Overview and pre-processing of Next Generation Sequencing data

Neerja Katiyar
Contents

• Overview of NGS workflow
• Illumina sequencing technology
• HiSeq, MiSeq, NextSeq
• Demultiplexing
• High Throughput Sequencing website
• FASTQC
• PacBio
• Applications of NGS technology
a | Spike-in controls can be added to DNA or RNA samples for combined library preparation and sequencing

b | Well-characterized biological reference materials act as valuable process controls but cannot be directly added to samples.

c | Finally, in silico sequencing libraries can be used to rapidly evaluate key bioinformatics steps (indicated in blue).

Reference standards can be used to assess various biases and errors in the next-generation sequencing workflow (indicated in beige). rDNA, ribosomal DNA.

Reference standards for next-generation sequencing, Nature 2016 Simon A. Hardwick1,2, Ira W. Deveson1,3 and Tim R. Mercer
Applications of next generation sequencing in fish ecotoxicogenomics

Alvine C. Mehinto, Christopher J. Martyniuk, Daniel J. Spade and Nancy D. Denslow

Non-model fish species
- Dab
- Largemouth bass
- Medaka
- Stickleback
- Fathead minnow

Molecule of interest
- DNA
- RNA
- miRNA

Sequencing strategy
- 454 GS FLX
- Illumina
- Ion Torrent
- PacBio
- SOLID

Sequence assembly/gene annotation
- ABySS
- Newbler
- PTA
- Blast2GO

Bioinformatics tools
- ARACNE
- Cytoscape
- GO, GSEA
- KEGG
- IPA, Pathway Studio

Research questions
- Transcriptome profiling
- Toxicant-specific biomarkers
- Toxicity pathways
- AOPs
- Toxicant-resistant population (e.g., SNPs)

Applications of next-generation sequencing in fish ecotoxicogenomics
Alvine C. Mehinto, Christopher J. Martyniuk, Daniel J. Spade and Nancy D. Denslow
Sequencing by synthesis
Illumina sequencing

HiSeq2500

NextSeq500

MiSeq
## HiSeq 2500

### High-Output Run Mode

<table>
<thead>
<tr>
<th>Read Length</th>
<th>Dual Flow Cell</th>
<th>Single Flow Cell</th>
<th>Dual Flow Cell Run Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 × 36</td>
<td>129–144 Gb</td>
<td>64–72 Gb</td>
<td>29 hours</td>
</tr>
<tr>
<td>2 × 50</td>
<td>360–400 Gb</td>
<td>180–200 Gb</td>
<td>2.5 days</td>
</tr>
<tr>
<td>2 × 100</td>
<td>720–800 Gb</td>
<td>360–400 Gb</td>
<td>5 days</td>
</tr>
<tr>
<td>2 × 125*</td>
<td>900 Gb–1 Tb</td>
<td>450–500 Gb</td>
<td>6 days</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reads Passing Filter*</th>
<th>Up to 4 billion</th>
<th>Up to 2 billion</th>
</tr>
</thead>
</table>

| Quality               | ≥ 95% of bases above Q30 at 2 × 50 bp | ≥ 80% of bases above Q30 at 2 × 100 bp | ≥ 80% of bases above Q30 at 2 × 125 bp |

### Rapid-Run Mode

<table>
<thead>
<tr>
<th>Read Length</th>
<th>Dual Flow Cell</th>
<th>Single Flow Cell</th>
<th>Dual Flow Cell Run Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 × 36</td>
<td>18–22 Gb</td>
<td>9–11 Gb</td>
<td>7 hours</td>
</tr>
<tr>
<td>2 × 50</td>
<td>50–60 Gb</td>
<td>25–30 Gb</td>
<td>16 hours</td>
</tr>
<tr>
<td>2 × 100</td>
<td>100–120 Gb</td>
<td>50–60 Gb</td>
<td>27 hours</td>
</tr>
<tr>
<td>2 × 150</td>
<td>150–180 Gb</td>
<td>75–90 Gb</td>
<td>40 hours</td>
</tr>
<tr>
<td>2 × 250*</td>
<td>250–300 Gb</td>
<td>125–150 Gb</td>
<td>60 hours</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reads Passing Filter*</th>
<th>Up to 600 million</th>
<th>Up to 300 million</th>
</tr>
</thead>
</table>

| Quality               | ≥ 85% of bases above Q30 at 2 × 50 bp | ≥ 80% of bases above Q30 at 2 × 100 bp | ≥ 75% of bases above Q30 at 2 × 150 bp |
# HiSeq 2500

