Illumina's next generation sequencing technology

Presented by field applications scientist
Pernille Albertus
Denmark/Norway
Illumina

- headquarter in San Diego, California
- 1800+ employees globally
- develop and sell innovative technologies for studying genetic variation and function enabling rapid advances in disease research, drug development, and the development of molecular tests in the clinic
- founded in 1998 (GoldenGate genotyping)
- acquired Solexa in 2006 (Sequencing By Synthesis)
Illumina Sequencers

Next Generation Sequencing made accessible.

Two proven technologies. One powerful platform.

Most widely adopted NGS platform.

Redefining the trajectory of sequencing.

GA\textsubscript{IIe}  HiScanSQ  GA\textsubscript{IIx}  HiSeq2000
Illumina Array Platforms

- Low- to mid-plex molecular testing.
- Dedicated array instrument.
- Sequencing-compatible array instrument.
- Two proven technologies. One powerful platform.

BeadXpress  
iScan  
HiScan  
HiScanSQ
Sequencing by synthesis chemistry
Workflow

SAMPLE PREP

cBot CLUSTER GENERATION

Genome Analyzer SEQUENCING

DATA PROCESSING & ANALYSIS
The flow cell - a core component

EVERYTHING EXCEPT SAMPLE PREPARATION IS COMPLETED ON THE FLOW CELL

- template annealing (1 - 96 samples)
- template amplification
- sequencing primer hybridization
- Sequencing-by-synthesis reaction
- generation of fluorescent signal
The flow cell surface is coated with oligos
Preparation of template

template DNA
↓
fragment
↓
repair ends
↓
add A overhang
↓
ligate adaptors & purify on gel
↓
enrich
↓
genomic library & library QC

Connect Nebulizer to compressed air or nitrogen using this port

Vinyl Tubing

Atomizer

DNA Solution

[Sequence]

[Image: Molecular sequence diagram]
The flow cell is mounted on the cBot

AUTOMATICALLY
loads library into the lanes of the flow cell
amplifies templates
anneals sequencing primer to templates

FEATURES
intervention-free clonal amplification in 4 hours
simple touch screen operation
Hybridization of template

Grafted flowcell

Template
Hybridization

Initial extension
(Taq Polymerase)

Denaturation
(Formamide)
Amplification of template

1st cycle
Denaturation (Formamide)

1st cycle
Annealing

1st cycle
Extension (Bst Polymerase)

2nd cycle
Denaturation (Formamide)

n=35 total

2nd cycle extension

2nd cycle annealing
Annealing of sequencing primer to template
Summary - "cluster generation"

1. Grafting

2. Hybridization & Amplification

3. Linearization

4. Blocking with ddNTP

5. Denature and Hyb SBS3

Sequencing on Genome Analyzer
The flow cell is mounted on the sequencer.

CCD camera collects laser-excited fluorescence.

Sequencing reagents pass through the 8 lanes inside the flow cell.

Sequencing reaction is temperature controlled.
1. Incorporation

Incorporation sequencing
Scanning

1. Incorporation
2. Scan
Cleavage

1. Incorporation
2. Scan
3. Cleavage
Millions of clusters are sequenced in parallel
A picture is taken every time a new base is added

Sequencing
36bp – 100bp

Image acquisition

Base calling
"Paired-end" sequencing - a core concept

- allows unique mapping of more data
- combined with single reads and mate pair complex structural changes can be discovered

**Repetitive regions in the genome**

If one of the paired reads is unique we can still map the non-unique read because we know the size of the insert.
Hybridization of second sequencing primer is done *in-situ* on the sequencer.
Instrument specifications and throughput
Illumina Sequencer for Everyone!

- Next Generation Sequencing made accessible
- Unique combination of sequencing & arrays
- Most widely adopted NGS platform
- Changing the trajectory of sequencing

GA\textsubscript{1le}, iScanSQ, GA\textsubscript{lix}, HiSeq 2000
Genome Analyzer IIx

Most widely adopted NGS platform
# Genome Analyzer IIx Performance Specifications

<table>
<thead>
<tr>
<th>Performance Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 Gb of high quality data / run</td>
</tr>
<tr>
<td>5 Gb / day</td>
</tr>
<tr>
<td>500 M reads per paired-end run</td>
</tr>
<tr>
<td>2 x 100 bp supported read length</td>
</tr>
<tr>
<td>Raw Accuracy:</td>
</tr>
<tr>
<td>≥ 98% (2 x 100)</td>
</tr>
<tr>
<td>≥ 99% (2 x 50)</td>
</tr>
<tr>
<td>Run Time:</td>
</tr>
<tr>
<td>2 x 100 bp in 9.5 days</td>
</tr>
<tr>
<td>2 x 50 bp in 5 days</td>
</tr>
<tr>
<td>1 x 35 bp in 2 days</td>
</tr>
<tr>
<td>Consensus accuracy 99.999%</td>
</tr>
<tr>
<td>12 to 96 multiplex sequencing/channel</td>
</tr>
</tbody>
</table>
How much can you do with just one lane of GA data?

