Introduction of RNA-Seq Analysis

Jiang Li, MS
Bioinformatics System Engineer I
Center for Quantitative Sciences (CQS)
Vanderbilt University

September 21, 2012
Goal of this talk

1. Act as a practical resource for those new to the topic of RNA-Seq analysis

2. Introduce our pipelines for the analysis of RNA-Seq in different aspects
Outline

1. Introduction to RNA-Seq (advantages and challenges)
2. Raw data (format, QC and manipulating)
3. Mapping (SAM, evaluation)
4. Differential Gene expression
   RNASeqDEPipeline by CQS
5. Fusion gene detection
   viewFusion by CQS
6. Resources
Introduction to RNA-Seq

RNA-seq, also called "Whole Transcriptome Shotgun Sequencing" ("WTSS"), refers to the use of high-throughput sequencing technologies to sequence cDNA in order to get information about a sample's RNA content.

We begin RNA-Seq analysis here
Introduction to RNA-Seq

RNA-Seq VS microarray (advantages)

Zhong Wang, Mark Gerstein and Michael Snyder 2009, Nature Genetics

<table>
<thead>
<tr>
<th>Technology specifications</th>
<th>Tiling microarray</th>
<th>RNA-Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principle</td>
<td>Hybridization</td>
<td>High-throughput sequencing</td>
</tr>
<tr>
<td>Resolution</td>
<td>From several to 100 bp</td>
<td>Single base</td>
</tr>
<tr>
<td>Throughput</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Reliance on genomic sequence</td>
<td>Yes</td>
<td>In some cases</td>
</tr>
<tr>
<td>Background noise</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>

Application

| Simultaneously map transcribed regions and gene expression | Yes | Yes |
| Dynamic range to quantify gene expression level | Up to a few-hundredfold | >8,000-fold |
| Ability to distinguish different isoforms | Limited | Yes |
| Ability to distinguish allelic expression | Limited | Yes |

Practical issues

| Required amount of RNA | High | Low |
| Cost for mapping transcriptomes of large genomes | High | Relatively low |
Introduction to RNA-Seq

Challenges

1. Bioinformatically
   - Extensive alternative splicing makes mapping reads to genome challenge
   - Storage and performance

2. Library construction
   - RNA is fragile compared to DNA
   - FFPE RNAseq is possible but not ideal
   - Ribosomal and mitochondrial RNAs contamination

3. Statistically more complicated than microarray
   - Large dynamic range (0-50000 comparing to 2-15 in microarray)
   - Large amount of zeros, conventional statistical methods don't apply
Outline

1. Introduction to RNA-Seq (advantages and challenges)
2. Raw data (format, QC and manipulating)
3. Mapping (SAM, evaluation)
4. Differential Gene expression
   RNASEqDEPPipeline by CQS
5. Fusion gene detection
   viewFusion by CQS
6. Resources
Raw data -- format (fastq)

**Example Read**
1 @HWI-ST508:152:D06G9ACXX:2:1101:1160:2042 1:Y:0:ATCACG
2 NAAGACCGAATTCTCAAGCTATGGTAAACATTGCACTGGCCTTTCATCTG
3 +
4 #11??+2<<CCB4AC?32@+1@AB1**1?AB<4=4>=BB<9=>?######

**Description**

**Line1:** Title and optional description

**Line2:** Sequence

**Line3:** begin with a “+” and is optionally followed by the same sequence identifier

**Line4:** encodes the quality values for the sequences in Line2

<table>
<thead>
<tr>
<th>Description, OBF name</th>
<th>ASCII characters</th>
<th>Quality score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Range</strong></td>
<td><strong>Offset</strong></td>
<td><strong>Type</strong></td>
</tr>
<tr>
<td>Sanger standard</td>
<td>fastq-sanger</td>
<td>33–126</td>
</tr>
<tr>
<td>Solexa/early Illumina</td>
<td>fastq-solexa</td>
<td>59–126</td>
</tr>
<tr>
<td>Illumina 1.3+</td>
<td>fastq-illumina</td>
<td>64–126</td>
</tr>
</tbody>
</table>

\[
Q_{PHRED} = -10 \times \log_{10}(P_e) \\
Q_{Solexa} = -10 \times \log_{10}\left(\frac{P_e}{1 - P_e}\right)
\]
Raw data -- QC