<table>
<thead>
<tr>
<th>Application</th>
<th>Rapid-Run Mode</th>
<th>High-Output Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription Factor</td>
<td>40 Samples</td>
<td>260 Samples</td>
</tr>
<tr>
<td>ChIP-Seq</td>
<td>7 Hours</td>
<td>29 Hours</td>
</tr>
<tr>
<td>mRNA-Seq</td>
<td>12 Samples</td>
<td>80 Samples</td>
</tr>
<tr>
<td>50 M Reads</td>
<td>16 Hours</td>
<td>2.5 Days</td>
</tr>
<tr>
<td>Nextera Rapid Capture Exome</td>
<td>20 Samples</td>
<td>150 Samples</td>
</tr>
<tr>
<td>37 Mb Region</td>
<td>27 Hours</td>
<td>5 Days</td>
</tr>
<tr>
<td>Human Whole Genome</td>
<td>1 Sample</td>
<td>8 Samples</td>
</tr>
<tr>
<td>&gt;50x Coverage</td>
<td>27 Hours</td>
<td>6 Days</td>
</tr>
<tr>
<td>De Novo Sequencing</td>
<td>1 Sample</td>
<td>60 Hours</td>
</tr>
<tr>
<td>2.5 Gb Genome, 100x Coverage</td>
<td>2 x 250 bp</td>
<td></td>
</tr>
</tbody>
</table>
## NextSeq 500

<table>
<thead>
<tr>
<th>Application</th>
<th>High-Output Flow Cells</th>
<th>Mid-Output Flow Cells</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of Samples</td>
<td>Time</td>
</tr>
<tr>
<td>Gene Expression Profiling</td>
<td>40</td>
<td>11 hours</td>
</tr>
<tr>
<td>&gt; 10 M Reads</td>
<td>1 x 75 bp</td>
<td></td>
</tr>
<tr>
<td>mRNA-Seq</td>
<td>16</td>
<td>18 hours</td>
</tr>
<tr>
<td>&gt; 25 M Reads</td>
<td>2 x 75 bp</td>
<td></td>
</tr>
<tr>
<td>Enrichment Panel</td>
<td>36</td>
<td>29 hours</td>
</tr>
<tr>
<td>12 Mb Region</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; 20x coverage at &gt; 95% targets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole-Exome Sequencing</td>
<td>12</td>
<td>18 hours</td>
</tr>
<tr>
<td>50x mean coverage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small Whole-Genome Sequencing</td>
<td>30</td>
<td>29 hours</td>
</tr>
<tr>
<td>130 Mb Genome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; 30x coverage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 x 150 bp</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## MiSeq System Performance Parameters

### MiSeq Reagent Kit v2

<table>
<thead>
<tr>
<th>Read Length</th>
<th>Total Time*</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 x 36 bp</td>
<td>~4 hours</td>
<td>540-810 Mb</td>
</tr>
<tr>
<td>2 x 25 bp</td>
<td>~5.5 hours</td>
<td>750-850 Mb</td>
</tr>
<tr>
<td>2 x 150 bp</td>
<td>~24 hours</td>
<td>4.5-5.1 Gb</td>
</tr>
<tr>
<td>2 x 250 bp</td>
<td>~39 hours</td>
<td>7.5-8.5 Gb</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reads Passing Filter†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Reads</td>
</tr>
<tr>
<td>Paired-End Reads</td>
</tr>
</tbody>
</table>

