RNA-seq with reference genome: What about transcripts reconstruction?

Coline Billerey, Alexandre Cormier, Matthieu Defrance, Marc Deloger, Justine Guegan, Carl Herrmann, Rachel Legendre, Denis Puthier, Matthias Zytnicki
Splicing variants analysis

RNA-Seq reads

Align reads to genome

Assemble transcripts de novo

Genome

Assemble transcripts from spliced alignments

More abundant

Less abundant

Erwan CORRE : Jeudi 09/10
14h-17h30
Splicing variants analysis

ALREADY DONE

Bowtie
Extremely fast, general purpose short read aligner

TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

Cufflinks package

Cufflinks
Assembles transcripts

Cuffcompare
Compares transcript assemblies to annotation

Cuffmerge
Merges two or more transcript assemblies

Trapnell & al, Nature protocol (2012) "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks"
Objective of the project: Identify splicing variants differentially expressed in 2 human cell lines

- Data retrieved from the ENCODE project: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeCaltechRnaSeq

- 2 human cell lines:
  - Gm12878 (lymphoblastoid cell line)
  - Hct116 (colorectal carcinoma cell line)

- Each experiment is performed 2 times leading to a total of 4 samples.

- Illumina sequencing, paired-ends 2x75 bp, insert size ~200 bp - the selected genome region is the chromosome 22.
Splicing variants analysis: step 1

Get Data:

• Import published history called "Splicing_Variants_TP_Datasets"

WARNING: ALL THE FOLLOWING DIFFERENT STEPS HAS TO BE DONE ON THE 4 SAMPLES
Splicing variants analysis : step2

Transcript assembly with Cufflinks:

SAM or BAM file of aligned RNA-Seq reads:
16: TopHat2 on Gm12678_rep1 accepted_hits

Max Intron Length:
300000

Min Isoform Fraction:
0.1

Pre mRNA Fraction:
0.15

Perform quartile normalization:
No

Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Use Reference Annotation:
No

Perform Bias Correction:
No

Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Use multi-read correct:
No

Tells Cufflinks to do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome.

Use effective length correction:
No

Cufflinks will not employ its ‘effective’ length normalization to transcript FPKM.
Minimum path cover

Transcripts
Splicing variants analysis : step2

Visualization of the "assembled transcripts file" on IGV (Alternative splicing case) => chr22:31,721,000-31,742,000
Splicing variants analysis: step 2

Visualization of the "assembled transcripts file" on IGV (New splicing isoforms case) => chr22:31,518,000-31,531,000
Splicing variants analysis : step2

Visualization of the "assembled transcripts file" on IGV (New assembly case) => chr22:37,340,000-37,370,000
Splicing variants analysis : step 3

Merge of replicates assembled transcripts from Cufflinks with Cuffmerge:

- GTF file produced by Cufflinks:
  - 82: Cufflinks on Gm12878_rep1 assembled transcripts

- Additional GTF Input Files 1
  - GTF file produced by Cufflinks:
    - 86: Cufflinks on Gm12878_rep2 assembled transcripts
Splicing variants analysis: step 4

Comparison of assembled transcripts from Cufflinks with known transcripts (RefSeq) with Cuffcompare:

- GTF file produced by Cufflinks: 116: Cuffmerge on Gm12878_rep1 and Gm12878_rep2 merged transcripts
- Add new Additional GTF Input Files
- Use Reference Annotation: Yes
- Reference Annotation: 5: chr22.gtf
  Requires an annotation file in GFF3 or GTF format.
- Ignore reference transcripts that are not overlapped by any transcript in input files: ✓
- Use Sequence Data: Yes
- Choose the source for the reference list: History
- Using reference file: 7: chr22.fasta

Execute
Splicing variants analysis : step4

Comparison of assembled transcripts from Cufflinks with known transcripts (RefSeq) with Cuffcompare:

Annotation (gff or gtf)

Cufflinks results

Cuffcompare results

\[ o = u \]
Splicing variants analysis: step 5

Understand the Cuffcompare classification of the transcripts:

