An introduction to “Next Generation” DNA Sequencing

Jakob Hedegaard
Post doc
Dept. of Genetics and Biotechnology
Molecular Genetics and Systembiology
Blichers Allé 20, P.O. BO 50, DK-8830 Tjele
Researchcentre Foulum
Tlf (+45)89991363
E-mail: Jakob.Hedegaard@agrsci.dk
<table>
<thead>
<tr>
<th>Day</th>
<th>a.m.</th>
<th>p.m.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td></td>
<td><strong>NGS intro</strong> (RNA to sequences)</td>
</tr>
<tr>
<td>Tuesday</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wednesday</td>
<td><strong>NGS, pre-processing</strong> (Sequences to counts)</td>
<td><strong>RNA- seq analysis</strong> (Counts to list(s))</td>
</tr>
<tr>
<td>Thursday</td>
<td><strong>Sequence based analysis</strong></td>
<td><strong>Supervised and unsupervised learning</strong></td>
</tr>
<tr>
<td>Friday</td>
<td><strong>Gene set analysis</strong></td>
<td></td>
</tr>
</tbody>
</table>
RNA to sequences

The aims of today are to understand (or get an idea about):

• the potential use of NGS technology
• the basic principles of the Illumina technology
• the sample prep methods for mRNA-Seq (and RNA-Seq)
• the options for experimental design for mRNA-Seq
• the basic principles of processing the raw data from an Illumina GA
• the output and format of the data produced by an Illumina GA

➢ Case study
➢ Software check

➢ Questions and breaks!
DNA Sequencing technology

Pharmacia ALF DNA Sequencer

Sanger method (chain-terminator method)
- Fluorescent labeled primer
- 4 reactions/sample

- 10 sequences of 3-500 bp
- sequence preparation and ON run
Short history of DNA Sequencing

1977
– Maxim-Gilbert
– Sanger

1986
– First Automated DNA Sequencer ABI 370 (373)

1988
– Pharmacia ALF

1995
– ABI 377
Up to 96 lanes

1996
– First Capillary DNA Sequencer ABI 310

1998
– First 96 Capillary instruments
MegaBace, ABI 3700

2000
– ABI 3100, 16 Capillary

2002
– ABI 3730, 48 or 96 Capillary

2005
– Genome Sequencer GS20 (454)

2006
– Solexa (Illumina)

2007
– SOLiD

2011
– ?
## Short history of DNA Sequencing

<table>
<thead>
<tr>
<th>Method</th>
<th>Read length (max bp)</th>
<th>Sequences per run</th>
<th>Run time</th>
<th>Output/day</th>
<th>Output/run</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dye-Terminator (3730xl)</td>
<td>1,500</td>
<td>96</td>
<td>1 hr</td>
<td>~1-2 Mbp</td>
<td>~70 Kbp</td>
</tr>
<tr>
<td>454/Roche (GS FLX)</td>
<td>400</td>
<td>~1 million</td>
<td>10 hrs</td>
<td>~1 Gbp</td>
<td>~400 Mbp</td>
</tr>
<tr>
<td>Illumina (GA)</td>
<td>150</td>
<td>~300 millions</td>
<td>14 days (2x150 bp)</td>
<td>~7 Gbp</td>
<td>~90 Gbp</td>
</tr>
<tr>
<td>Illumina (HiSeq)</td>
<td>100</td>
<td>~1 billion</td>
<td>8 days (2x100 bp)</td>
<td>~25 Gbp</td>
<td>~200 Gbp</td>
</tr>
<tr>
<td>Applied (SOLiD)</td>
<td>75</td>
<td>~2.8 billions</td>
<td>7 days (60x60 pb)</td>
<td>~25 Gbp</td>
<td>~200 Gbp</td>
</tr>
</tbody>
</table>
NGS – also known as...