Information we check on raw data

a. The number of total reads
b. Per base sequence quality
c. Per sequence quality scores
d. More (GC%, duplication level, etc.)
Raw data -- QC

**FastQC**: A quality control for high throughput sequence data.
library(ShortRead)
fq<-readFastq("input.fastq")
phred.score<-as(quality(fq),"matrix")
boxplot(phred.score,col=rainbow(ncol(phred.score)),xlab="Position in read(bp)",ylab="Phred Scores",outline=FALSE,main="Quality scores across all bases")

per.sequece <- density(rowMeans(phred.score))
plot(per.sequece$x, per.sequece$y * per.sequece$n, type = "l",col = rainbow(per.sequece$n),xlab = "Mean Sequence Quality(Phred Score Sanger Format)", ylab = "Read counts", main = "Quality score distribution over all sequences")

http://goo.gl/H4RTO
Raw data -- Pre-processing

a. Filter out sequences with missing base call
b. Filter out sequences with low scores
c. Trim sequences
d. More preprocessing (adapter removing, etc.)
Raw data -- preprocessing

**FASTX-Toolkit:** a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

**Available Tools (13)**

- **FASTQ Quality Filter**
  Filters sequences based on quality
- **FASTQ Quality Trimmer**
  Trims (cuts) sequences based on quality
- **FASTQ Masker**
  Masks nucleotides with 'N' (or other character) based on quality
- **FASTQ/A Clipper**
  Removing sequencing adapters / linkers
- Other 9 tools

[http://hannonlab.cshl.edu/fastx_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)
Outline

1. Brief introduction to RNA-Seq (advantages and challenges)
2. Raw data (format, QC and manipulating)
3. Mapping(SAM, evaluation)
4. Differential Gene expression
   RNASeqDEPipeline by CQS
5. Fusion gene detection
   viewFusion by CQS
6. Resources
Mapping -- Overview

1. Unlike DNA-Seq, when mapping RNA-Seq reads back to reference genome, we need to pay attention to those exon-exon junction reads.

http://www.bgisequence.com/eu/files/8712/9552/3422/rna3.JPG
**Mapping -- tophat**

*Sequence analysis*  

**TopHat: discovering splice junctions with RNA-Seq**

Cole Trapnell¹, Lior Pachter² and Steven L. Salzberg¹

¹Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742 and ²Department of Mathematics, University of California, Berkeley, CA 94720, USA

Received on October 23, 2008; revised on February 24, 2009; accepted on February 26, 2009  
Advance Access publication March 16, 2009

Associate Editor: Ivo Hofacker

---

**Step 1: Direct un-spliced mapping (use bowtie)**

**Step 2: Building splicing junctions**

- Assemble contiguous "coverage regions" as 'exon'
- Predict possible splicing junctions

**Step 3: Use bowtie again to map read to possible splices**

---

Map reads to whole genome with Bowtie  
Collect initially unmappable reads  
Assemble consensus of covered regions  
Generate possible splices between neighboring exons  
Build seed table index from unmappable reads  
Map reads to possible splices via seed-and-extend
Mapping -- Result format (SAM/BAM)

A year ago:
- 14 published aligners, 14 (primitive) alignment formats.
- No generic alignment viewers, no generic variant callers.
- Everyone writes their own pipeline from scratch.

Now (a year after the publication of SAM):
- Most popular aligners generate a single format: SAM.
- 10 SAM supported alignment viewers, 3 generic SNP/indel callers.
- Build pipeline upon high-performance tools as well as upon libraries in C, C++, Java, Perl, Python and Ruby.

The Sequence Alignment/Map format and SAMtools

Heng Li¹,‡, Bob Handsaker²,†, Alec Wysoker², Tim Fennell², Jue Ruan³, Nils Homer⁴, Gabor Marth⁵, Goncalo Abecasis⁶, Richard Durbin¹:* and 1000 Genome Project Data Processing Subgroup⁷

---

**Paired-end**
- r001+
- r002+
- r003+
- r004+
- r003-
- r001-

**Ins & padding**
**Soft clipping**
**Splicing**
**Hard clipping**

---

```
coor 12345678901234 5678901234567890123456789012345
ref AGCATGTTAGATA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
```

---

```
rs001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAAGGATACTA *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *
```

