RNA-seq Approach to Study Gene Expression Profiles in Non-Model Organisms

Asela Wijeratne (wijeratne.1@osu.edu)
Saranga Wijeratne (wijeratne.3@osu.edu)
Tea Meulia (Meulia.1@osu.edu)
Why we sequence transcriptomes?

- Show repertoire of expressed sequences, including rare transcripts
- Gene expression
- SNPs
- Alternative splicing
- Structural variation
- Practical alternative to genome sequencing for non-model organisms
Genomic model vs. non-model organisms

- **Model organism** is a non-human species that is extensively studied to understand particular biological phenomena.

- Genomic model organisms:
  - Occupy a pivotal position in the evolutionary tree
  - Some quality of their genome makes them ideal to study

- Non-Genomic model organisms
  - Rest of them
  - Important for many reasons:
    - Human pathogens
    - Agricultural pathogens and pests
RNA-seq for non-model organisms

- No sequenced genomes most of the time
- Most analytical tools are designed for model organisms
- Present unique challenges for quality control for data analyses
What are we going to learn

- Basics of initial designing the experiment
- Analyses pipeline
  - Primary analyses
    - Read preprocessing
    - Transcript assembly
    - Assessing quality of assembly
    - Mapping reads back assembled transcripts
  - Secondary analyses
    - Transcript characterization and annotation
    - Comparative gene expression
Sequencing
Platform of choice

<table>
<thead>
<tr>
<th>Platform</th>
<th>Library/template preparation</th>
<th>NGS chemistry</th>
<th>Read length (bases)</th>
<th>Run time (days)</th>
<th>Gb per run</th>
<th>Machine cost (US$)</th>
<th>Pros</th>
<th>Cons</th>
<th>Biological applications</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche/454's GS FLX Titanium</td>
<td>Frag, MP/emPCR</td>
<td>PS</td>
<td>330*</td>
<td>0.35</td>
<td>0.45</td>
<td>500,000</td>
<td>Longer reads improve mapping in repetitive regions; fast run times</td>
<td>High reagent cost; high error rates in homopolymer repeats</td>
<td>Bacterial and insect genome de novo assemblies; medium scale (&lt;3 Mb) exome capture; 16S in metagenomics</td>
<td>D. Muzny, pers. comm.</td>
</tr>
<tr>
<td>Illumina/Solexa’s GA II</td>
<td>Frag, MP/solid-phase</td>
<td>RTs</td>
<td>75 or 100</td>
<td>4*, 9*</td>
<td>18*, 35*</td>
<td>540,000</td>
<td>Currently the most widely used platform in the field</td>
<td>Low multiplexing capability of samples</td>
<td>Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics</td>
<td>D. Muzny, pers. comm.</td>
</tr>
<tr>
<td>Life/APG's SOLiD 3</td>
<td>Frag, MP/emPCR</td>
<td>Cleavable probe SBL</td>
<td>50</td>
<td>7*, 14*</td>
<td>30*, 50*</td>
<td>595,000</td>
<td>Two-base encoding provides inherent error correction</td>
<td>Long run times</td>
<td>Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics</td>
<td>D. Muzny, pers. comm.</td>
</tr>
</tbody>
</table>
Different read types for Illumina sequencing

- Paired end reads
- Single end reads

Final PCR enrichment

Size selection 200 - 600bp
Replication and sequencing depth

- **Replication:**
  - At least two biological replicates for expression analyses

- **Sequencing depth:**
  - Depends:
    - Goals of the experiments
    - Samples and conditions; sample preparation
    - Sequencing type
    - Number and length of the genes
No of transcripts assembled increases as the no of reads increase

| Combination | Reads (millions) | Expressed genes $|$ | Number of reference cDNA with matches to assembled sequences* |
|-------------|-----------------|---------------------|---------------------------------------------------------------|
| Sample1     | 40              | 13416               | 10500                                                         |
| Sample 1+2  | 72              | 15646               | 14250                                                         |
| Sample 1+2+3| 112             | 16916               | 19223                                                         |