- Next generation sequencing (NGS)
- High-throughput sequencing (HTS or HT-Seq)
- Flow cell sequencing (FCS)
- Massively parallel sequencing (MPS)
- Deep sequencing
- Many other synonyms!
NGS technologies

Applied Biosystems
5500xl SOLiD

Roche/454
Genome Sequencer FLX

Illumina
HiSeq 2000

Illumina
Genome Analyser IIx

Additional technologies – and more to come!
Illumina Sequencing technology

<table>
<thead>
<tr>
<th>Applications</th>
<th>Library Generation (hours-days)</th>
<th>Cluster Generation (~5 hours)</th>
<th>Sequencing (days, 1.1 hrs/cycle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole genome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted genome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long PCR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pull-down</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epigenome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bisulfite</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Restriction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Restriction + bisulfite</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antibody pull-down</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ab. + bisulfite</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole transcriptome</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All RNA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNA minus rRNA, tRNA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PolyA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ncRNA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene expression (DGE)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small RNA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RISC RNA products</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein:DNA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein:RNA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DNA Fragments
- + Adapters
  ~200 bp
Illumina Sequencing Technology
Reversible Terminator Chemistry

DNA (0.1-1.0 ug)

Sample preparation

Single molecule array
Cluster growth

Sequencing

Image acquisition

Base calling
A **flow cell** contains eight lanes.

Each lane contains **two columns** of tiles.

Each column contains **multiple tiles** – total 120.

Each tile is imaged four times per cycle – one image per base.

- \(~340,000\) clusters/tile \(\rightarrow\)
- \(~40,000,000\) clusters/lane \(\rightarrow\)
- \(~320,000,000\) clusters/flowcell
# Illumina GA output

## SYSTEM SPECIFICATIONS

<table>
<thead>
<tr>
<th>READ LENGTH</th>
<th>RUN TIME (DAYS)</th>
<th># OF READS (PER FLOW CELL)</th>
<th>HIGH-QUALITY OUTPUT (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 x 35 bp</td>
<td>~2</td>
<td>138–168 million</td>
<td>4.5–6.0</td>
</tr>
<tr>
<td>2 x 35 bp</td>
<td>~4</td>
<td>276–336 million</td>
<td>9.5–11.5</td>
</tr>
<tr>
<td>2 x 50 bp</td>
<td>~5</td>
<td>276–336 million</td>
<td>13.5–16.5</td>
</tr>
<tr>
<td>2 x 75 bp</td>
<td>~7.5</td>
<td>276–336 million</td>
<td>20.5–25.0</td>
</tr>
<tr>
<td>2 x 100 bp</td>
<td>~9.5</td>
<td>276–336 million</td>
<td>27.5–33.0</td>
</tr>
</tbody>
</table>

**Throughput**: eight channels per flow cell, up to 12 samples per instrument per day

**Input requirement**: 0.1–1.0 μg (single- and paired-end reads)

**Genomic DNA sample prep**: 3 hours hands-on, 6 hours total

---

**Genome Analyzer 1lx Performance Parameters**

<table>
<thead>
<tr>
<th>Read Length</th>
<th>Run Time (Days)</th>
<th>Output (Gb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 x 35 bp</td>
<td>~2</td>
<td>10 – 12</td>
</tr>
<tr>
<td>2 x 50 bp</td>
<td>~5</td>
<td>25 – 30</td>
</tr>
<tr>
<td>2 x 75 bp</td>
<td>~7</td>
<td>37.5 – 45</td>
</tr>
<tr>
<td>2 x 100 bp</td>
<td>~9.5</td>
<td>54 – 60</td>
</tr>
<tr>
<td>2 x 150 bp</td>
<td>~14</td>
<td>85 – 96</td>
</tr>
</tbody>
</table>

*Sequencing output generated using TruSeq SBS V5 kit with PhiX library and cluster densities between 506,000–630,000 clusters/mm\(^2\) that pass filtering on a GA\(\text{nx}\).

**Throughput**

Up to 6.5 Gb per day for a 2 x 100 bp run

**Reads**

Up to 320 million cluster passing filter and up to 640 million paired-end reads

**Performance**

The Genome Analyzer 1lx generates a significant yield of bases greater than Q30

- Greater than 90% bases higher than Q30 at 2 x 50 bp
- Greater than 85% bases higher than Q30 at 2 x 100 bp

---

1. gDNA sequencing output generated with cluster densities between 506,000–630,000 clusters/mm\(^2\) that pass filtering on a GA\(\text{nx}\).
2. 2 x 100 bp reads supported.
3. Data generated from clusters that pass Pipeline software v1.5 quality controls.
Illumina GA Sequencing technology

**Pipeline Computer  
"Donkey"**
DELL PowerEdge 2900
8 cores (2 x 4 CPU, 2.6 GHz)
16 GB RAM
4.7 TB hard drive

- **Storage**
- **Analysis**

**Switch**

- 100 Mbps

**GAIIx, “Oban”**

**Cluster Station**

**GAIIx, “Lagavulin”**

Illumina GA Sequencing technology
Illumina GA Sequencing technology

Switch and 100 Mbps network to pipeline computer

Paired End (PE) module

GA PC
- 2.66 GHz cpu
- 3 GB RAM
- 80 GB hard drive

Cooling unit!