**Quality Scores††**

- > 90% bases higher than Q30 at 1 x 36 bp
- > 90% bases higher than Q30 at 2 x 25 bp
- > 80% bases higher than Q30 at 2 x 150 bp
- > 75% bases higher than Q30 at 2 x 250 bp

### MiSeq Reagent Kit v3

<table>
<thead>
<tr>
<th>Read Length</th>
<th>Total Time*</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 x 75 bp</td>
<td>~21 hours</td>
<td>3.3-3.8 Gb</td>
</tr>
<tr>
<td>2 x 300 bp</td>
<td>~56 hours</td>
<td>13.2-15 Gb</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reads Passing Filter†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Reads</td>
</tr>
<tr>
<td>Paired-End Reads</td>
</tr>
</tbody>
</table>

**Quality Scores††**

- > 85% bases higher than Q30 at 2 x 75 bp
- > 70% bases higher than Q30 at 2 x 300 bp

---

*Total times include cluster generation, sequencing, and base calling on a MiSeq system enabled with dual surface scanning.
† Install specifications based on Illumina PhIX control library at supported cluster densities between 865–965 k/mm² clusters passing filter for v2 chemistry and 1200–1400 k/mm² clusters passing filter for v3 chemistry. Actual performance parameters can vary based on sample type, sample quality, and clusters passing filter.
†† The percentage of bases > Q30 is averaged across the entire run.
bp = base pairs, Mb = megabases, Gb = gigabases, M = millions
# Illumina sequencing

<table>
<thead>
<tr>
<th>Instrument</th>
<th>MiSeq</th>
<th>NextSeq 500</th>
<th>HiSeq 2500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run Modes</td>
<td>NA</td>
<td>Mid Output</td>
<td>High Output</td>
</tr>
<tr>
<td>Output (Gb)</td>
<td>0.3-15Gb</td>
<td>20-39Gb</td>
<td>30-120Gb</td>
</tr>
<tr>
<td>Lanes per flowcell</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Max reads per lane (single read)</td>
<td>25 mil (V3 kit)</td>
<td>130 mil</td>
<td>400 mil</td>
</tr>
<tr>
<td>Max read length</td>
<td>2 X 300bp</td>
<td>2 X 150bp</td>
<td>2 X 150bp</td>
</tr>
<tr>
<td>Run quality</td>
<td>&gt;70% Q30 at 2X300bp</td>
<td>&gt;80% Q30 at 2X150bp (V2)</td>
<td>&gt;80% Q30 at 2X150bp (V2)</td>
</tr>
<tr>
<td>Applications</td>
<td>Amplicons, 16S metagenomics, small RNA, small genomes</td>
<td>Small genomes, RNASeq, small RNA, ChIPSeq, exomes and many others</td>
<td>Genomes, RNASeq, small RNA, ChIPSeq, exomes and many others</td>
</tr>
</tbody>
</table>
HiSeq, MiSeq, NextSeq

1. Library Prep
2. Quality check BioAnalyzer
3. Sequencing
4. Demultiplexing
5. Fastq files
6. Analysis
Demultiplexing

- `bcl2fastq`
Demultiplexing

Multiplexing is using sample specific adaptors in library preparation step, then sequencing them all together in one lane.
SampleSheet for demultiplexing

[Header],
Sample Name,Holly Eckelhoefer,
Experiment Name,142S,
Date,11/15/2016,
Workflow,GenerateFASTQ,
Application,NextSeq FASTQ Only,
Assay,TruSeq LT,
Description,NEB Pool,
Chemistry,Default,

[Reads],
76,

[Settings],
Adapter,AGATCGGAAGAGCGACACGCTCTTCCGATCTCA,
AdapterRead2,AGATCGGAAGAGCGACACGCTCTTCCGATCTAGGAAAGTGT,