---

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>QNAME</td>
<td>Query NAME of the read or the read pair</td>
</tr>
<tr>
<td>2</td>
<td>FLAG</td>
<td>Bitwise FLAG (pairing, strand, mate strand, etc.)</td>
</tr>
<tr>
<td>3</td>
<td>RNAME</td>
<td>Reference sequence NAME</td>
</tr>
<tr>
<td>4</td>
<td>POS</td>
<td>1-Based leftmost POSition of clipped alignment</td>
</tr>
<tr>
<td>5</td>
<td>MAPQ</td>
<td>MAPping Quality (Phred-scaled)</td>
</tr>
<tr>
<td>6</td>
<td>CIGAR</td>
<td>Extended CIGAR string (operations: MIDNSHP)</td>
</tr>
<tr>
<td>7</td>
<td>MRNM</td>
<td>Mate Reference NaMe (=’ if same as RNAME)</td>
</tr>
<tr>
<td>8</td>
<td>MPOS</td>
<td>1-Based leftmost Mate POSition</td>
</tr>
<tr>
<td>9</td>
<td>isize</td>
<td>Inferred Insert SIZE</td>
</tr>
<tr>
<td>10</td>
<td>SEQ</td>
<td>Query SEQUENCE on the same strand as the reference</td>
</tr>
<tr>
<td>11</td>
<td>QUAL</td>
<td>Query QUALity (ASCII-33=Phred base quality)</td>
</tr>
</tbody>
</table>

---

Mapping evaluation

1. Coverage Statistics:
   - number of reads that are properly mapped
   - Among the mapped reads, the percentage of reads in exons, introns, etc.

2. Coverage quality control
   - 5' or 3' bias due to poly-A tail selection.
   - The percentage of expressed genes. (too low means library construction may have failed)
RNA-SeqQC: RNA-seq metrics for quality control and process optimization

David S. DeLuca*, Joshua Z. Levin, Andrey Sivachenko, Timothy Fennell, Marc-Danie Nazaire, Chris Williams, Michael Reich, Wendy Winckler and Gad Getz*
The Broad Institute of MIT and Harvard, Cambridge, MA, USA

- Read Metrics
  - Total, unique, duplicate reads
  - Alternative alignment reads
  - Read Length
  - Fragment Length mean and standard deviation
  - Read pairs: number aligned, unpaired reads, base mismatch rate for each pair mate, chimeric pairs
  - Vendor Failed Reads
  - Mapped reads and mapped unique reads
  - rRNA reads
  - Transcript-annotated reads (intragenic, intergenic, exonic, intronic)
  - Expression profiling efficiency (ratio of exon-derived reads to total reads sequenced)
  - Strand specificity

- Coverage
  - Mean coverage (reads per base)
  - Mean coefficient of variation
  - 5'3' bias
  - Coverage gaps: count, length
  - Coverage Plots

- Downsampling
- GC Bias
- Correlation:
  - Between sample(s) and a reference expression profile
  - When run with multiple samples, the correlation between every sample pair is reported

https://confluence.broadinstitute.org/display/CGATools/RNA-SeQC
The following plot shows the mean coverage for expressed transcripts from 5' to 3' end, with the lengths of transcripts normalized to 1-100.

No 5' or 3' bias

5' bias
Outline

1. Brief introduction to RNA-Seq (advantages and challenges)
2. Raw data (format, QC and manipulating)
3. Mapping(SAM, evaluation)
4. **Differential Gene expression**
   RNASeqDEPipeline by CQS
5. Fusion gene detection
   viewFusion by CQS
6. Resources
Differential Gene expression

1. Based on FPKM (cuffdiff)

\[ FPKM = \frac{\text{Counts of mapped fragments}}{\text{Total mapped fragments (million) } \times \text{Exon length of transcript (KB)}} \]

2. Based on count table (R packages like DESeq, baySeq, etc.)
What is cuffdiff

Cuffdiff is a program that uses the Cufflinks transcript quantification engine to calculate gene and transcript expression levels in more than one condition and test them for significant differences.