Genome Analyzer (GAIIx)

Uninterruptible Power Supply (UPS)
- Back up for ~10 min

Illumina Genome Analyzer IIx, “Oban”
Illumina GA Sequencing technology

- Pieltter element
- Liquid outlet
- Liquid inlet
- Camera
- Laser
- Flowcell & Prism
Illumina Sequencing technology

Single-end

Paired-end

Multiplex
Illumina Sequencing technology

**Single End**
- Genomic DNA
- Fragment (200–500 bp)
- Ligate Adaptors
- Generate Clusters
- Sequence

**Paired End**
- Genomic DNA
- Fragment (200–500 bp)
- Ligate Adaptors
- Generate Clusters
- Sequence First End
- Regenerate Clusters and Sequence Paired End
Illumina Sequencing technology

Mate Pair

1. Genomic DNA
2. Fragment (2-5 kb)
3. Biotinylate Ends
4. Circularize
5. Fragment (400-600 bp)
6. Enrich Biotinylated Fragments
7. Ligate Adaptors
8. Generate Clusters
9. Sequence First End
10. Regenerate Clusters and Sequence Paired End
SR and PE flowcells

Single Read
Periodate Linearization

Paired End
Linearization 1 Enzyme
Linearization 2 Enzyme
Single-Read flowcells

Overview: Standard method for single reads

1. Grafted flowcell

2. Template hybridisation

3. Initial extension

4. Denaturation

5. Annealing

6. Extension

n=35
Single-Read flowcells

Overview: Standard method, early cycles of amplification

1. 1st cycle denaturation
2. 1st cycle annealing
3. 1st cycle extension
4. 2nd cycle denaturation
5. 2nd cycle annealing
6. 2nd cycle extension

n=35 total
Single-Read flowcells

Overview: Processing clusters for sequencing

1. Grafting

2. Cluster amp

3. Periodate Linearisation

4. Blocking with ddNTP

5. Denature and hyb

Sequencing on Genome Analyzer
Single-Read flowcells

Cluster Station

Amplification

Linearization

Blocking

Primer hybridization

Sequencing – single read

GAI, “Oban”
Paired-End flowcells

Cluster Generation: Initial Extension

- Grafted flowcell
- Template hybridization
- Initial extension
- Denaturation
Paired-End flowcells

Cluster Generation: Amplification

1st cycle denaturation → 1st cycle annealing → 1st cycle extension → 2nd cycle denaturation

n=35 total

2nd cycle extension → 2nd cycle annealing
Paired-End flowcells

Cluster Generation:

Cluster Amplification → P5 Linearization → Block with ddNTPS → Denaturation and Sequencing Primer Hybridization
Paired-End flowcells

Sequencing

Denaturation and Hybridization

Sequencing First Read

Denaturation and De-Protection

Resynthesis of P5 Strand

Sequencing Second Read

Denaturation and Hybridization

Block with ddNTPs

P7 Linearization
Paired-End flowcells

Amplification
↓
Linearization
↓
Blocking
↓
Primer hybridization
↓
Sequencing – read 1
↓
Denaturation + de-protection
↓
Re-synthesis

Sequencing – read 2
↑
Primer hybridization
↑
Blocking
↑
Linearization

Cluster Station

GAI, “Oban”
Using NGS for expression profiling: mRNA-Seq

**Digital gene expression**
- Sequencing of tags
- Mapping tags to transcriptome -> counts
- Counts, 0 - ∞
- All genes/transcripts
- Added information (Alternative transcripts, SNPs, novel genes,...)
RNA-Seq – library preparation

- Coding
  - PolyA mRNA
  - Non-PolyA mRNA
- Structural
  - DNA associated
  - RNA associated
  - Ribosome associated
- Regulatory
  - MicroRNA
    - Transcriptional start site associated
    - Anti-sense
    - Enhancer RNA
- Non-coding
- Replisome
- DNA Repair
- Telomeric
- DNA methylation
- rRNA
mRNA-Seq – library preparation

