Illumina's next generation sequencing technology

Presented by field applications scientist
Pernille Albertus
Denmark/Norway
Illumina

- headquarters 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\(_{\text{IIe}}\)  HiScanSQ  GA\(_{\text{IIx}}\)  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

1. template DNA
2. fragment
3. repair ends
4. add A overhang
5. ligate adaptors &
   purify on gel
6. enrich
7. genomic library &
   library QC
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

clusteration
Annealing of sequencing primer to template

Cluster Amplification → Periodate Linearization → Blocking with ddNTP (⊗) → Denature and Hybridization

SBS3
Summary - "cluster generation"

1. Grafting

2. Hybridization & Amplification

3. Linearization

4. Blocking with ddNTP

5. Denature and Hyb with SBS3

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

- CCD camera collects laser-excited fluorescence.
- Sequencing reaction is temperature controlled.
- Sequencing reagents pass through the 8 lanes inside the flow cell.
Incorporation

1. Incorporation
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
"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

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{IIe}  iScanSQ  GA\textsubscript{IIx}  HiSeq 2000
Genome Analyzer IIx

Most widely adopted NGS platform
## Genome Analyzer™ Performance Specifications

**Performance Parameters**

- 50 Gb of high quality data / run
- 5 Gb / day
- 500 M reads per paired-end run
- 2 x 100 bp supported read length
- **Raw Accuracy:**
  - $\geq 98\%$ (2 x 100)
  - $\geq 99\%$ (2 x 50)
- **Run Time:**
  - 2 x 100 bp in 9.5 days
  - 2 x 50 bp in 5 days
  - 1 x 35 bp in 2 days
- Consensus accuracy 99.999%
- 12 to 96 multiplex sequencing/channel
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
- Analyze two human methylomes
- Profile 200 gene expression samples
- In four days
- In one week
- In less than two days

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

**One Sequencing Run**
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_{IIx} (at 50G)</th>
<th>GA_{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$^2$**</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$^2$)*</td>
<td>2880</td>
<td>510</td>
<td>510</td>
</tr>
</tbody>
</table>

*GA_{IIx} with single surface, single FC, HiSeq 2000 with dual surface, dual FC

**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>

Point Imaging

Scan Area

BeadArray

Illumina Decoding, GA

GA

Illumina Next Gen Decoding,
HiScan, HiSeq 2000

Area Imaging

Scan Area

Line Imaging

Scan Area

Object

Line Scan Camera
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
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

GENERATING SEQUENCING IMAGES

PERFORMING IMAGE ANALYSIS
  cluster positions / intensities / noise

BASE CALLING
  cluster sequence
  quality calibration
  filtering results

DEMULTIPLEXING

ALIGNING TO REFERENCE GENOME

DETECTING VARIANTS AND COUNTING
  expression levels of exons, genes, splice variants

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

INSTRUMENT PC
PRIMARY ANALYSIS
SCS

LINUX SERVER
SECONDARY ANALYSIS
CASAVA

ANY PC
GENOMESTUDIO
# qseq.txt file

- **Tab-delimited**: easy to parse, easy to 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