What are those R packages

1. **DESeq** -- based on negative binomial distribution
2. **edgeR** -- use an overdispersed Poisson model
3. **baySeq** -- use an empirical Bayes approach
4. **TSPM** -- use a two-stage Poisson model
1. Our goal
   a. Make it simple
   b. Reproducible

2. What's the input
   a. A configure file

3. Where you can get it
   a. https://github.com/riverlee/RNASeqDEPipeline
An example Using RNASeqDEPipeline

1. Test data

John etc. paper (RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays)

2. The Configure file

```bash
[[general]]
#Bowtie index file for tophat
BowtieIndex = /demo/hg19/bowtie/hg19
#A file describes all the input files
FileList = files.list
#GTF format file for specified species, could be download from ensembl
TranscriptGtf = /demo/hg19/Homo_sapiens.gtf
#Threads for running tophat simultaneously
Threads = 1

[[tophat]]
#parameter setting for tophat
#details are available at http://tophat.cbcb.umd.edu/manual.html
#Non-recognized parameters will generate warnings
#-h/--help, -v/--version, -o/--output-dir will not be recognized
#Use this many threads to align reads
-p = 8

[[cuffdiff]]
#parameter setting for cuffdiff
#details are available at http://cufflinks.cbcb.umd.edu/manual.html#cuffdiff
#Non-recognized parameters will generate warnings
#-h/--help, -v/--version, -o/--output-dir will not be recognized
#Use this many threads to run cuffdiff
-p = 8
--upper-quartile-norm=
```

files.list

```bash
#Use data from paper RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays.
Genome Research 2008 Sep;18(9):1509-17.
#Here only use data on run 2, with concentration 1.5pm
#whichend value in 0,1,2; 0 corresponding to $paired=0 (single)
#sample group is_paired whichend filename
R2L4Kidney 0 0 0 /demo/rawdata/kidney1.5pm_run2_lane4.txt
R2L8Kidney 0 0 0 /demo/rawdata/kidney1.5pm_run2_lane8.txt
R1L1Kidney 0 0 0 /demo/rawdata/kidney3pm_run1_lane1.txt
R1L3Kidney 0 0 0 /demo/rawdata/kidney3pm_run1_lane3.txt
R1L7Kidney 0 0 0 /demo/rawdata/kidney3pm_run1_lane7.txt
R2L2Kidney 0 0 0 /demo/rawdata/kidney3pm_run2_lane2.txt
R2L6Kidney 0 0 0 /demo/rawdata/kidney3pm_run2_lane6.txt
R2L1Liver 1 0 0 /demo/rawdata/liver1.5pm_run2_lane1.txt
R2L7Liver 1 0 0 /demo/rawdata/liver1.5pm_run2_lane7.txt
R1L2Liver 1 0 0 /demo/rawdata/liver3pm_run1_lane2.txt
R1L4Liver 1 0 0 /demo/rawdata/liver3pm_run1_lane4.txt
R1L6Liver 1 0 0 /demo/rawdata/liver3pm_run1_lane6.txt
R1L8Liver 1 0 0 /demo/rawdata/liver3pm_run1_lane8.txt
R2L3Liver 1 0 0 /demo/rawdata/liver3pm_run2_lane3.txt
```
Differential expressed gene (DEG) report (RNA-Seq) by RNASeqDEPipeline

Table of contents
1. Alignment (Tophat)
2. DEG by Cuffdiff
3. DEG analysis in R
   1. Estimate count level expression
   2. DEG detection by DESeq, edger and baySeq
   3. Significant DEG overlaps among different methods
4. Session info

1. Alignment (Tophat)

TopHat is a fast splices junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

Running Time: 00:03:57.46

Running Code:

```
# TopHat
# o tophat/RLLKidney/demo/hg19-bowtie/hg19/demo/rawdata/kidney3m_run1_lane1.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run1_lane2.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run1_lane3.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run1_lane4.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run1_lane5.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run1_lane6.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run1_lane7.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run1_lane8.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run2_lane1.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run2_lane2.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run2_lane3.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run2_lane4.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run2_lane5.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run2_lane6.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run2_lane7.txt
# o tophat/RLLLiver/demo/hg19-bowtie/hg19/demo/rawdata/liver3m_run2_lane8.txt
```

2. DEG by Cuffdiff

Cuffdiff tests for differential expression and is part of cufflinks which assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples.