$|These is based on the alignments to reference cDNA; Number of sequences with more than 200 Illumina reads aligned

*number of reference cDNA that showed sequence similarity to assembled transcripts that showed at least 80% coverage and had at least 95% sequence identity
This shows number of reference cDNA that showed sequence similarity to assembled transcripts that showed at least 80% coverage and had at least 95% sequence identity. X axis is percentile ranking of the expression and Y axis is the number of sequences.
Analyzing the data

- Manually using UNIX terminal prompt
- Automated using PERL or Python scripts
- Using Makefile
- Reproducibly Workflow environment
  - Galaxy
  - Taverna
Steps of data analyses

Raw data, Images, signals → Basecalling → Short reads, Quality

Short reads, Quality → Assembly → Alignment, Assembled reads

Alignment, Assembly → Comparison → SNP detection

Database searches → Statistical analysis → Annotations

Core lab

Secondary

Tertiary
Base calling from raw data

The identity of each base of a cluster is read off from sequential images.

From Debbie Nickerson, Department of Genome Sciences, University of Washington, http://tinyurl.com/6zbzh4
FASTQ

@BILLIEHOLIDAY:1:1:6:768#0/1
CATGATGCCAGAGGCAGAGGACAGGTTGCCAAAGCTCTCGCTTCTGGAACGTCTGAGGTAT
CAATAAGCTC
+BILLIEHOLIDAY:1:1:6:768#0/1
abbbababbbbaa_`a`_aa``^_aaa_][_aa``^`]`a`_abbabbbbbbabbabbbbabbba`abbb
bbbbbbbbbbb

<table>
<thead>
<tr>
<th>Whatever_name</th>
<th>the unique instrument name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>flowcell lane</td>
</tr>
<tr>
<td>1</td>
<td>tile number within the flowcell lane</td>
</tr>
<tr>
<td>6</td>
<td>'x'-coordinate of the cluster within the tile</td>
</tr>
<tr>
<td>768</td>
<td>'y'-coordinate of the cluster within the tile</td>
</tr>
<tr>
<td>#0</td>
<td>index number for a multiplexed sample (0 for no indexing)</td>
</tr>
<tr>
<td>1</td>
<td>the member of a pair, /1 or /2 (paired-end or mate-pair reads only)</td>
</tr>
</tbody>
</table>
Sources of error

- De-phasing
  - Lagging strand de-phasing from incomplete extension
  - Leading strand de-phasing from over-extension
- Polymerase errors ($10^{-5}$ to $10^{-7}$)
- More likely to have an error after G
- PCR induced errors (AT or GC rich regions)
- Cross-cluster bridge formation
BER=base error rate

- BER: Estimated probability of a base being wrong
- Phred quality score (as of Illumina pipeline 1.3):

\[ Q = -10 \cdot \log_{10}(BER) \]

- In a FASTQ file they are encoded as ASCII: Q+64
- May be used to filter out poor quality reads, and to improve alignments

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of Incorrect Based 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 10000</td>
<td>99.99%</td>
</tr>
</tbody>
</table>
Quality assessment - FASTQC
Preprocessing

- Why
  - Get rid of artifacts from library preparations (PCR) and sequencing steps (errors/adapters)
    - Resulted in fragmented assemblies or mis-assemblies
    - Inflate dataset

- Remove
  - Adapter sequences
  - Low quality bases
    - FASTX-toolkit
    - Cutadapt
    - Custom perl script from UC Davis for paired-end reads

- Remove contaminates
  - DeconSeq
    - Web based
    - Standalone
Assembling sequence reads into transcripts
Transcript Assembly of Illumina reads

- **Challenges**
  - Short reads (76-150bp)
  - Large datasets
  - Needs a large number of reads
  - Transcript expression is not uniform

- **De novo assembly**
  - No reference
    - Rnnotator, Trinity
  - Highbred assembly
    - Uses Longer reads (previous cDNA or 454 data) with Illumina short reads (Best approach)

- **Reference based assembly**
  - TopHat/Cufflinks, ERANGE, and Scripture
Excluded from further analysis

Raw Illumina Paired end reads

Quality filter

Low Quality Reads

High Quality Reads

BWA

Alignments

Samtools

Count reads

Previously assembled contigs

Rnnotator

Assembled Contigs

Minimus2

Reference

Novel gene discoveries and annotations

Differential gene expression Analysis (DESeq)
The Rnnotator assembly pipeline

- Input preprocessed data
- Strand information
- Remove duplicates
- K-mer based filtering

Minumus2 in Amos package
Assessing the quality of the assembly
A comparison of the performance between the Rnnotator assembly and a single Velvet assembly.

