Next Generation DNA sequencing: why should we care

Shoudan Liang  
(Keith Baggerly & Bradley Broom)  
Microarray Analysis Class, Nov 23, 2010
SEQUENCING CAPACITY IS GROWING EXPONENTIALLY

- first human genome sequenced over ten years at $3 billion.
- 2007, Watson’s genome was sequenced in two months by 454 at $2 million.
- Last year, the cost (list price of reagent) of human genome re-sequencing using Solexa is $250,000.
- ABI SOLiD claim to be able to re-sequence at $10,000 this year.
The cost of sequencing DNA has dropped by more than a million folds in the last ten years.
CPU Transistor Counts 1971-2008 & Moore’s Law

Curve shows ‘Moore’s Law’: transistor count doubling every two years.
SEQUENCING IS EXPECTED TO FOLLOW MOORE’S-LIKE LAW

- Moore’s law: computing power a dollar can buy doubles every 18 months

- rate limiting step in NEXT GEN sequencing is imaging. CCD camera in sequencer will increase in capacity following Moore’s law.

- DNA sequencing with semiconductor : merging of two technologies?
Pixels per dollar of Kodak digital cameras
Applications in genome research

- a DNA sequence is a bar-code, and therefore an addressing system of a genome
- share similarities with microarray in measuring amount of DNA by genome locations
Steps in Preparing an RNA-Seq Library

1. Purify RNA
2. Bind polyA fraction (mRNA)
3. Fragment RNA (200 bp)
4. Convert to cDNA by random priming
5. Apply adaptors and sequence
6. Analyze millions of 25 bp reads
Digital gene expression
Solexa vs Gold Standard

787 RefSeq human transcripts in brain and UHR TaqMan is considered a gold standard

Diversity of The Human Genome

Amount of diversity observed per individual genome:

- 250-300 loss-of-function variants
- 50-100 variants implicated in inherited disorders
- ~10,000 non-synonymous cSNP differences compared to a published reference genome

{15 million SNPs; 1 million short indels; 20,000 structural variants were identified from total sequence data}
TECHNOLOGY LOOKING FOR PROBLEMS

• currently, USA has 600 next-generation sequencers. The rest of the world another 500 or so.

• Number of human genomes to be sequenced by the end of next year is about 30,000.
<table>
<thead>
<tr>
<th>platform</th>
<th>Feature generation</th>
<th>Cost per mega base</th>
<th>Cost per instrument</th>
<th>most common error</th>
<th>Read-length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche GS-FLX (454)</td>
<td>Emulsion PCR</td>
<td>$20</td>
<td>$500,000</td>
<td>Indel</td>
<td>400 bp</td>
</tr>
<tr>
<td>Illumina GA (Solexa)</td>
<td>Bridge PCR</td>
<td>$2</td>
<td>$430,000</td>
<td>Subst.</td>
<td>36 bp</td>
</tr>
<tr>
<td>ABI SOLiD</td>
<td>Emulsion PCR</td>
<td>$2</td>
<td>$591,000</td>
<td>Subst.</td>
<td>35 bp</td>
</tr>
<tr>
<td>HeliScope</td>
<td>Single molecule</td>
<td>$1</td>
<td>$1,350,000</td>
<td>Del</td>
<td>30 bp</td>
</tr>
<tr>
<td>Pacific Biosciences</td>
<td>Single molecule</td>
<td>-</td>
<td>-</td>
<td>Del/Subst.</td>
<td>long</td>
</tr>
<tr>
<td>Complete Genomics</td>
<td>Nanoball</td>
<td>$0.01</td>
<td>NA</td>
<td>Subst.</td>
<td>35 bp</td>
</tr>
</tbody>
</table>

NEXT GEN SEQUENCING: HARDWARE

• Sanger
• 454
• ABI SOLiD
• Illumina Solexa
• Complete Genomics
• Pacific Biosciences
Sanger
ELONGATION AFTER PRIMER

http://web.utk.edu/~khughes/SEQ
ENLONGATION STOPS WHEN DIDEXOY BASE IS ENCOUNTERED

http://web.utk.edu/~khughes/SEQ
PRODUCING A LADDER

http://web.utk.edu/~khughes/SEQ
That can be read on gel

DNA fragments are loaded in wells in the gel.

• Fragments separate by size as they migrate through the gel.

