RNA-Seq strategies

• Align against a known transcriptome:
  – **good**: efficient, well defined answers
  – **bad**: unable to discover novel transcripts, may align reads that would map better in noncoding regions

• Align against genome:
  – **good**: discover novel transcripts
  – **bad**: more false positives, more uncertainty

Many methods try to make use of a combination of both
RNA-Seq against known transcriptome

Simplest approach (somewhat frowned upon):

• Extract transcripts sequences and treat them as the “reference” (bedtools getfasta gfffile)

• OR: create a database of all possible known exon combinations

• Use a short read aligner to against this reference

• Post-process the results (some programming may be required)
Alexa-Seq approach

Alternative expression analysis by RNA sequencing

Malachi Griffith, Obi L Griffith, Jill Mwenifumbo, Rodrigo Goya, A Sorana Morrissy, Ryan D Morin, Richard Corbett, Michelle J Tang, Ying-Chen Hou, Trevor J Pugh, Gordon Robertson, Suganthi Chittaranjan, Adrian Ally, Jennifer K Asano, Susanna Y Chan, Haiyan I Li, Helen McDonald, Kevin Teague, Yongjun Zhao, Thomas Zeng, Allen Delaney, Martin Hirst, Gregg B Morin, Steven J M Jones, Isabella T Tai et al.

Affiliations | Contributions | Corresponding author

Received 04 May 2010 | Accepted 20 August 2010 | Published online 12 September 2010
Alexa-Seq Pipeline
(well it is complicated)
RNA-seq against genome

Methodology

1. Separate reads (read-pairs) that map “correctly” to exonic locations

2. Process “incorrectly” mapped reads:
   – Distant pairs could indicate the presence of an intron

3. Re-align reads that did not map in step 1 to “potential” junction sites
   – create a putative transcriptome by fusing sequences at the border of mapped reads
   – identify intron splicing indicator base pairs: GT --- AG, etc.
   – train machine learning algorithms to predict junction sites
Download and install **tophat** and **cufflinks**

• There are great many rnaanalysis packages. Many of which don’t seem work at this moment.

• A consistently good performer, a tools that works immediately with minimal fuss and therefore is becoming the almost a de-facto standard is called the **tuxedo** package:

  – **bowtie** → short read mapper
  – **tophat** → RNA-seq mapping
  – **cufflinks** → counting the transcripts
    • **cuffdiff** → establish differential expression
    • **cuffmerge** → merge experiments
  – **cummeRbund** → R package to facilitate RNA-Seq analysis
Exercise: produce the transcriptome
Generate transcript database

```bash
# create directory to store data
mkdir -p data

# the gffread tool from cufflinks generates the
# transcripts sequences based on the GFF file
gffread -w data/demo.fa -g refs/mini.fa demo.gff
```

```bash
ialbert@porthos ~/work/lec27
$ head data/demo.fa
>T1
ATACTATATATCTTTTTCCATCATTCCATATGCTAACCAGCAATATCCCTAAAAGCATATACTGATGATCTTT
TAATCTTGATGTGACACTACTCATAAGAAGGACTATATCTAGTCAAGACGATACTGTGATAGGTACGT
TTAAATAGGATCTATAAGAAGATATTATTTTTTTCTACGGTGATATTAAATCTATGATAGGTACGT
TATCGAGCTTTATTTCCTCGTAGTCACTGACTCCCTCCTCCTTCCTCTCACTGATGATGATGATGATG
TATCGAGCTTTATTTCCTCGTAGTCACTGACTCCCTCCTCCTTCCTCTCACTGATGATGATGATGATG
```

Data generation script

```bash
# create directory to store data
mkdir -p data

# the gffread tool from cufflinks generates the
# transcripts sequences based on the GFF file
# simulator reads from this transcript
# we only need to cover the transcript so we
# don't need that many reads
wgsim -N 500 -d 100 -l 50 -Z 50 data/demo.fa data/s1.fq data/s2.fq

# build the index
# align against this index
sh bowtie-aln.sh data/s1.fq data/s2.fq refs/mini
```
This already exhibits the underlying structure – too many errors though.
Map against the transcriptome

```bash
# create directory to store data
mkdir -p data

# the gffread tool from cufflinks generates the transcripts sequences based on the GFF file
gffread -w data/demo.fa -g refs/mini.fa demo.gff

# simulate the reads from this transcript
# we only need to cover the transcript so we don't need that many reads
wgsim -N 500 -d 100 -1 50 -2 50 data/demo.fa data/s1.fq data/s2.fq

# build the index
bowtie2-build data/demo.fa refs/demo

# align against this index
sh bowtie-aln.sh data/s1.fq data/s2.fq refs/demo
```
Visualize against the transcriptome

Import the transcript as your reference genome
Align with tophat

```bash
# create directory to store data
mkdir -p data

# the gffread tool from cufflinks generates the
# transcripts sequences based on the GFF file
# gffread tool requires the GFF file
# gffread -w data/demo.fa -g refs/mini.fa demo.gff

# simulate the reads from this transcript
# we only need to cover the transcript so we
# don't need that many reads
# wg_sim -N 500 -d 100 -1 50 -2 50 data/demo.fa data/s1.fq data/s2.fq

# align with tophat
# tophat tool requires the GFF file
# tophat -G demo.gff -r 100 refs/mini data/s1.fq data/s2.fq

# create an index of the results file
# samtools index tool_out/accepted_hits.bam
```
Mapping with TopHat
Simulating realistic reads

Modelling and simulating generic RNA-Seq experiments with the flux simulator

Thasso Griebel\textsuperscript{1,2}, Benedikt Zacher\textsuperscript{2,3}, Paolo Ribeca\textsuperscript{1,4}, Emanuele Raineri\textsuperscript{5}, Vincent Lacroix\textsuperscript{1,6}, Roderic Guigó\textsuperscript{1} and Michael Sammeth\textsuperscript{1,2,}*

+ Author Affiliations

*To whom correspondence should be addressed. Tel: +34 696 635 659; Fax: +34 931 761 537; Email: micha@sammeth.net

Received October 23, 2011.\nRevision received May 25, 2012.\nAccepted June 16, 2012.

Everyone investing into RNA-Seq analysis needs to start here. It explains in great detail the source of experimental bias and measurement error.

Simulate and analyze known transcripts of your genome of interest before diving into analyzing real data.
Flux-Generator

• **Positive:** responsive authors, good underlying concepts

• **Negative:**
  1. it is more complicated than it should be,
  2. invents its own data formats

*You shall not invent a new weakly defined, internally redundant, ambiguous, bulky fruit salad of a data format. Again.*

From the Biostar thread: **What are the most common stupid mistakes in bioinformatics?**
Run the flux simulator

• Turn the transcript into GFF into GTF

• Put the references into a separate file per each chromosome

• Simulate each step:
  – Simulate library construction
  – Simulate gene expression
  – Simulate sequencing
Simulation Parameters

```
1. REF_FILE_NAME /Users/ialbert/work/lec27/demo.gtf
2. GEN_DIR /Users/ialbert/work/lec27/refs/fasta
3. NB_MOLECULES 10000
4. READ_NUMBER 1000
5. READ_LENGTH 76
6. FASTA true
7. ERR_FILE 76

# run the simulator for based on the parameters
~src/flux-simulator-1.1/bin/flux-simulator -l -x -s -p sim/simulate.par --force

# align with tophat
tophat -G demo.gtf refs/mini sim/simulate.fastq

# create an index of the results file
samtools index tophat_out/accepted_hits.bam
```
Simulation results