Isoform discovery and quantification from RNA-Seq data

Claire Toffano-Nioche, Thibault Dayris

November 2016
Quantification from RNA-Seq data

Previous talk: quantification within the gene level

Quantification from RNA-Seq data

Previous talk: quantification within the gene level

Condition 1

but

Genes may be differentially spliced $\rightarrow$ many different mRNAs from a single locus $\rightarrow$ isoforms

Condition 2

Quantification from RNA-Seq data

And isoforms may be differentially expressed between 2 conditions:

Classification and usage of splicing events

Histogram: ASTalavista\(^1\) + lastests RefSeq versions available of species annotations, ce2, dm3, hg18, tair10 (number of splicing events)

1. Foissac S, Sammeth M (2007) ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in...
A real need?

- transcriptome from new condition
- tissue-specific transcriptome
- different development stages
- transcriptome from non model organism
- cancer cell
- RNA maturation mutant
- ...

How to manage RNA-Seq data with genes subjected to differential splicing?

- Is it possible to discover new isoforms?
- Is it possible to quantify abundance of each isoform?
A real need?

- transcriptome from new condition
- tissue-specific transcriptome
- different development stages
- transcriptome from non model organism
- cancer cell
- RNA maturation mutant
- ...

How to manage RNA-Seq data with genes subjected to differential splicing?

- Is it possible to discover new isoforms? Cufflinks, Cuffmerge
- Is it possible to quantify abundance of each isoform? RSEM, EBSeq
Isoforms reconstruction and quantification from RNA-Seq

- Genomic Annotation (gtf)
- Genomic Sequence (fasta)
- Reads (fastq)

Mapping (TopHat)
- Alignments (bam)
- Annotation Management (Cuffmerge)
- Transcript Management (Cufflinks)

Discovery of new Isoforms
- Enriched Genomic Annotation

Mapping and Quantification (RSeqM)
- Counts
- Workflows
- Differential Expression (EBSeq)

Quantification and differential Expression
- Differentially Expressed Isoforms
RNA-Seq Data: Profiling of sex-biased expression in *Drosophila melanogaster*

- tissue: whole flies
- developmental stage, age: adult, 5-7 days post eclosion
- conditions: sex, female or male

<table>
<thead>
<tr>
<th>SRA¹</th>
<th>Female rep1</th>
<th>Female rep2</th>
<th>Male rep1</th>
<th>Male rep2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRA¹</td>
<td>GSM694258</td>
<td>GSM694259</td>
<td>GSM694260</td>
<td>GSM694261</td>
</tr>
</tbody>
</table>
- PolyA+ mRNA, paire-ends 2x75bp, insert size +/- 200bp
- genome reduction: chr 3R (autosom), from 377 to 13 947 890, biological duplicate

TP 1st step: Data importation from Published histories

Welcome to Gustave Roussy's Galaxy!!!

You're currently connected to prod environment.
Isoforms reconstruction protocol

Discovery of new transcript

Genomic Annotation (gtf) → Genomic Sequence (fasta) → Reads (fastq) → Mapping (TopHat) → Alignments (bam) → Transcript Management (Cufflinks) → Annotation Management (Cuffmerge) → Enriched Genomic Annotation

Tuxedo suite:

- **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

TP 2\textsuperscript{nd} step: Cufflinks

**Parameters**

- "SAM or BAM file of aligned RNA-Seq reads": Your mapping file
- "Use Reference Annotation": Set to "Use reference annotation as guide"
- "Reference Annotation": Your genome annotation
- "Use effective length correction": No

We are interested in Isoform detection, but not in their quantification

[✓ Execute]
Cufflinks algorithm

Our cufflinks usage

- for the discovery of isoform and without quantification aims no matter of the parameters related to quantification (normalization, length correction)

- 2 thresholds (signal to noise ratio), isoform and splicing event:
  - minimum expression ratio: given isoform / majority isoform
  - number of reads ratio: splicing site / intron

- with a well-known genome (fruitfly) "use reference annotation as guide" (but may be used with no reference annotation)