• Dyes attached to the dideoxy terminators mark their position in the gel.

http://web.utk.edu/~khughes/SEQ
TRACING OF THE LADDER

http://web.utk.edu/~khughes/SEQ
Sanger vs Next Gen Sequencing

**Sanger Sequencing**
- DNA fragmentation
- In vivo cloning and amplification
- Cycle sequencing:
  - Template: 3'-... GACTAGATACGAGCGTGA...-5'
  - Primer: 5'-... CTGAT
  - Polymerase, dNTPs, Labeled ddNTPs
  - Electrophoresis (1 read/capillary)

**Next Generation Sequencing**
- DNA fragmentation
- In vitro adaptor ligation
- Generation of polony array
- Cyclic array sequencing (>10^6 reads/array)

TWO METHODS OF SINGLE MOLECULE PCR

a: emulsion PCR (454 & SOLiD)
b: bridge PCR (Solexa)
SEQUENCING BY SYNTHESIS

base extension

ligation
Schematic representation of the progress of the enzyme reaction in solid-phase pyrosequencing

Ronaghi M. Genome Res. 2001;11:3-11
Pyrosequencing

5–100ng DNA

2–5 μg DNA

www.454.com
GS FLX Data

Image Processing Overview

1. Raw data is a series of images.
2. Each well's data is extracted, quantified, and normalized.
3. Read data converted into "flowgrams"
Flowgram
ABI SOLiD
PREPARE LIBRARY OF SINGLE STRANDED DNA
SINGLE MOLECULE PCR
Beads on surface
ABI SOLiD Sequencing
16 di-nucleotides probes in 4 steps

Possible Dinucleotides Encoded By Each Color

1st Base

2nd Base

A C G T

A
C
G
T

Template Sequence

TA
CG
GC
TA
AC
CA
CC
TG
AG
CG
TT
CT

Double Interrogation

With 2 base encoding each base is defined twice

A ← T ← G ← G ← A
ABI SOLiD CYCLING

[Diagram showing dual interrogation of each base with primer rounds and read positions]
Illumina Solexa
BRIDGE PCR
• 8 lanes per run
• 200 pictures per lane
• 4X36 pictures for 36-mer
• 1/4 million pictures per 3-day run -> 0.5TB of data
NEW HISEQ 2000

- sequence up to 100bp
- 1 billion tags per experiment
- 25Gbase per day
- reagent cost is about 10 times cheaper than the current product
Complete Genomics
CONSTRUCT A CIRCULAR DNA
ROLLING CIRCLE AMP
HIGH DENSITY PACKING
• each DNA nano-ball is 80 bp genomic DNA (plus 4 adaptors) repeated 200 times.
• reads are equivalent to two 35-bp on paired ends of 500bp DNA.
• rolling circle amplification replaces emulsion or bridge PCR
• 1″X3″ silicon slide holds one billion DNA nano-balls
• reagent cost is 1/1000 of Solexa
• demonstrated 8.8 Gb per machine run per day.
• a completed genome sequence on company’s web site
• June 2009, launch of commercial run: 200Gb per machine run lasting 8 days.
• data center: 60,000 processors and 30 petabytes storage.
• according to Dr Drmanac, CSO of Complete Genomics, Inc

• next generation of machine will have
  • $10 per genome reagent cost
  • $20 per genome of instrument cost
Pacific Biosciences
Unlike Sanger sequencing, which average over many molecules, in next gen sequencing PCR errors do not average away
Application: 3D genome
Chromosome conformation capture (3C) and its derivatives have been used to detect long-range interactions within and between chromosomes, entire chromosomes, and even whole genomes layered on top of information conveyed by DNA sequence and tissue-specific expression. By centromeres and include interactions among transfer RNA genes, among origins of early DNA replication and among sites where chromosomal breakpoints occur. Finally, we constructed a three-dimensional model of the yeast genome. Our findings provide a glimpse of the interface between the form and function of a eukaryotic genome.

In budding yeast, chromosomes remain poorly understood. Here we developed a method’s reliability by assessing: (1) random intermolecular ligations from each of five control libraries (Fig. 2a, Supplementary Tables 1 and 2 and Supplementary Methods); (2) restriction site-based biases and 2 and Supplementary Fig. 2); (4) consistency between the HindIII and EcoRI libraries (Supplementary Figs 3–5 and Supplementary Tables 3); (3) reproducibility between independent sets of experimental libraries (Fig. 2b, Supplementary Figs 1 and 2 and Supplementary Table 3); (5) a set of 24 chromosomal interactions using conventional Supplementary Methods). Because all 3C-based technologies are encumbered by low signal-to-noise ratios.

Vol 465 | 20 May 2010

Figure 1

Our method relies on the 4C procedure by using cross-linking, two rounds of alternating restriction enzyme (such as EcoP15I). Following EcoP15I digestion, adaptors, one of which permits digestion with a type IIS or type III cutter RE2 for the 4C-step digestion) and intra-molecular ligation. At step 7, adapters are produced that incorporate interacting partner sequence at the restricted RE1 site.
Application: tumorigenesis
Geographic mapping of metastatic clones within the primary carcinoma and proposed clonal evolution of Pa08.
• happy thanksgiving