1. Total RNA (10 µg)
2. Purify and fragment mRNA
3. ds cDNA synthesis (random primed)
4. Repair ends
5. Add ‘A’ bases to 3’ ends
6. Ligate adapters
7. Purify ligation product (200-500 bp)
8. PCR amplification
RNA – library preparations

**mRNA-Seq**
- Purified Total RNA
- Poly-A Selection
- RNA Fragmentation
- cDNA Synthesis
- Adapter Ligation
- PCR

**TotalRNA-Seq I**
- Purified Total RNA
- RNA Fragmentation
- cDNA Synthesis
- Adapter Ligation
- PCR

**TotalRNA-Seq II**
- Purified Total RNA
- Removal of rRNA
- RNA Fragmentation
- cDNA Synthesis
- Adapter Ligation
- PCR

Normalization
- e.g. DSN to remove very abundant elements
RNA – library preparations

**smallRNA-Seq**
- Purified Total RNA
  - PAGE Selection
  - RNA Adapter Ligation
  - 1. strand cDNA Synthesis
    - PCR
    - PAGE Selection

**Directional mRNA-Seq**
- Purified Total RNA
  - Poly-A Selection
  - RNA Fragmentation
  - RNA Adapter Ligation
    - 1. strand cDNA Synthesis
      - PCR
RNA – library preparations

Confused?

More variations:
• Multiplexing/indexing (more than 1 sample/lane)
• PCR free protocols
• Robot compatible protocols
• Tagged RT primer + terminal tagged oligo + index-PCR (Epicentre)
• Amplification based methods for samples with limited RNA
• ...... more to come!
mRNA-Seq – library preparation

1. Total RNA (10 µg)
2. Purify and fragment mRNA
3. ds cDNA synthesis (random primed)
4. Repair ends
5. Add ‘A’ bases to 3’ ends
6. Ligate adapters
7. Purify ligation product (200-500 bp)
8. PCR amplification
Illumina - multiplex

Multiplexed Sequencing on the Genome Analyzer

Multiplexed paired-end sequencing utilizes three sequencing reads as described in Figure 4.

Figure 4
Three-Read Multiplexed Sequencing
Intensity versus Cycle

- Genomic DNA
- mRNA (RNA-Seq)
- "random" priming?
- Small-RNA

Array vs NGS

- Hybridization to probes
- Intensities, 0 – N (analog)
- Fixed gene-set
- Proven technology
- Costs ?
- ......

- Sequencing of tags
- Counts, 0 - ∞ (digital)
- All genes/transcripts
- Added information
- Costs ?
- ......
Array vs NGS

**mRNA-Seq data is information rich**

- mRNA Expression Profiling
- Alternative Splicing Analysis
- Analysis of expressed SNPs and mutations
- Analysis of Allelic-specific Expression
- Chimeric Transcript Discovery
- Gene Discovery and Annotation
- ......

(Module 11, Thursday: Sequence based analysis)
Array vs NGS

As in the early days of microarray technology....

• The NGS technology is still under development
• The library prep methods (technology, biases, robot assisted, ......)
• Development of analytical tools and methods
• Education and training to handle the data
• ......
Experimental design (mRNA-Seq)

Objective(s) of the study -> number of reads needed/sample

- **Counting/profiling**: 5-10 million reads/sample
- **Alternative splicing**: 50-100 million reads/sample
- **Novel discovery**: >100 million reads/sample

**Multiplexing?**
- Illumina GAIIx: 35-40 million reads/lane
- Done after 1 lane? <-> repeat in several lanes/flowcells(runs)?

**Replication?**
- Biological vs. technical

**Batch effects**
- RNA purification
- library preparation (and method)
- flowcell/sequencing run
- chemistry batches
- ?
Experimental design (mRNA-Seq)

- Replication
- Randomization
- Blocking

Statistical Design and Analysis of RNA Sequencing Data

Paul L. Auer and R. W. Doerge

Department of Statistics, Purdue University, West Lafayette, Indiana 47907
Manuscript received January 31, 2010
Accepted for publication March 15, 2010

ABSTRACT

Next-generation sequencing technologies are quickly becoming the preferred approach for characterizing and quantifying entire genomes. Even though data produced from these technologies are proving to be the most informative of any thus far, very little attention has been paid to fundamental design aspects of data collection and analysis, namely sampling, randomization, replication, and blocking. We discuss these concepts in an RNA sequencing framework. Using simulations we demonstrate the benefits of collecting replicated RNA sequencing data according to well known statistical designs that partition the sources of biological and technical variation. Examples of these designs and their corresponding models are presented with the goal of testing differential expression.