Merge transcripts from many samples

- Cufflinks done for each sample → different lists of transcripts
- necessary to unify lists between them
- in connection with the reference annotations

TP 3rd step: Cuffmerge

Parameters

- "GTF file produced by Cufflinks": Your first genomic annotation produced by Cufflinks (gtf)
- "Additional GTF Input Files": Repeat up to your last annotation!
- "Use Reference Annotation": Set to "Yes", then insert your initial genomic annotation (gtf)
- "Use Sequence Data": Set it to "Yes"
- "Choose the source for the reference list": Set it to "History"
- "Using Reference file": Your genomic sequence (fasta)

✔️ Execute
Understanding the "cuff" 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>

http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/
The following reads are mapped to an existing transcript in the fly genome (here female sample 2 and male sample 1) without any differential expression, nor differential processing.
Another example of "=" class which is differentially expressed in relation to the male condition.
"j" class code

The following reads are mapped to a part of an existing transcript. This is a potential novel isoform in female sample.
"u" class code

The following reads are mapped to an intergenic region.
Isoforms reconstruction and quantification from RNA-Seq
Differential expression, transcript level

Isoforms differential expression: RSEM

RSEM aligns reads on a transcript reference:

- computed from the genome annotations (gtf file)
- directly from transcript assembly (in case of non-model organism, cancer cell, etc)

RSEM adds a polyA tail to each transcript (reads from 3’ end mRNA) and uses indexation (gain time)

RSEM prepare reference
TP 4\textsuperscript{th} step: RSEM prepare reference

**Parameters**

- "Reference transcript source": Set it to "reference genome and gtf"
- "reference fasta file": Your genome sequence (fasta)
- "gtf or gff3 file": Your enhanced and merged genome annotation
- "Use Bowtie2": Hit "Yes"
RSEM features

RSEM estimates the uncertainty due to both multiread allocation and random sampling effect using all valid mappings of the read (mapping scores, probability for a read to come from a locus)

Need of a specific mapping (sam/bam) file: reporting of **all** the valid mappings for each read

→ relaunch the mapping step (bowtie)

Some RSEM features:

- strand-specificity
- highly 5’ or 3’ biased distribution of read positions
- in case of single-end, fix the fragment length
- does not support gapped mapping (no indel)
**First 3 cycles of EM algorithm.**
Abundance of red isoform estimated after the 1st M-step: 
\[
\frac{1/3 \text{ read } a + 1/2 \text{ read } c + 1 \text{ read } d + 1/2 \text{ read } e}{\text{(total read number)}}, \text{ i.e. } 0.47 \ \frac{(0.33 + 0.5 + 1 + 0.5)}{5}
\]
- proved to converge
- stop criterion: when all probabilities that a fragment is derived from a transcript $\geq 10^{-7}$ have a relative change $\leq 10^{-3}$

RSEM calculate expression

TP 5th step: RSEM isoform abundances

Parameters

- "RSEM Reference Source": Set it to "From your history"
- "RSEM reference": Your previous reference
- "Library type": Set it to "Paired End Reads"
- "Read 1 fastq file" and "Read 2 fastq file": Your reads (fastq)
- "Use bowtie 1 or 2 ?": Set it to Bowtie 2
- "Is the library strand specific?": Set it to forward orientation

 Execute
Empirical Bayesian approach that models a number of features observed in RNA-Seq data.

Runs EBSeq to find DE isoforms across two conditions:
Isoform level DE test across two conditions
Mapping uncertainty increases due to the presence of multiple isoforms of a given gene.

EBSeq:

- Expected count for an isoform is distributed as Negative Binomial
- Isoform-specific means and variances are estimated via the Expectation-Maximization (EM) algorithm
- EBSeq accommodates isoform expression estimation uncertainty by modeling the differential variability observed in distinct groups of isoforms.
- 3 groups: following the number of isoforms associated to each gene (1, 2 or 3 and more)

EBSeq directly models isoform expression

A collective analysis of isoforms:

▶ reduces the power for identifying isoform in the "1" group (the true variance in that group are lower, on average, than those derived from the full collection of isoforms)

▶ increases the false discoveries in the 2 other groups (true variances are higher).