[Data],
Sample_ID,Sample_Name,Sample_Plate,Sample_Well,Index_ID,index,Sample_Project,Description
1A,,,Index_1,ATGCAG,
1B,,,Index_2,CATAGT,
2A,,,Index_3,TTAGGC,
2B,,,Index_4,TCAGCA,
3A,,,Index_5,GACAGT,
3B,,,Index_6,CCTACG,
4A,,,Index_7,CTAGTC,
4B,,,Index_8,ACTTGA,
5A,,,Index_9,GATCAG,
Demultiplexing

- bcl2fastq
a | Spike-in controls can be added to DNA or RNA samples for combined library preparation and sequencing

b | Well-characterized biological reference materials act as valuable process controls but cannot be directly added to samples.

c | Finally, in silico sequencing libraries can be used to rapidly evaluate key bioinformatics steps (indicated in blue).

Reference standards can be used to assess various biases and errors in the next-generation sequencing workflow (indicated in beige). rDNA, ribosomal DNA.

Reference standards for next-generation sequencing, Nature 2016 Simon A. Hardwick1,2, Ira W. Deveson1,3 and Tim R. Mercer
High throughput sequencing website

IIGB HT Sequencing

Main Page

HiSeq2500 HT Sequencing

The IIGB Genomics Core facility provides a next-generation sequencing service using Illumina's HiSeq2500 platform. Combined with the latest in cluster generation and sequencing reagents, the HiSeq provides the yield of the HiSeq2000 (high yield mode) plus the ability to deliver 300bp reads (150bp paired-end) in rapid mode. It achieves this due to improved flowcell and hardware design and the ability to run two flowcells independently and simultaneously. The HiSeq2500 can generate genome, small RNA and transcriptome level sequences with unprecedented speed for a cost per cycle. The most common research applications of this technology are:

- Whole genome sequencing/re-sequencing
- Genome-wide detection of SNPs and mutations
- DNA methylation profiling
- DNA-protein interactions
- RNA expression profiling
- Small RNA profiling and discovery
- ChIP sequencing
Projects and Flowcells

- View and Submit Projects
- Download Project Data
- View and Submit Flowcells
- View Flowcell QC Reports
- View All Samples
High throughput sequencing website – Click here to submit a new project
### Sample Information (Click + to enter data for each sample). Please note that all fields are mandatory.

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample 1</strong></td>
<td></td>
</tr>
<tr>
<td>Label</td>
<td></td>
</tr>
<tr>
<td>Project Description</td>
<td></td>
</tr>
<tr>
<td>Organism(s) (full scientific name)</td>
<td></td>
</tr>
<tr>
<td>Sequencer and read options</td>
<td></td>
</tr>
<tr>
<td>Sample Type (RNASeq, DNASeq, small RNA, etc)</td>
<td></td>
</tr>
<tr>
<td>DNA Concentration (ng/µl)</td>
<td>ng/µl</td>
</tr>
<tr>
<td>Number of SMRT cells (for PacBio only)</td>
<td></td>
</tr>
<tr>
<td>Movie_length (in mins) (for PacBio only)</td>
<td></td>
</tr>
<tr>
<td>Concentration measured by</td>
<td></td>
</tr>
<tr>
<td>DNA Concentration (nM)</td>
<td>nM</td>
</tr>
<tr>
<td>Average length or size range of library, including adaptors (bp)</td>
<td>bp</td>
</tr>
<tr>
<td>Sample Volume submitted (µl)</td>
<td>µl</td>
</tr>
<tr>
<td>Kit or custom library-making protocol used</td>
<td></td>
</tr>
</tbody>
</table>

*Please note this field is mandatory and will assist with flowcell loading estimates.*
File formats

- FASTA file
- FASTQ file
- SAM format
- BAM format
- VCF
- BCF
FASTA format

- FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences.
- A sequence in FASTA format begins with a single-line description, followed by lines of sequence data.
- The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. It is recommended that all lines of text be shorter than 80 characters in length.