<table>
<thead>
<tr>
<th>Priority</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>=</td>
<td>Complete match of intron chain</td>
</tr>
<tr>
<td>2</td>
<td>c</td>
<td>Contained</td>
</tr>
<tr>
<td>3</td>
<td>j</td>
<td>Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript</td>
</tr>
<tr>
<td>4</td>
<td>e</td>
<td>Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment</td>
</tr>
<tr>
<td>5</td>
<td>i</td>
<td>A transfrag falling entirely within a reference intron</td>
</tr>
<tr>
<td>6</td>
<td>o</td>
<td>Generic exonic overlap with a reference transcript</td>
</tr>
<tr>
<td>7</td>
<td>p</td>
<td>Possible polymerase run-on fragment (within 2Kbases of a reference transcript)</td>
</tr>
<tr>
<td>8</td>
<td>r</td>
<td>Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case</td>
</tr>
<tr>
<td>9</td>
<td>u</td>
<td>Unknown, intergenic transcript</td>
</tr>
<tr>
<td>10</td>
<td>x</td>
<td>Exonic overlap with reference on the opposite strand</td>
</tr>
<tr>
<td>11</td>
<td>s</td>
<td>An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)</td>
</tr>
</tbody>
</table>
Splicing variants analysis: IGV

"=" class code

"c" class code
Splicing variants analysis: IGV

"u" class code

"j" class code
Splicing variants analysis: DONE

YAHOO (again)!!!

What else? (BONUS 1)
Isoforms differential expression

Remember Julie Aubert talk (Tuesday 8h30 – 10h)
Isoform detection and quantification

From E. Bernard
Isoforms differential expression

RADIANT : Azra Krek (Memorial Sloan Kettering Cancer Center, New York, USA)
"Comparison of isoform quantification methods for short reads sequencing data using SEQC data" (unpublished)

Methods :

• Which require alignment to the genome: Cufflinks, BitSeq, FluxCapacitor
• Which require alignment to the transcriptome: eXpress, RSEM
• Alignment free : Sailfish
Isoforms differential expression

Data: 2 RNA samples sequenced on different platforms in different labs

Comparison PacBio / Illumina:
• Direct comparison of counts is not ideal
• Better correlation in the comparison of RIE (relative isoform expression) = transcript expression / sum (expression of all transcripts corresponding to this gene)

CONCLUSION:
Methods based on transcriptome alignment (eXpress and RSE M) are generally better
Isoforms differential expression

Top 5 actual softwares:

• RSEM => Erwan Corre talk (Thursday 09/10 16h15-17h30)
• eXpress => has to be installed on galaxy
• DEXSeq => but exon-based method not transcript-based
• Cuffdiff2 : easy to use, Galaxy usable but not the best one unfortunately
Isoforms differential expression

Cuffdiff version 0.0.7

Transcripts:
121: Cuffcompare on Gm12878_rep1_and_rep2 combined transcripts

Conditions
Condition 1
Name: Gm12878
Replicates
Replicate 1
Add replicate:
16: Tophat2 on Gm12878_rep1 accepted_hits
Remove Replicate 1

Replicate 2
Add replicate:
21: Tophat2 on Gm12878_rep2 accepted_hits
Remove Replicate 2

Add new Replicate
Isoforms differential expression

Condition 2

Name:
Hct116

Replicates

Replicate 1

Add replicate:
26: Tophat2 on Hct116_rep1 accepted_hits

Remove Replicate 1

Replicate 2

Add replicate:
31: Tophat2 on Hct116_rep2 accepted_hits

Remove Replicate 2

Add new Replicate

Add new Condition

Library normalization method:
geometric

Dispersion estimation method:
pooled

If using only one sample per condition, you must use 'blind.'

False Discovery Rate:
0.05
The allowed false discovery rate.

Min Alignment Count:
10
The minimum number of alignments in a locus for needed to conduct significance testing on changes in that locus observed between samples.
Isoforms differential expression

Use multi-read correct:
No

Tells Cufflinks to do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome.

Perform Bias Correction:
No

Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Include Read Group Datasets:
No

Read group datasets provide information on replicates.