<table>
<thead>
<tr>
<th></th>
<th>Rnnotator (non-stranded)</th>
<th>Rnnotator</th>
<th>Velvet</th>
<th>Oases</th>
<th>Multiple-k</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. albicans SC531</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>94.0</td>
<td>95.0</td>
<td>97.4</td>
<td>92.3</td>
<td>96.6</td>
</tr>
<tr>
<td>Completeness</td>
<td>81.9</td>
<td>80.4</td>
<td>66.7</td>
<td>79.9</td>
<td>85.9</td>
</tr>
<tr>
<td>Contiguity</td>
<td>58.4</td>
<td>58.0</td>
<td>46.6</td>
<td>47.9</td>
<td>37.3</td>
</tr>
<tr>
<td>Gene fusions</td>
<td>1.73</td>
<td>0.26</td>
<td>1.18</td>
<td>1.31</td>
<td>0.20</td>
</tr>
</tbody>
</table>

1 Accuracy: the percentage of contigs that share at least 95% identity with the reference genome;  
2 Completeness: percentage of known genes covered by the contigs to at least 80% of the gene length;  
3 Contiguity: percentage of complete genes covered by a single contig over at least 80% of the gene length.  
4 Gene fusions: the percentage of contigs that contain more than 50% of two or more annotated genes.  

Length distribution

No of sequences

Frequency

Reference
Assembled
**Ortholog Hit Ratio indicates the completeness of assembly**

Table:

<table>
<thead>
<tr>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total expressed</td>
<td>20721</td>
</tr>
<tr>
<td>More than 200 reads</td>
<td>17068</td>
</tr>
</tbody>
</table>

"Ortholog Hit Ratio" = (#Bases in $\alpha$)/(#Bases in $\beta$)

About 20,000 assembled ESTs had more than 0.8 OHR

O’Neil, 2010; Ewen-Campen, 2011 Python script
Depth of sequencing and assembly can be assessed using Ultra Conserved Othologs (UCO)

- Compare with eukaryotic ultraconserved genes

Annotation of assembled transcripts
Gene Ontology Consortium

Provide a controlled vocabulary to describe gene and gene product attributes in any organism

Includes both the development of the Ontology and the maintenance of a Database of annotations

Adapted from Blast2Go presentation
THE ONTOLOGY

✓ Annotations are given to the most specific (low) level.

✓ True path rule: annotation at a term implies annotation to all its parent terms

✓ Annotation is given with an Evidence Code:
  - IDA: inferred by direct assay
  - TAS: traceable author statement
  - ISS: inferred by sequence similarity
  - IEA: electronic annotation
  - ....

Adapted from Blast2Go presentation
Blast2GO

- Suite for functional annotation and data mining on functional data
  - Considerations for annotation
    - Similarity
    - Length of the overlap
    - Percentage of hit sequence spanned by the overlap
    - Evidence original annotation
    - Blast hits and motif hits
    - Refinement by additional methods
  - Visualization:
    - Annotation charts
    - Knowledge discovery on the DAG
- Desktop Java application
- Web interface @ Babelomics: Babelomics for non-model

Adapted from Blast2Go presentation
Blast2GO Annotation Strategy

More information:
http://www.blast2go.com/data/blast2go/b2g_tutorial_23062009.pdf
# Blast2go output