>gi|186681228|ref|YP_001864424.1| phycoerythrobilin:ferredoxin oxidoreductase
MNSERSDVTLYQPFLDYAIAYMRSRLDLEPYPIPTGFESNSAVVGKKNQEEVVTTSYAFQTAKLRQIRA
AHVQGGNSLQQLNVIFPHLYNDLPFFGADLVTPGGLIAALDMQPLFRDSAYQAAYTEPILPFIHAHQ
QHLSWGGDFPEEAQPFFEAPFLWTRPQETAVETQVFAAFKDYLKAYLDFVEQEAEVTDSONLVAIKQAQ
LRYLRYRAEKPARGMFKRFYGAETEYIHGFIIFDLKERKLTVVK
FASTQ format

A FASTQ file normally uses four lines per sequence.

- Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description (like a FASTA title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

```
@NB501124:103:HLK7WBGXY:1:11101:7384:1061 1:N:0:AGTCAA
CCCATNGATTGCACGCAGTCTCGAAAGGTGACTTATCTCGGGAATCTCCGGGATCTACGTATTTTCAACTCCCG
+
AAAAA#EEEEEEEEEEEEEX/EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB501124:103:HLK7WBGXY:1:11101:14405:1063 1:N:0:AGTCAA
GCAGCAATGATGCGATAATAAAAAACAGAAGTTGCGTGCAATAAGGTTAGGGATCATCAAAAAACACCAACACCATCCA
+
AAAAAEEEEEEEEEEEEEE/EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE/EEEEEE666666EEAE
```
FASTQ quality

<table>
<thead>
<tr>
<th>Quality Score</th>
<th>Phred+33</th>
<th>Phred+64</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 40</td>
<td>Sanger</td>
<td>Solexa</td>
</tr>
<tr>
<td>(-5, 40)</td>
<td>Illumina 1.3+</td>
<td>Illumina 1.5+</td>
</tr>
<tr>
<td>(3, 40)</td>
<td>Illumina 1.8+</td>
<td></td>
</tr>
</tbody>
</table>

Additional notes:
- With 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
- (Note: See discussion above)
Phred Quality Score

\[ Q = -10 \log_{10} P \]

or

\[ P = 10^{\frac{-Q}{10}} \]

Phred quality scores are logarithmically linked to error probabilities

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</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 1000</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>99.999%</td>
</tr>
<tr>
<td>60</td>
<td>1 in 1,000,000</td>
<td>99.9999%</td>
</tr>
</tbody>
</table>
Basic data processing

- Quality check
  - FASTQC

- Adapter removal
  - Trim-galore – cutadapt + FASTQC
FASTQC continued…

Per sequence quality scores

Per sequence GC content
Check for contamination in the data

- NCBI (National Center for Biotechnology Information)
- BLAST (Basic Local Alignment Search Tool)
- NR (non-redundant) database
PacBio RSII sequencer
PacBio – SMRT Analysis Server

Single molecule resolution in real time

• Short waiting time for result and simple workflow
  – Generate basecalls in <1 day
  – Polymerase speed ≥1 base per second

• No amplification required
  – Bias not introduced
  – More uniform coverage

• Long reads
  – Identify repeats and structural variants
  – Less coverage required

• Information content
  – One assay, multiple applications
    • Genetic variation (SVs to SNPs)
    • Methylation
    • Enzymology

HOW IT WORKS

DNA is copied by an enzyme in PacBio’s machine

The DNA letters used to make the copy have been tagged to emit tiny flashes of colored light.

A camera can catch these tiny flashes thanks to a 56-nanometer hole that screens out other light.