NEXT-GENERATION sequencing (NGS) has emerged as a revolutionary tool in genetics, genomics and epigenomics.Bringing together throughput and certain size specification (e.g., 200-300 bases long) are retained for amplification using polymerase chain reaction (PCR). After amplification, the cDNA is
Data processing
Data processing

- Generating Sequencing Images
  - Performing Image Analysis
    - Cluster positions
    - Cluster intensities
    - Cluster noise
  - Base Calling
    - Cluster sequence
    - Quality calibration
    - Filtering results
- .bcl Conversion
- Demultiplexing
  - Separate multiplexed sequence runs by index
- Aligning
  - Align to reference genome
- Analysis System
  - HiSeq, Genome Analyzer and Real Time Analysis (RTA)
  - bcl to qseq converter
  - CASAVA demultiplex.pl script
  - CASAVA GERALD module

- Illumina software
- Alternative software
Analysis

1. GA-PC (SCS software, RTA)
   - image analysis
   - base calling

2. Pipeline Computer, “Donkey”
   - sequence analysis alignment to reference
     produces fastq files
Data - structure

Real Time Analysis (RTA) output
Data - structure

Disk space use
- Depends on the run (SE, PE, #sequences)
- Images are deleted
- Complete run folder ~ 1-2 TB (no images!)
- Minimum data set ~ 20-40 GB
- More space needed for analyzing the data!

BCL Converter output
Data, the qseq files

1 x _qseq.txt file/tile/lane -
- single-end: 960 _qseq.txt files (120x8)
- paired-end: 1,920 _qseq.txt files (120x8x2)
- paired-end, multiplex: 2,880 _qseq.txt files (120x8x3)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>O 11 1 20 18235 1120 0 1</td>
<td>GGC CGC GGC GGC . . . GAG GGG GGG G</td>
<td>BBBBB BBBBB BBBBB BBBBB BBBBB 0</td>
<td></td>
</tr>
<tr>
<td>O 11 1 20 18284 1125 0 1</td>
<td>ATG CC ACC CGC . . . ATG A GAATA</td>
<td>BBBBB BBBBB BBBBB BBBBB BBBBB 0</td>
<td></td>
</tr>
<tr>
<td>O 11 1 20 18449 1128 0 1</td>
<td>ACT TG AAATA . . . ATTT GGA TAT</td>
<td>`BBBB BBBBB BBBBB BBBBB BBBBB BBBBB 0</td>
<td></td>
</tr>
<tr>
<td>O 11 1 20 18499 1123 0 1</td>
<td>CT TAT TCG AGG . . . GTG A TGA AG</td>
<td>e_cS becbe . . . BBBBB BBBBB BBBBB BBBBB 0</td>
<td></td>
</tr>
<tr>
<td>O 11 1 20 18546 1128 0 1</td>
<td>CGC AT AT CGT . . . TG A CAT GAA</td>
<td>[XUY KOMM . . . BBBBB BBBBB BBBBB BBBBB 0</td>
<td></td>
</tr>
<tr>
<td>O 11 1 20 18607 1120 0 1</td>
<td>CC C AT TG CGA . . . TAA T GTG A TC</td>
<td>SRS R [Q W . . . BBBBB BBBBB BBBBB 1</td>
<td></td>
</tr>
<tr>
<td>O 11 1 20 18669 1125 0 1</td>
<td>CTT TT TAC GCA . . . AG A CT GGA AA</td>
<td>c_Sc ca [] cc . . . BBBBB BBBBB BBBBB 1</td>
<td></td>
</tr>
<tr>
<td>O 11 1 20 18783 1121 0 1</td>
<td>CCG GT GTG AT . . . AG TA A AT GGA</td>
<td>cccc R <code>Za</code>a . . . ^ ^ ZY b ^ Hb ^ 1</td>
<td></td>
</tr>
</tbody>
</table>