<table>
<thead>
<tr>
<th>Seq. Name</th>
<th>Seq. Description</th>
<th>Seq. Length</th>
<th>#Hits</th>
<th>min. eValue</th>
<th>mean Similarity</th>
<th>#GOs</th>
<th>GOs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>---NA---</td>
<td>165</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Gene 2</td>
<td>pathogenesis-related protein 1</td>
<td>480</td>
<td>20</td>
<td>9.57E-94</td>
<td>89.95%</td>
<td>3</td>
<td>C:extracellular region; P:defense response to fungus; P:killing of cells of another organism</td>
</tr>
</tbody>
</table>
Molecular function

- Catalytic activity (4)
- Binding (4)
- Enzyme regulator activity (2)
- Transporter activity (1)
Mapping
Mapping reads to assembled transcripts

- Aligners
  - Burrow Wheeler Transform (BWT)
    - Fast
    - Need good quality data
    - Bowtie, BWA, SOAP2
  - Hash tables
    - Slower
    - More sensitive – better for SNP finding
    - PerM, SHRIMP, BFAST, ELAND

- Mapping back to a reference genome
  - Splice aware aligners:
    - TOPHAT, MapSplice

- Mapping back to assembled transcripts
  - No splice junctions
mRNA seq reads mapped to cDNA using BWA

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total Reads</th>
<th>Reads Mapped</th>
<th>% Reads Mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>36,856,702</td>
<td>29,479,287</td>
<td>80%</td>
</tr>
<tr>
<td>Sample 2</td>
<td>38,967,190</td>
<td>34,688,060</td>
<td>89%</td>
</tr>
<tr>
<td>Sample 3</td>
<td>50,018,335</td>
<td>37,275,875</td>
<td>75%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total Reads</th>
<th>Reads Mapped</th>
<th>% Reads Mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>39,584,939</td>
<td>30,702,758</td>
<td>78%</td>
</tr>
<tr>
<td>Sample 2</td>
<td>32,622,057</td>
<td>27,450,258</td>
<td>84%</td>
</tr>
<tr>
<td>Sample 3</td>
<td>39,060,469</td>
<td>31,627,675</td>
<td>81%</td>
</tr>
</tbody>
</table>

- Treatment 1
- Treatment 2
Differential gene expression analyses
Estimate the number of reads in each transcript

- Short reads aligned to a transcript are counted

- Count each read at most once

- Reads are discarded
  - Not uniquely mapped
  - Aligned to several genes
  - Poor alignment quality score
  - (for paired-end reads) the mates matches to different genes

Simon Anders
Normalization

- Number of reads (coverage) vary between samples (Sequencing depth)
- Other technical effects
- **RPKM** (Reads Per KB per Million mapped reads)
  - divide counts per million reads and by gene length
- RPKM assumes:
  - Total amount of RNA per cell is constant
  - Most genes do not change expression
- RPKM is invalid if there are a few very highly expressed genes that have dramatic change in expression (dominate the pool of reads)
- Quantile normalization (Bullard, 2010)
- Produces non-integer counts, not good for Poisson or Negative Binomial model based methods
Scaling factor normalization as in DEseq

- Reference sample
  - The geometric mean of the counts in all samples for each gene
- Get the sequencing depth of a sample relative to the reference
  - Calculate for each gene the quotient of the counts in the test sample divided by the counts of the reference sample
- Median of all the quotients is the depth of the library
Noise

- Shot noise
  - The variance in counts that persists even if everything is exactly equal
    - unavoidable, appears even with perfect replication
    - dominant noise for weakly expressed genes

- Technical noise
  - from sample preparation and sequencing

- Biological noise
  - Dominant noise for strongly expressed genes
Statistical methods for DEG analyses

- Mathematically shown:
  - If
    - number of reads is large
    - Probability of a read mapped to a gene is small
  - Binomial distribution is well approximated by Poisson distribution
- Poisson distribution: mean = variance
- counts for the same gene from different technical replicates have a variance equal to the mean (Poisson)
- counts for the same gene from different biological replicates have a variance exceeding the mean (overdispersion)
- The negative-binomial distribution
  - A commonly used generalization of the Poisson distribution with two parameters
- Estimate a scaling factor to use with the statistical model
Biological vs technical replicates

RNA-Seq of yeast [Nagalakshmi et al, 2008]
Packages for testing for differential

- **Parametric**
  - Based on negative-binomial distribution:
    - edgeR (Robinson, McCarthy, Smyth)
    - DESeq (Anders, Huber)
    - BaySeq (Hardcastle, Kelly)
  - Based on Binomial distribution:
    - DEGSeq (Wang et al.)