Pacbio: reads

DNA fragment with hairpins

Polymerase Read

Subread

Read of Insert or CCS
Example Data: 1 SMRT cell
FASTQC on PacBio data
Genome Assembly using NGS

• Short-read *de novo* assembly by NGS
  – Requires mate-pair sequences
    • Ideally with different insert sizes
  – Complicated analysis
    • Assembly, scaffolding, finishing
    • Maybe even some manual steps
  => Rather expensive and time consuming

• Long reads really makes a difference!!
  – We can assemble genomes using PacBio data only!
SMRT analysis portal

RECENT JOBS

<table>
<thead>
<tr>
<th>Job Name</th>
<th>Protocol</th>
<th>Reference Sequence</th>
<th>Started</th>
<th>Status</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>mapping_sample_520</td>
<td>RS_Resequencing.1</td>
<td>nucmer</td>
<td>2013-09-13T14:21:21</td>
<td>Completed</td>
<td>ugc_admin</td>
</tr>
<tr>
<td>test11a</td>
<td>RS_HuAP_Aseembly.1</td>
<td>nucmer</td>
<td>2013-09-13T13:25</td>
<td>Completed</td>
<td>ugc_admin</td>
</tr>
<tr>
<td>mapping_sample_519</td>
<td>RS_Resequencing.1</td>
<td>nucmer</td>
<td>2013-09-13T12:55</td>
<td>Completed</td>
<td>ugc_admin</td>
</tr>
<tr>
<td>mapping_sample_518</td>
<td>RS_Resequencing.1</td>
<td>nucmer</td>
<td>2013-09-13T11:59</td>
<td>Completed</td>
<td>ugc_admin</td>
</tr>
<tr>
<td>sample_50_mapping</td>
<td>RS_Resequencing.1</td>
<td>nucmer</td>
<td>2013-09-13T11:35</td>
<td>Completed</td>
<td>ugc_admin</td>
</tr>
</tbody>
</table>
SMRT analysis pipelines

- Mapping
- Variant calling
- Assembly
- Scaffolding
- Base modifications
NGS applications

- RNA-Seq analysis
- Var-Seq analysis
- ChIP-Seq analysis
- Methyl-Seq analysis
- Microbial analysis
- Genome Assembly
Alignment Tools

- BWA
- Bowtie
- Tophat
- STAR
- gSNAP
RNA-Seq analysis

1. mRNA Isolation

2. Illumina Sequencing

3. Align Sequences against Genome

4. Generate Sequence Counts for all Genes in Genome

Gene A: \( \frac{30}{10} = 3 \text{ fold change} \)

Gene B: \( \frac{10}{5} = 2 \text{ fold change} \)
Var-Seq analysis

Variant calling (SNPs and short Indels)

- Genomic reads
- Aligned reads
- Reference genome
- Conditional variant filtration
- Variant files
- Annotated variants
- Variant annotation database
ChIP-Seq analysis

1: Data processing and quality control
- Raw data
- Quality control
- Genome alignment
- Sequence visualization
- Quality control
- Peak detection

2: Statistical analysis
- Nucleosome positioning
- Differential peak calling
- Motif discovery
- Interpretation
- Gene target
- Data visualization

3: Functional analysis
- GSA
- PPIs
- Gene ontologies
- Signaling pathways
- Metabolic pathways
- Other databases:
  - Epigenomes
  - GRNs
  - Transcription factors

Legend:
- □ ChIP-seq tools
- □ Common tools
- □ Functional analysis
DNA methylation

- Addition of a methyl group to DNA strand itself, often to the 5 carbon position of a cytosine ring
- Typically occurs at CpG dinucleotides in vertebrates
- Act as epigenetic control mechanism for gene regulation
- Hinder binding of transcription factors
- Associated with many human malignancies
- Essential to normal development, X-chromosome activation, genetic imprinting, gene suppression, and carcinogenesis
- Hypo-methylation is associated with gene transcription
- Hyper methylation is associated with gene expression

http://helicase.pbworks.com/w/page/17605615/DNA%20Methylation
Bisulphite sequencing (Methyl-Seq)

- Technique to determine DNA methylation patterns
- Treated with bisulphite which converts cytosine residues to uracil, but leaves 5-methylcytosine residues unaffected.
- Call methylation status of a base – percentage
- Determine how many of the bases that are aligning to a given cytosine location in the genome have actual C bases in the reads
- Percentage methylation score on that base
Thank you!