- 2X Human Genome
- 50X Arabidopsis
- 500X Yeast Genome
- 50X Drosophila
- 1150X E. coli
- 3000X BRCA1+BRCA2, 12 samples per lane
What if, in one sequencing run you could...

**SIMULTANEOUSLY**
Run multiple applications requiring different read lengths.

- Sequence one cancer & one normal genome
- Unravel 20 whole transcriptomes
- Profile 200 gene expression samples
- Analyze two human methylomes

**Whole genome sequencing**
**Targeted resequencing**
**Gene expression**
**Methylation**
**De novo**
**Metagenomics**
**ChIP-seq**
**Whole transcriptome**

In four days
In one week
In less than two days
HiSeq 2000

**OUTPUT**
Initially capable of up to 200 Gb per run

**DATA RATE**
~25 Gb/day
7-8 days for 2 x 100 bp

**NUMBER OF READS**
One billion single-end reads*
Two billion paired-end reads*

*Based on one billion clusters passing filter
HiSeq 2000
Comparison with the Genome Analyzer

<table>
<thead>
<tr>
<th></th>
<th>HiSeq 2000 (at launch)</th>
<th>GA\textsubscript{IIx} (at 50G)</th>
<th>GA\textsubscript{IIx} (at 95G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gb per run</td>
<td>150-200</td>
<td>50</td>
<td>95</td>
</tr>
<tr>
<td>Gb per day</td>
<td>20-25</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Cluster density in KClusters/mm\textsuperscript{2}\textsuperscript{**}</td>
<td>260-350</td>
<td>490</td>
<td>620</td>
</tr>
<tr>
<td>Read length</td>
<td>2 x100</td>
<td>2 x100</td>
<td>2 x150</td>
</tr>
<tr>
<td>Available surface area (mm\textsuperscript{2})\textsuperscript{*}</td>
<td>2880</td>
<td>510</td>
<td>510</td>
</tr>
</tbody>
</table>

\textsuperscript{*}GA\textsubscript{IIx} with single surface, single FC, HiSeq 2000 with dual surface, dual FC

\textsuperscript{**}Clusters passing filter
HiSeq 2000

New flow cell design

LARGER, DUAL-SURFACE ENABLED

>5x increase in imaging area

Retains 8 lane format
HiSeq 2000 dual flow cell design

**TWO INDEPENDENT FLOW CELLS**
Simultaneously run applications that require different read lengths
Run in single or dual flow cell mode

**SIMPLE FLOW CELL LOADING**
Flow cells held by vacuum
No oil needed
LED switch ensures correct connection
Dual surface imaging

*Cutting-edge imaging technology*

TDI line-scanning technology with four CCDs for imaging
Fastest scanning and imaging method

Images clusters grown on both surfaces of flow cell
Huge gain in number of reads and sequence output
The power of line scanning

*Maximizing data rate*

<table>
<thead>
<tr>
<th></th>
<th>Point Imaging</th>
<th>Area Imaging</th>
<th>Line Imaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage &amp; filter movement delays</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Data transfer delays</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Practical data acquisition limit</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Data quality/background rejection</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>
HiSeq 2000

Plug-and-play reagents

PRE-CONFIGURED SEQUENCING REAGENTS
Only two minutes hands-on time
Up to 200 cycles per flow cell
Bar-coded for tracking
Temperature-controlled compartment
Integrated paired-end fluidics
Workflow

SIMPLIFIED SAMPLE PREP

cBot CLUSTER GENERATION

Genome Analyzer SEQUENCING

DATA PROCESSING & ANALYSIS

CATAGATGTCATCTAGTGCATGATCAGTACGATGATCG

CATAGATGTCATCTAGTGCATGATCAGTACGATGATCG

CATAGATGTCATCTAGTGCATGATCAGTACGATGATCG

CATAGATGTCATCTAGTGCATGATCAGTACGATGATCG
Data management and analysis
Instrument computer specifications

**INSTRUMENT CONTROL COMPUTER (HISEQ)**

Base Unit: 2x Intel Xeon X5560 2.8 GHz CPU
Memory: 48 GB RAM
Hard Drive: 4x 1.0 TB 7200 RPM SATA
Operating System: Windows Vista

**DATA ANALYSIS COMPUTER**

HP ProLiant DL580 G5 Rack Server (any 64-bit Unix)
Red Hat Linux
Four quad-core 2.93GHz 64-bit Intel Xeon processors
32 GB fault-tolerant RAM
## Data analysis flow

<table>
<thead>
<tr>
<th>GENERATING SEQUENCING IMAGES</th>
<th>INSTRUMENT PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERFORMING IMAGE ANALYSIS</td>
<td>PRIMARY ANALYSIS</td>
</tr>
<tr>
<td>BASE CALLING</td>
<td>SCS</td>
</tr>
<tr>
<td>cluster positions / intensities / noise</td>
<td></td>
</tr>
</tbody>
</table>