- **Non parametric**
  - Cuffdiff
  - NOIseq

Performance was compared by Kvam, 2011
BaySeq was slightly better
Correlation between replicates
### DEseq output

<table>
<thead>
<tr>
<th>id</th>
<th>Base Mean</th>
<th>Base MeanA</th>
<th>Base MeanB</th>
<th>Fold Change</th>
<th>log2FoldChange</th>
<th>p val</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene 1</td>
<td>110</td>
<td>22</td>
<td>198</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gene 2</td>
<td>544</td>
<td>860</td>
<td>227</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Base Mean**: Mean of Base MeanA and B
- **Base MeanA**: Reads from sample A divided by the size factor
- **Base MeanB**: Reads from sample A divided by the size factor
- **padj**: adjusted P value for multiple testing with the Benjamini-Hochberg procedure
Tutorial

- [http://galaxy.oardc.ohio-state.edu](http://galaxy.oardc.ohio-state.edu)
- email: mcic@gmail.com
- pwd: glbiouse

Shared data:

- **Published Pages:** [GLBIO RNA-seq Analysis Exercise](http://galaxy.oardc.ohio-state.edu)

Data Libraries

- [DESeq Sample Data](http://galaxy.oardc.ohio-state.edu): to use with DEseq
- [Rnnotator - final_contigs.fa](http://galaxy.oardc.ohio-state.edu): For mapping
- [Rnnotator_Contigs_blastx_with_nr](http://galaxy.oardc.ohio-state.edu): For annotation
<table>
<thead>
<tr>
<th>Step</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1: Input dataset</td>
<td>sample_2_1.fastq</td>
</tr>
<tr>
<td>sample_2_1.fastq</td>
<td>select at runtime</td>
</tr>
<tr>
<td>Step 2: Input dataset</td>
<td>sample_2_2.fastq</td>
</tr>
<tr>
<td>sample_2_2.fastq</td>
<td>select at runtime</td>
</tr>
<tr>
<td>Step 3: Input dataset</td>
<td>sample_1_1.fastq</td>
</tr>
<tr>
<td>sample_1_1.fastq</td>
<td>select at runtime</td>
</tr>
<tr>
<td>Step 4: Input dataset</td>
<td>sample_1_2.fastq</td>
</tr>
<tr>
<td>sample_1_2.fastq</td>
<td>select at runtime</td>
</tr>
<tr>
<td>Step 5: Input dataset</td>
<td>blastx_input</td>
</tr>
<tr>
<td>blastx_input</td>
<td>select at runtime</td>
</tr>
<tr>
<td>Rnorthor Contigs blast against nt</td>
<td></td>
</tr>
</tbody>
</table>
Step 6: Rannotator

Library

Libraries 1
Library
non strand-specific paired-end library
Insert Length
300
Filename
select at runtime

Libraries 2
Library
non strand-specific paired-end library
Insert Length
300
Filename
select at runtime

Use Default General Options
Yes
Use Default Read Pre-processing Options
Yes
Use Default Assembly Options
No

[-a assembler] Assembler to use (velvet, oases) (default: velvet)
Thank you !!!!
RNA families

Coding
  - PolyA mRNA
  - Non-PolyA mRNA

Non-coding
  - Structural
    - DNA associated
    - RNA associated
  - Regulatory
    - Ribosome associated
    - Micro RNA
    - TSS associated
    - Anti-sense
    - Enhancer RNA
  - Replisome
  - DNA Repair
  - Telomeric
  - DNA methylation (piRNA)
  - rRNA
Second-gen sequencers

- **Illumina/GAII short-read sequencers**
  - (10+Gb in 50-100 bp reads, >100M reads, 4-8 days)