Result Files
1. Differentially expressed genes
2. Differentially expressed transcripts

Running Time: 00:00:32.00

Running Code:

cuffdiff --upper-quartile-norm -p 8 -o DE/cuffdiff /demo/annotation/hg19/Refseq_sapiens.gtf
tophat/RLLKidney/accepted_hits.bam, tophat/RLLKidney/accepted_hits.bam, tophat/RLLKidney/accepted_hits.bam,
tophat/RLLLiver/accepted_hits.bam, tophat/RLLLiver/accepted_hits.bam, tophat/RLLLiver/accepted_hits.bam,
3.2 Differential gene expression by DESeq, edgeR, baySeq and TSPM

**Method brief description**


**Results files**

1. **DE genes by P**
2. **DE transcripts by P**

**Running code**

```r
source('/home/lij17/Dropbox/Documents/RNA-Seq/RNASEqDEPipline/DEUtl.R')
group <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3)
mc.cores <- 4
degeneoutfilename <- 'DE/R/de_gene_expr_count.csv'
detranscriptoutfilename <- 'DE/R/de_transcript_expr_count.csv'

## Do differential gene expression
countData <- read.csv('DE/R/de_gene_expr_count.csv', row.names = 1)
system.time(wrapDourDE(countData, group, degeneoutfilename, mc.cores - mc.cores))

## Do differential transcript expression
countData <- read.csv('DE/R/de_transcript_expr_count.csv', row.names = 1)
system.time(wrapDourDE(countData, group, detranscriptoutfilename, mc.cores - mc.cores))
```

4. **Session Info**

```r
sessionInfo()

# R version 2.14.1 (2011-12-22)
# Platform: x86_64-pc-linux-gnu (64-bit)
#
# locale:
# LC_CTYPE=en_US.UTF-8
# LC_NUMERIC=C
# LC_TIME=en_US.UTF-8
# LC_COLLATE=en_US.UTF-8
# LC_MONETARY=en_US.UTF-8
# LC_MESSAGES=en_US.UTF-8
# LC_PAPER=C
# LC_NAME=C
# LC_ADDRESS=C
# LC_TELEPHONE=C
# LC_MEASUREMENT=en_US.UTF-8
# LC_IDENTIFICATION=C
```
Outline

1. Brief introduction to RNA-Seq (advantages and challenges)
2. Raw data (format, QC and manipulating)
3. Mapping(SAM, evaluation)
4. Differential Gene expression
   RNASeqDEPipeline by CQS
5. **Fusion gene detection**
   viewFusion by CQS
6. Resources
Fusion Gene-- overview

A fusion gene is a gene made by joining parts of two different genes

Fusion Gene-- How to detect from RNA-Seq (FusionHunter)

Yang Li¹, Jeremy Chien³, David I. Smith³ and Jian Ma¹,2,*

¹Department of Bioengineering, ²Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801 and ³Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA
Fusion Gene -- Visualization

Example source: FusionHunter supplement based on Berger et al. (2010) datasets
Fusion Gene -- Visualization

1. The solution is 'circos' plot
2. viewFusion at https://github.com/riverlee/viewFusion/

perl viewFusion.pl -i FusionHunter.result
Outline

1. Brief introduction to RNA-Seq (advantages and challenges)
2. Raw data (format, QC and manipulating)
3. Mapping(SAM, evaluation)
4. Differential Gene expression
   RNASeqDEPipeline by CQS
5. Fusion gene detection
   viewFusion by CQS
6. Resources
Resources

  List software packages for next generation sequence analysis
- [http://manuals.bioinformatics.ucr.edu/home/ht-seq](http://manuals.bioinformatics.ucr.edu/home/ht-seq)
  Give examples of R codes to deal with next generation sequence data
  A blog publishes news related to RNA-Seq analysis.
- [http://www.bioconductor.org/help/workflows/high-throughput-sequencing](http://www.bioconductor.org/help/workflows/high-throughput-sequencing)
  Give examples using bioconductor for sequence data analysis
- [http://goo.gl/H4RTO](http://goo.gl/H4RTO)
  An example of QC and DE Gene for RNA-Seq data in R
- [https://github.com/riverlee/RNASeqDEPipeline](https://github.com/riverlee/RNASeqDEPipeline)
  RNASEqDEPipeline
- [https://github.com/riverlee/viewFusion/](https://github.com/riverlee/viewFusion/)
  viewFusion
Thank you