Set Additional Parameters? (not recommended for paired-end reads):
No

Execute
### Isoforms differential expression

<table>
<thead>
<tr>
<th>test_id</th>
<th>gene_id</th>
<th>gene</th>
<th>locus</th>
<th>sample_1</th>
<th>sample_2</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCONS_00000001</td>
<td>XLOC_000001</td>
<td>-</td>
<td>chr22:16414984-16415930</td>
<td>Gm12878</td>
<td>Hct116</td>
<td>OK</td>
</tr>
<tr>
<td>TCONS_00000002</td>
<td>XLOC_000002</td>
<td>NR_015352</td>
<td>chr22:17518167-17551557</td>
<td>Gm12878</td>
<td>Hct116</td>
<td>OK</td>
</tr>
<tr>
<td>TCONS_00000003</td>
<td>XLOC_000002</td>
<td>NR_015352</td>
<td>chr22:17518167-17551557</td>
<td>Gm12878</td>
<td>Hct116</td>
<td>OK</td>
</tr>
<tr>
<td>TCONS_00000004</td>
<td>XLOC_000002</td>
<td>NR_015352</td>
<td>chr22:17518167-17551557</td>
<td>Gm12878</td>
<td>Hct116</td>
<td>OK</td>
</tr>
<tr>
<td>TCONS_00000005</td>
<td>XLOC_000002</td>
<td>NR_015352</td>
<td>chr22:17518167-17551557</td>
<td>Gm12878</td>
<td>Hct116</td>
<td>OK</td>
</tr>
<tr>
<td>TCONS_00000006</td>
<td>XLOC_000002</td>
<td>NR_015352</td>
<td>chr22:17518167-17551557</td>
<td>Gm12878</td>
<td>Hct116</td>
<td>OK</td>
</tr>
<tr>
<td>TCONS_00000007</td>
<td>XLOC_000002</td>
<td>NR_015352</td>
<td>chr22:17518167-17551557</td>
<td>Gm12878</td>
<td>Hct116</td>
<td>OK</td>
</tr>
<tr>
<td>TCONS_00000008</td>
<td>XLOC_000003</td>
<td>NM_014339</td>
<td>chr22:17565831-17585698</td>
<td>Gm12878</td>
<td>Hct116</td>
<td>OK</td>
</tr>
<tr>
<td>TCONS_00000009</td>
<td>XLOC_000003</td>
<td>NM_014339</td>
<td>chr22:17586749-17594333</td>
<td>Gm12878</td>
<td>Hct116</td>
<td>OK</td>
</tr>
<tr>
<td>TCONS_00000010</td>
<td>XLOC_000003</td>
<td>NM_014339</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>test_id</th>
<th>gene_id</th>
<th>gene</th>
<th>locus</th>
<th>sample_1</th>
<th>sample_2</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCONS_00000011</td>
<td>XLOC_000004</td>
<td>NM_015367</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCONS_00000012</td>
<td>XLOC_000004</td>
<td>NM_001270731</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCONS_00000013</td>
<td>XLOC_000005</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCONS_00000014</td>
<td>XLOC_000005</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCONS_00000015</td>
<td>XLOC_000006</td>
<td>NM_001127649</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCONS_00000016</td>
<td>XLOC_000006</td>
<td>NM_017929</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCONS_00000017</td>
<td>XLOC_000006</td>
<td>NM_017929</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>gene_id</th>
<th>value_1</th>
<th>value_2</th>
<th>log2(fold_change)</th>
<th>test_stat</th>
<th>p_value</th>
<th>q_value</th>
<th>significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLOC_000001</td>
<td>251.205</td>
<td>430.703</td>
<td>0.777829</td>
<td>2.85428</td>
<td>0.0002</td>
<td>0.000765462</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000002</td>
<td>38.7989</td>
<td>7.04635</td>
<td>-2.46107</td>
<td>-1.69074</td>
<td>0.1233</td>
<td>0.219103</td>
<td>no</td>
</tr>
<tr>
<td>XLOC_000002</td>