**A** Read number (1 or 2 for paired end runs; 1,2 or 3 for indexed paired-end)

**B** Sequence

**C** Quality string

**D** Filtering (1 for pass and 0 for failed)
Data, the fastq files

Conversion of qseq -> fastq
1 x fastq file/lane/read ->
• single-end: 8 x fastq files (s_4_sequence.txt)
• paired-end: 16x fastq files(s_4_1_sequence.txt, s_4_2_sequence.txt)
• paired-end, multiplex: as paired-end (index included in fastq)

Fastq: same information as in qseq files, but
• different format
• 4 lines/sequence -> .txt files with 4x~40 millions lines
• only sequences passing the signal purity filter

Pass: the second lowest chastity score in the first 25 cycles of a read must be greater than 0.6

Chastity is the ratio of the highest of the four (base type) intensities to the sum of highest two

Further quality (phred) based filtration can be applied on the fastq files
Data, the fastq files

Fastq file

```
@L_0023:1:1:16103:1200#0/1
TTATGTGTTTATTACGTNTTTG......AATGTGTTTATTACGGGTATTTTATTTA
+L_0023:1:1:16103:1200#0/1
hhhhhhghhhhhghgdeBee_...... hfhghhhgghghfffaehghhcX
@L_0023:1:1:16318:1210#0/1
TGTGAAAGAATTGTTAGTTGGGA...... TTAGGTGTTTATATGTTAAATT
+L_0023:1:1:16318:1210#0/1
Y_fXfdfafc`a_g_g_gdd\af...... ffddfcffffffffgccccccccccff
```

qseq file

```
O 11 1 20 18235 1120 0 1 GGCUGCUGGC......GAGGGGGGGG BBBBBBBBBBB....BBBBBBBBBBB 0
O 11 1 20 18284 1125 0 1 ATGCCACCGC......GTTGAATTA BBBBBBBBBBB....BBBBBBBBBBB 0
O 11 1 20 18449 1128 0 1 ACTTGAATA......ATTTGGATAT `BBBBBBBBBB....BBBBBBBBBBB 0
O 11 1 20 18499 1123 0 1 CTTATTCAGG......ATTTGGATAT `]XUYKOMM....BBBBBBBBBB 0
O 11 1 20 18546 1128 0 1 CGCCATCTCG......TGACTGAGA e_cS]becbe....BBBBBBBBBBB 1
O 11 1 20 18607 1120 0 1 CCCATTCGCA......TAATGGTATC SRSRQ[QW\.....BBBBBBBBBBB 1
O 11 1 20 18669 1125 0 1 CTTTTACGCA......AGACTGAAA c_Scca[]cc....BBBBBBBBBBB 1
O 11 1 20 18783 1121 0 1 CCGGTGTGAT......AGTTAATGGA ccccR`Za`a....^^ZYb^]Hb^ 1
```
Data, the fastq files

@L_0023:1:1:16103:1200#0/1
TTATGTGTATTACGGTTNTTTG......AATGTTTATTTACGGTTATTTA
+L_0023:1:1:16103:1200#0/1
hhhhhhghhhhhhhghdeeBee_...... hfhhghhghhghfffaehghhcX
@L_0023:1:1:16318:1210#0/1
TGTGAAAGAATTGTTTAGTGGGA...... TTAGGTGTTTATATGTTAATT
+L_0023:1:1:16318:1210#0/1
Y_fXfdfafc`a_g_g_gdd\af...... ffddfcfffff\gggfffcff

@L_0017:1:1:7461:1105#CGATGT/1
GGGTGATCATTTAATATTNCCGCCGTGCAGGTAGCTAATCAACTAAATA
+L_0017:1:1:7461:1105#CGATGT/1
cdddaa`c[addccda`\Bba^aZ`\UR_a^K[\c^a]Ldd\Wad`\]
@L_0017:1:1:15703:1104#CGATGT/1
CTGCTTTCAGATGCTGTCTTTTTTTATTAGTGATGATGATGTTTGCTAAT
+L_0017:1:1:15703:1104#CGATGT/1
hfhfhhhhhhhhfhfhhdbhBbfffdb`h_cecf_bacdeedegfffcaca]
Data, the fastq files

The Illumina fastq header - sequence

Paired-end read:
@L_0020:4:4:4:1052#0/1
@L_0020:4:4:4:1052#0/2

Indexed read:
@L_0001:3:1:4:1303#GATCAG/1

L_0001     unique instrument name and run number
3          flowcell lane (1-8)
1          tile number within the flowcell lane (1-120)
4          'x'-coordinate of the cluster within the tile
1303       'y'-coordinate of the cluster within the tile
#0         index sequence for a multiplexed sample (0 for no indexing)
/1         the member of a pair, /1 or /2 (paired-end)
Data, the fastq files