- **454 GS FLX pyrosequencer**
  - (100-500 Mb in 100-400 bp reads, 0.5-1M reads, 5-10 hours)

- **ABI capillary sequencer**
  - (0.04-0.08 Mb in 450-800 bp reads, 96 reads, 1-3 hours)

From John McPherson, OICR
Developments

Automated Library Construction  In response to the increasing demand for constructing Illumina libraries, a semi-automated process which enables us to construct 96 Illumina libraries in approximately 6-8 hours has been developed. With a few simple modifications, the library production efficiency has doubled. The modifications include the shearing of DNA with a Covaris E210, and the cleaning of enzymatic reactions and fragment size selection with SPRI beads and a magnetic plate holder. Recently, a BioMek FX robot has been programmed to carry out the library construction process.

BioMek FX Robot  The Beckman-Coulter Biomek FX robot is used to construct 96 Illumina libraries in parallel. This process automates the repetitive pipetting involved in library construction process and enables a single operator to construct 96 Illumina libraries in 3 days with minimal ergonomic risk.

Illumina Library Quality  The efficiency, quality and reproducibility of libraries created on BioMek are currently being optimized.

Library Quantification

Using the Agilent Bioanalyzer High Sensitivity DNA chip, the quantity of libraries was assessed. A sample is required to have a concentration of at least 10 nM. The graph below shows the Bioanalyzer traces of 13 libraries constructed by the BioMek FX robot.
Different types of platforms

- ABI capillary sequencer (First generation)
- Current generation (Second generation)
  - Illumina
  - 454
  - AB/SOLiDv3
  - Ion torrent
- Next generation (Third generation)
  - PacBio
## BWT aligners

<table>
<thead>
<tr>
<th>Input</th>
<th>All Rotations</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>^BANANA</td>
<td>^BANANA</td>
<td>ANANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BANANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BANANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>^BANANA</td>
<td></td>
<td>ANANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BANANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BANANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BANANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NANA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A</td>
</tr>
</tbody>
</table>

http://en.wikipedia.org/wiki/Burrows%E2%80%93Wheeler_transform
Sequence reads

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>ACGTGTTatc</td>
</tr>
<tr>
<td>02</td>
<td>CTACGTgtc</td>
</tr>
<tr>
<td>03</td>
<td>ACGTGTCcg</td>
</tr>
<tr>
<td>04</td>
<td>AGGCTAaat</td>
</tr>
<tr>
<td>05</td>
<td>GGTCAAAggc</td>
</tr>
<tr>
<td>06</td>
<td>GTCACCtgc</td>
</tr>
<tr>
<td>07</td>
<td>AAGTCGgag</td>
</tr>
<tr>
<td>08</td>
<td>TCGGCAact</td>
</tr>
</tbody>
</table>

Hash index

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>00AA</td>
<td></td>
</tr>
<tr>
<td>00AC</td>
<td></td>
</tr>
<tr>
<td>00AG</td>
<td></td>
</tr>
<tr>
<td>00AT</td>
<td></td>
</tr>
</tbody>
</table>

Read identifiers associated with each hash index

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>05</td>
<td></td>
</tr>
<tr>
<td>02 06</td>
<td></td>
</tr>
<tr>
<td>02</td>
<td></td>
</tr>
<tr>
<td>05 06 08</td>
<td></td>
</tr>
</tbody>
</table>
CONCEPTS OF FUNCTIONAL ANNOTATION

- **Gene/Protein function**
  - Refers to the molecular function of a gene or a protein:
    - Tyrosine kinase

- **Functional annotation**
  - More general, refers to the characterization of functional aspect of the protein:
    - Stress-related, cytoplasm, ABC transporter
  - Also refers to the process of assignment of a function label
  - Habitually, standard vocabularies are used to assign function
v of count values is modelled as $v = s\mu + \alpha s^2\mu^2$,

where $\mu$ is the expected normalized count value (estimated by the average normalized count value), $s$ is the size factor for the sample under consideration, and $\alpha$ is the dispersion value for the gene under consideration.