### BASE CALLING
- cluster sequence
- quality calibration
- filtering results

<table>
<thead>
<tr>
<th>DEMULTIPLEXING</th>
<th>LINUX SERVER</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALIGNING TO REFERENCE GENOME</td>
<td>SECONDARY ANALYSIS</td>
</tr>
<tr>
<td>DETECTING VARIANTS AND COUNTING</td>
<td>CASAVA</td>
</tr>
<tr>
<td>expression levels of exons, genes, splice variants</td>
<td></td>
</tr>
</tbody>
</table>

### VIEWING RESULTS
- build consensus sequence
- call SNPs
- detect indels
- count RNA reads

### Any PC
- Genomestudio
The qseq.txt file is tab-delimited, making it easy to parse and import into databases.
Base calling quality score

- A quality score is a prediction of the probability of an error in base calling
  - produced by a model that uses quality predictors as inputs and produces Q-values as outputs

- \( Q = -10 \log_{10} (\text{probability that the base is wrong}) \)
  - \( Q_{40} \): 1 error in 10,000 base calls
  - \( Q_{30} \): 1 error in 1,000 base calls
  - \( Q_{20} \): 1 error in 100 base calls

- The Phred score is a method for assigning quality scores to sequencing data, using numerical predictors of base quality

- Q score are represented as ASCII characters
  - from ASCII to phred = ASCII value + 64

- Why not use the capillary sequencing standard Phred algorithm/predictors?
  - Phred depends crucially on the quality predictors and their statistical distributions
  - good predictors for SBS data are much different than good predictors for capillary sequencing data
Alignment and alignment scoring

- ELAND v2
- reference genome is squashed
- multiseed, gapped alignment allows for detection of indels (<20 bp)
- each candidate position gets a probability
  - Base quality scores and mismatches are used in this calculation
  - Alignment score is expressed on the Phred scale (log odds ratio)
Data quality is assessed by checking a set of metrics and plots

<table>
<thead>
<tr>
<th>Lane info</th>
<th>Tile Mean +/- SD for Lane</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lane</td>
<td>Lane Yield (kbases)</td>
</tr>
<tr>
<td>---</td>
<td>----------------</td>
</tr>
<tr>
<td>1</td>
<td>384339</td>
</tr>
<tr>
<td>2</td>
<td>359280</td>
</tr>
<tr>
<td>3</td>
<td>363356</td>
</tr>
<tr>
<td>4</td>
<td>384594</td>
</tr>
<tr>
<td>5</td>
<td>382373</td>
</tr>
<tr>
<td>6</td>
<td>358367</td>
</tr>
<tr>
<td>7</td>
<td>380393</td>
</tr>
<tr>
<td>8</td>
<td>374541</td>
</tr>
</tbody>
</table>
GenomeStudio
visualization of paired-end reads

From the TPTE gene on Chromosome 21
Applications
Broadest range of customer demonstrated applications

"The Genome Analyzer is enabling our clients to do things that used to be impossible, experiments that they only dreamed of doing, but can do now at a reasonable cost. The Genome Analyzer has completely changed our business."

- Laurent Farinelli, Ph.D., Fasteris
A comprehensive catalogue of somatic mutations from a human cancer genome

Pleasance et al - Nature 2010

"..provides insights into the forces that shape a cancer genome."

"..reveal traces of the DNA damage, repair, mutation and selection processes that were operative years before the cancer became symptomatic"

Method

- combined 2x75bp PE reads and 2x50bp mate pair libraries (2/3/4 kb)
- COLO-829 cancer cell line from a metastasis of a malignant melanoma and COLO-829BL lymphoblastoid line from same patient
- obtained > 40x average haploid genome coverage from COLO-829 and 32-fold from COLO-829BL
The catalogue of somatic mutations in COLO 829

Results

- 33,345 single base substitutions
  - 292 coding
- 1018 small indels
  - 14 coding
- 37 structural rearrangements
  - 34 intrachromosomal
  - 3 interchromosomal
  - 19 breakpoints in genes
- 198 changes in copy number

The sequence and *de novo* assembly of the giant panda genome

Ruiqiang Li *et al*, *Nature* 2010 Jan 21;463(7279):311-7

The Giant Panda lives in bamboo forests high in the mountains of Western China. It eats 12 - 38 kg bamboo per day.

1,600 individuals remained in the wild in 2004.

**Method**
- insert sizes of 150 bp, 500 bp, 2 kb, 5 kb and 10 kb
- generated 176 gigabases of usable sequence (equal to 73x coverage of the whole genome)
- average read length of 52 bp
- assembled short reads using "SOAPdenovo"

**Results**
- genome size 2.40 gigabases
- dietary preferences seem to be related to gut microbiome; genetically speaking the Panda is carnivorous
Ancient DNA sequencing

- DNA isolated from 4000 year old permafrost-preserved hair
- 20x coverage
- provides evidence for a migration from Siberia into the New World some 5,500 years ago, independent of that giving rise to the modern Native Americans and Inuit
The impact of scale in sequencing

$10^4$ scale in throughput; $10^7$ scale in parallelisation in 5 years