- \( Q = -10 \log_{10} \text{(probability that the base is wrong)} \)
  - Q40: 1 error in 10,000 base calls
  - Q30: 1 error in 1,000 base calls
  - Q20: 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</th>
<th>Lane Yield (kbases)</th>
<th>Clusters (raw)</th>
<th>Clusters (PF)</th>
<th>1st Cycle Int (PF)</th>
<th>% intensity after 20 cycles (PF)</th>
<th>% PF Clusters</th>
<th>% Align (PF)</th>
<th>Alignment Score (PF)</th>
<th>% Error Rate (PF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>384339</td>
<td>374821 +/- 2479</td>
<td>317111 +/- 2469</td>
<td>500 +/- 15</td>
<td>80.94 +/- 1.82</td>
<td>84.60 +/- 0.52</td>
<td>84.61 +/- 0.25</td>
<td>326.08 +/- 8.17</td>
<td>1.24 +/- 0.46</td>
</tr>
<tr>
<td>2</td>
<td>359280</td>
<td>335506 +/- 4367</td>
<td>296436 +/- 3392</td>
<td>495 +/- 13</td>
<td>80.28 +/- 1.27</td>
<td>88.36 +/- 0.28</td>
<td>99.10 +/- 0.11</td>
<td>520.47 +/- 3.21</td>
<td>0.68 +/- 0.10</td>
</tr>
<tr>
<td>3</td>
<td>363356</td>
<td>351148 +/- 12389</td>
<td>299799 +/- 6395</td>
<td>476 +/- 17</td>
<td>78.23 +/- 1.37</td>
<td>85.42 +/- 1.34</td>
<td>98.47 +/- 0.00</td>
<td>460.72 +/- 11.11</td>
<td>1.28 +/- 0.36</td>
</tr>
<tr>
<td>4</td>
<td>384594</td>
<td>374144 +/- 21316</td>
<td>317322 +/- 12469</td>
<td>515 +/- 17</td>
<td>80.03 +/- 2.38</td>
<td>84.90 +/- 1.75</td>
<td>84.47 +/- 0.27</td>
<td>322.65 +/- 4.62</td>
<td>1.20 +/- 0.16</td>
</tr>
<tr>
<td>5</td>
<td>382373</td>
<td>377654 +/- 10799</td>
<td>315489 +/- 8581</td>
<td>494 +/- 21</td>
<td>79.81 +/- 1.09</td>
<td>83.54 +/- 0.51</td>
<td>84.51 +/- 0.28</td>
<td>322.09 +/- 4.96</td>
<td>1.22 +/- 0.19</td>
</tr>
<tr>
<td>6</td>
<td>358367</td>
<td>357009 +/- 11548</td>
<td>295682 +/- 6826</td>
<td>451 +/- 27</td>
<td>78.09 +/- 1.85</td>
<td>82.85 +/- 1.09</td>
<td>98.46 +/- 0.28</td>
<td>461.38 +/- 11.17</td>
<td>1.27 +/- 0.41</td>
</tr>
<tr>
<td>7</td>
<td>380393</td>
<td>375250 +/- 13209</td>
<td>313856 +/- 7465</td>
<td>494 +/- 18</td>
<td>79.75 +/- 1.93</td>
<td>83.67 +/- 1.14</td>
<td>84.69 +/- 0.10</td>
<td>324.06 +/- 5.23</td>
<td>1.19 +/- 0.19</td>
</tr>
<tr>
<td>8</td>
<td>374541</td>
<td>372954 +/- 3179</td>
<td>309027 +/- 3476</td>
<td>535 +/- 83</td>
<td>78.60 +/- 2.66</td>
<td>82.86 +/- 0.53</td>
<td>84.65 +/- 0.00</td>
<td>321.03 +/- 8.67</td>
<td>1.24 +/- 0.30</td>
</tr>
</tbody>
</table>
GenomeStudio
visualization of paired-end reads

From the TPTE gene on Chromosome 21
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

Ancient human genome sequence of an extinct Palaeo-Eskimo

Morten Rasmussen1,2,3, Yingrui Li2,3, Stinus Lindgreen1,2,4, Jakob Skou Pedersen4, Anders Albrechtsen1, Ida Moltke1, Mait Metspalu5, Ene Metspalu5, Toomas Kivisild5,6, Ramnesh Gupta7, Marcelo Bertalan7, Kasper Nielsen1, M. Thomas P. Gilbert1,2, Yong Wang8, Maanasa Raghavan1,9, Paula F. Campos1, Hanne Munkholm Kamp1,4, Andrew S. Wilson10, Andrew Gledhill11, Silvana Tridico12, Michael Bunce12, Eline D. Lorenzen1, Jonas Binladen1, Xiaosen Guo2,3, Jing Zhao2,3, Xiuxing Zhang2,3, Hao Zhang2,3, Zhou Li2,3, Minfeng Chen2,3, Ludovic Orlando13, Karsten Kristiansen3,4, Mads Bak14, Niels Tommerup14, Christian Bendixen15, Tracey L. Pierre16, Bjarne Grennow17, Morten Meldgaard18, Claus Andreassen19, Sardana A. Fedorova20, Ludmila P. Osiptova21, Thomas F. G. Higham2, Christopher Bronk Ramsey2, Thomas v. O. Hansen22, Finn C. Nielsen23, Michael H. Crawford2, Søren Brunak24, Thomas Sicheritz-Ponten1, Richard Villems5, Rasmus Nielsen1,4, Anders Krogh3,4, Jun Wang2,3,4 & Eske Willerslev1,2

We report here the genome sequence of an ancient human. Obtained from ~4,000 year-old permafrost-preserved hair, the genome represents a male individual from the first known culture to settle in Greenland. Sequenced to an average depth of 20x, we recover 79% of the diploid genome, an amount close to the practical limit of current sequencing technologies. We identify 353,151 high-confidence single-nucleotide polymorphisms (SNPs), of which 6.8% have not been reported
The impact of scale in sequencing

Gb / run

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