<td>92.5817</td>
<td>55.7005</td>
<td>-0.733037</td>
<td>-1.56203</td>
<td>0.0379</td>
<td>0.082841</td>
<td>no</td>
</tr>
<tr>
<td>XLOC_000002</td>
<td>48.149</td>
<td>21.2259</td>
<td>-1.81868</td>
<td>-1.88263</td>
<td>0.01165</td>
<td>0.030381</td>
<td>no</td>
</tr>
<tr>
<td>XLOC_000002</td>
<td>57.0838</td>
<td>23.5231</td>
<td>-1.279</td>
<td>-0.921145</td>
<td>0.21875</td>
<td>0.312547</td>
<td>no</td>
</tr>
<tr>
<td>XLOC_000002</td>
<td>47.1751</td>
<td>22.4816</td>
<td>-1.06928</td>
<td>-2.22673</td>
<td>0.002</td>
<td>0.00650512</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000002</td>
<td>49.683</td>
<td>25.9823</td>
<td>-0.935223</td>
<td>-3.0174</td>
<td>5e-05</td>
<td>0.000202766</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000003</td>
<td>294.407</td>
<td>540.92</td>
<td>0.877603</td>
<td>3.93859</td>
<td>5e-05</td>
<td>0.000202766</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000003</td>
<td>184.472</td>
<td>243.054</td>
<td>0.397872</td>
<td>2.40144</td>
<td>0.001</td>
<td>0.00349084</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000003</td>
<td>28.0183</td>
<td>31.8453</td>
<td>0.18471</td>
<td>0.359776</td>
<td>0.6034</td>
<td>0.661726</td>
<td>no</td>
</tr>
<tr>
<td>XLOC_000003</td>
<td>301.092</td>
<td>278.939</td>
<td>-0.110257</td>
<td>-0.577937</td>
<td>0.4066</td>
<td>0.478974</td>
<td>no</td>
</tr>
<tr>
<td>XLOC_000003</td>
<td>128.591</td>
<td>215.398</td>
<td>0.744215</td>
<td>2.3587</td>
<td>0.0019</td>
<td>0.00624379</td>
<td>yes</td>
</tr>
<tr>
<td>XLOC_000003</td>
<td>77.1797</td>
<td>1.05239</td>
<td>-6.19649</td>
<td>-1.18886</td>
<td>0.1865</td>
<td>0.293986</td>
<td>no</td>
</tr>
<tr>
<td>XLOC_000003</td>
<td>189.101</td>
<td>2.67862e-05</td>
<td>-22.6473</td>
<td>-0.0048817</td>
<td>0.21685</td>
<td>0.311585</td>
<td>no</td>
</tr>
<tr>
<td>XLOC_000003</td>
<td>121.554</td>
<td>89.0477</td>
<td>-0.448945</td>
<td>-0.991948</td>
<td>0.15575</td>
<td>0.259947</td>
<td>no</td>
</tr>
<tr>
<td>XLOC_000003</td>
<td>10.3063</td>
<td>5.56992</td>
<td>-0.8878</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1 no</td>
</tr>
</tbody>
</table>

Output named: "transcript differential expression testing"
Isoforms differential expression: DONE

YAHOO (again²) !!!

Want to see more? (BONUS 2)
Variant calling in RNA-seq data using Varscan

Elodie Girard
U900 INSERM - Mines ParisTech - Institut Curie

ITMO Roscoff – 11/19/2013
SNP-calling in RNA-seq data

- Fastq
- Galaxy Format Conversion
  - Groomer
- Reference Genome + Transcriptome
  - Mapping
  - TopHat
- Mpileup
- Variant Calling
  - VarScan
- VCF Annotation
- VCF Filtering

Mapping with TopHat is already done
PCR duplicates are kept in RNA-seq analysis
SNP-calling in RNA-seq data (vSpring/2014)

Data Pre-processing >> Variant Discovery

- Raw RNAseq Reads
  - Map to Ref (STAR) & Add Read Groups
  - Mark Duplicates & Sort (Picard)
  - Split’N’Trim +ReassignMappingQuality
  - Indel Realignment
  - Base Recalibration

Analysis-Ready RNAseq Reads

Variant Calling
- HC in RNAseq mode

Variant Filtering
- see RNAseq-specific docs (separate by variant type)

Filtered Variants
- SNPs
- Indels

SNPs

Indels