The Illumina fastq header – quality score

+L:4:4:4:1052#0/1

abbbb`ba]aYD[``aa[\[_ZSVYLWOZ`aYTWKTUXR[ZTMZYNIZ^OWY_`aaa]LVNYab_BBBBBBBBBB

Quality score are represented as ASCII characters (to save space)
– One ASCII character per base

To get Phred score:

<table>
<thead>
<tr>
<th>ASCII value</th>
<th>Phred Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>−64</td>
<td>−64</td>
</tr>
</tbody>
</table>

Sanger quality scores use the same principle
– Same as a Phred score but the ASCII score calculation is different

<table>
<thead>
<tr>
<th>ASCII value</th>
<th>Sanger Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>−33</td>
<td>−33</td>
</tr>
</tbody>
</table>
Phred quality scores

<table>
<thead>
<tr>
<th>Phred score</th>
<th>$P_e$</th>
<th>Accuracy of the base call</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90 %</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99 %</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1,000</td>
<td>99.9 %</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10,000</td>
<td>99.99 %</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100,000</td>
<td>99.99 %</td>
</tr>
</tbody>
</table>

Phred quality score:

\[ QV = -10 \cdot \log_{10}(P_e) \]

where $P_e$ is the probability that the base call is an error.

<table>
<thead>
<tr>
<th>Character</th>
<th>ASCII value</th>
<th>Phred score</th>
</tr>
</thead>
<tbody>
<tr>
<td>^</td>
<td>94</td>
<td>30</td>
</tr>
<tr>
<td>_</td>
<td>95</td>
<td>31</td>
</tr>
<tr>
<td>”</td>
<td>96</td>
<td>32</td>
</tr>
<tr>
<td>a</td>
<td>97</td>
<td>33</td>
</tr>
<tr>
<td>b</td>
<td>98</td>
<td>34</td>
</tr>
<tr>
<td>c</td>
<td>99</td>
<td>35</td>
</tr>
<tr>
<td>d</td>
<td>100</td>
<td>36</td>
</tr>
<tr>
<td>e</td>
<td>101</td>
<td>37</td>
</tr>
<tr>
<td>f</td>
<td>102</td>
<td>38</td>
</tr>
<tr>
<td>g</td>
<td>103</td>
<td>39</td>
</tr>
</tbody>
</table>
QC - FASTX-Toolkit (Hannonlab)

quality vs cycle
QC - FASTX-Toolkit (Hannonlab)

Nucleotide distribution vs cycle

Genomic DNA

mRNA-Seq
(“random” priming!)
Methylation: Paired-end sequencing of bisulfite treated gDNA

- bisulfite treatment converts C to U (T), except for 5-methylcytosine
Analysis......
Discussions (support)

SeqAnswers (http://seqanswers.com/forums/index.php)
# How much does it cost?

<table>
<thead>
<tr>
<th>Project</th>
<th>Description</th>
<th>Cost</th>
</tr>
</thead>
</table>
| **Project A:** | 5 human genomes ~20x  
• 5x PE library of gDNA  
• 5x PE runs, 2x 101 bp | **Total cost:** $ 66.000 (+) |
| Library preparation | ~ $ 2000 |  
| Cluster kit (flowcell) | 5x Paired-end: $ 22.000 |  
| Sequence kits | 30x 36 cycles: $ 42.000 |  
| **Project B:** | 84 mRNA-Seq profiles  
• 84x RNA-Seq libraries (7x12 plex)  
• 3x SE runs, 50+7 bp | **Total cost:** $ 41.000 (+) |
| Library preparation | ~ $ 25.000 |  
| Cluster kit (flowcell) | 3x Single-read: $ 7.600 |  
| Sequence kits | 6x 36 cycles: $ 8.400 |  

(+): Additional costs may apply.
Summary

The aims of today were to understand (or get an idea about):

• the potential use of NGS technology
• the basic principles of the Illumina technology
• some sample prep methods for mRNA-Seq (and RNA-Seq)
• the options for experimental design for mRNA-Seq
• the basic principles of processing the raw data from an Illumina GA
• the output and format of the data produced by an Illumina GA

➢ Case study