phenotype**
- **Test Hypothesis By Genetic Manipulation**
- **Hypothesis**
- **Thinking**
- **Genotype**

Currently <$800$* /genome
Cost is rapidly dropping

Gene Deletion/Replacement
Recombinant Technology
Bp/US dollar: increases exponentially with time

Adapted from Shendure et al 2004
Three questions:

- How was this dramatic acceleration achieved?
- What has it meant?
- What is the future?
How was this achieved?

• Integration (Think about sequencing pipeline)

• Parallelization

• Miniaturization

Same concepts the revolutionarized integrated circuits

Plus one additional insight
Technology Overview: Solexa/Illumina Sequencing

Sequencing-By-Synthesis Demo

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

http://www.illumina.com/
Immobilize DNA to Surface

Sequencing-By-Synthesis Demo

- Adapter
- DNA fragment
- Dense lawn of primers
- Adapter

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

Source: www.illumina.com
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Sequence Colonies

Sequencing-By-Synthesis Demo

First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
Sequence Colonies

Image of second chemistry cycle is captured by the instrument. After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
Call Sequence

Sequencing-By-Synthesis Demo

Cycle 1: T G C A
Cycle 2: A C T G
Cycle 3: A U T C
Cycle 4: G A C
Cycle 5: T A G G

Sequence read over multiple chemistry cycles
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.
What do we do with 100 million, 100 base pair long reads?
Paired End Reads are Important!

Known Distance

Read 1       Read 2

Repetitive DNA

Unique DNA

Paired read maps uniquely

Single read maps to multiple positions
Paired Ends are Important Part 2

Deletion

Insertion

Inversion

Shendure et al 2005
Sequencing: Illumina

- Variety of workflows:
  - Single- or paired end reads
  - 0, 1, or 2 index reads

Credit: http://image.slidesharecdn.com/ngsmicrobiome-120717185901-phpapp02/95/ngs-microbiome-14-728.jpg?cb=1342551631
How can we generate longer paired end reads?
Library Preparation: Mate Pair

- Why? Paired sequences are a known distance apart; improves genome assembly

- Note: 454 calls these “paired end libraries”, not to be confused with Illumina’s “paired end sequencing”!

Can we get really long reads?

Haplotyping Contigs with 10X

Long read information from short reads using 14bp bar codes

Very low input DNA (0.625 ng for ACP)

1ng of DNA is split across 100,000 Gel Coated Beads (GEMs)

Chromium instrument
Where does the technology stand?

- Right now, fully loaded costs are roughly <$8/GB at Wash U GTAC.
Nuts and Bolts of Sequencing

• Resequencing Human Genomes
  – Map reads back to genome
  – Call bases

• RNA-seq
  – Map reads back to genome
  – Count tags to determine gene expression levels

• Chip Seq
  – Map reads back to genome
  – Peaks determine binding sites.

Nearly all experiments have the same first step!
Mapping Reads Back

• Hash Table (Lookup table)
  – FAST, but requires near perfect matches

• Array Scanning
  – Can handle mismatches, but not gaps

• Dynamic Programming (Smith Waterman, Forward, Viterbi)
  – Indels
  – Mathematically optimal solution
  – Slow (most programs use Hash Mapping as a prefilter)

• Burrows-Wheeler Transform (BW Transform)
  – FAST (memory efficient)
  – But for gaps/mismatches, it often lacks sensitivity
## Aligners Evaluated

<table>
<thead>
<tr>
<th>Aligner</th>
<th>Version</th>
<th>Developer</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFAST</td>
<td>0.3.1</td>
<td>UCLA</td>
<td>Academic</td>
</tr>
<tr>
<td>Bowtie</td>
<td>0.9.8</td>
<td>Salzberg/UMD</td>
<td>GNUPL</td>
</tr>
<tr>
<td>cross_match</td>
<td>1.080721</td>
<td>U. Wash.</td>
<td>Academic</td>
</tr>
<tr>
<td>CELL</td>
<td>2.0</td>
<td>CLC bio</td>
<td>Commercial</td>
</tr>
<tr>
<td>Maq</td>
<td>0.7.1</td>
<td>Sanger</td>
<td>GNUPL</td>
</tr>
<tr>
<td>Novoalign</td>
<td>2.00.12</td>
<td>Novocraft</td>
<td>Academic</td>
</tr>
<tr>
<td>RMAP</td>
<td>0.41</td>
<td>CSHL</td>
<td>Public</td>
</tr>
<tr>
<td>SeqMap</td>
<td>1.0.12</td>
<td>Stanford</td>
<td>Academic</td>
</tr>
<tr>
<td>SHRiMP</td>
<td>1.1.0</td>
<td>U. Toronto</td>
<td>Public</td>
</tr>
<tr>
<td>SOAP</td>
<td>2.01</td>
<td>BGI</td>
<td>Academic</td>
</tr>
</tbody>
</table>

Slide Courtesy of Dan Kobolt, Genome Sequencing Center
See www.massgenomics.org
CPU Time 2M Reads to Hs36: SE/PE

- Benchmark: Maq (~8 hours)

Slide Courtesy of Dan Kobolt, Genome Sequencing Center
• Most aligners place \(~80\%\) of reads uniquely in Hs36.

Slide Courtesy of Dan Kobolt, Genome Sequencing Center
PE Mode Increases Unique Mapping

Slide Courtesy of Dan Kobolt, Genome Sequencing Center
What we find

- BW transform algorithms (Bowtie) are great for RNA-seq, ChIP-Seq

- We prefer Novoalign, which uses a seed-based approach followed by DP to find SNPs and INDELs. If you really want to align indels, use full DP, with a hash based strategy. (SLOW)
Math Aside: Sequencing coverage calculations

• Let’s say you need a base to be sequenced 5x for an accurate base call

• If you sequence at 10x coverage how much of the genome will be sequenced at least 5 times?
Poisson Distribution

\[ f(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \]

Deriving the Poisson:
(derive from the binomial for large number of trials but small probability of success)


Average coverage = \lambda
Probability of getting \( k \) reads from a base given the average coverage \( \lambda \)
Example

• Average coverage = 10x
• Probability of a given base being sequenced exactly 5 times is:

\[ f(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \]

\[ 10^5 e^{-10}/5! = 0.038 \text{ or about } 4\% \text{ of bases will have } 5x \text{ coverage.} \]
Math Aside: Sequencing coverage calculations

• If you sequence at 10x coverage how much of the genome will be sequenced at least 5 times?

\[ 1 - [f(0,10) + f(1,10) + f(2,10) + f(3,10) + f(4,10)] = 0.97 \]
A crash course in probability
“3rd generation” sequencing

- Pac Bio
Pacific Biosciences: A Third Generation Sequencing Technology

Eid et al 2008
Real Time Sequencing

![Graphs showing real-time sequencing data with expected bases and fluorescence intensity over time.](image)
PacBio circa 2021

- Two modes, CLR (long reads) CCS (circular consensus sequencing)
- 4-6 Million reads per run
- Median read length = 35kb. Mean ~25kb
- 30GB (CCS) or 100 GB CLR
- ~88% raw accuracy  99.9% accuracy for CCS
- $25/GB CLR $110 CCS