Changes of the estimation uncertainty with the increase of isoform complexity

 Isoforms differential expression: EBseq

Empirical Bayesian approach that models a number of features observed in RNA-Seq data.

2 workflows:

- Create a vector with the related group for each isoform
  Create IG Vector
- 4 RSEM outputs → 1 EBSeq input
  Create Expression Table

Runs EBSeq to find DE isoforms across two conditions:
Isoform level DE test across two conditions
Differential expression, transcript level

Isoforms differential expression: EBSeq

Empirical Bayesian approach that models a number of features observed in RNA-Seq data.

2 workflows:

- Create a vector with the related group for each isoform
  - Create IG Vector
- 4 RSEM outputs → 1 EBSeq input
  - Create Expression Table

Runs EBSeq to find DE isoforms across two conditions:
- Isoform level DE test across two conditions

Arrangement of inputs and outputs of the EBSeq workflow:

- Genomic Annotation (gtf)
- Genomic Sequence (fasta)
- Reads (fastq)
- Mapping and Quantification (RSEM)
- Counts
- Workflows
- Differential Expression (EBSeq)
- Differentially Expressed Isoforms

Claire Toffano-Nioche, Thibault Dayris  Isoform discovery and quantification from RNA-Seq data  November 2016  33 / 61
We have:
4 files (1 per replicate)

Those files are identically ordered by transcript names

We need:
- 1 file containing the number of isoforms each gene owns: IG vector
- 1 file for all expected expression: Expression table

We have to convert RSEM output to fit EBSeq input’s requirements. To do such a thing, we will use two workflows.
TP 6\textsuperscript{th} step: EBSeq IG Vector (1/3)

What is the IG Vector?

- The IG Vector is a table with only one column of numbers (integers).
- Each row corresponds to a transcript on the same row in the Expression table. Each integer in the IG Vector corresponds to the group 1, 2 or 3, according to the number of isoforms of the gene related to the considered isoform.

Tools:

- "Cut" and "Remove beginning" from Text Manipulation section.
- "Get Ig vector from gene-isoform mapping for isoform level DE analysis", available in EBSeq section.
Differential expression, transcript level

TP 6\textsuperscript{th} step: EBSeq IG Vector (2/3)

RSEM Isoform Abundance table

<table>
<thead>
<tr>
<th>transcript_id</th>
<th>gene_id</th>
<th>length</th>
<th>effective_length</th>
<th>expected_count</th>
<th>TPM</th>
<th>FPKM</th>
<th>IsoPct</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCONS_00000001</td>
<td>XLOC_000001</td>
<td>3183</td>
<td>3130.37</td>
<td>114.34</td>
<td>446.42</td>
<td>200.32</td>
<td>96.70</td>
</tr>
<tr>
<td>TCONS_00000002</td>
<td>XLOC_000002</td>
<td>5393</td>
<td>5340.37</td>
<td>6.66</td>
<td>15.22</td>
<td>6.83</td>
<td>3.30</td>
</tr>
<tr>
<td>TCONS_00000003</td>
<td>XLOC_000002</td>
<td>1984</td>
<td>1931.37</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TCONS_00000004</td>
<td>XLOC_000002</td>
<td>1924</td>
<td>1871.37</td>
<td>2.00</td>
<td>13.09</td>
<td>5.87</td>
<td>100.00</td>
</tr>
<tr>
<td>TCONS_00000006</td>
<td>XLOC_000002</td>
<td>2047</td>
<td>1994.37</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TCONS_00000005</td>
<td>XLOC_000003</td>
<td>670</td>
<td>617.55</td>
<td>5.00</td>
<td>100.21</td>
<td>44.97</td>
<td>100.00</td>
</tr>
<tr>
<td>TCONS_00000007</td>
<td>XLOC_000004</td>
<td>2919</td>
<td>2866.37</td>
<td>57.23</td>
<td>244.11</td>
<td>109.54</td>
<td>19.58</td>
</tr>
<tr>
<td>TCONS_00000008</td>
<td>XLOC_000004</td>
<td>2965</td>
<td>2913.37</td>
<td>222.63</td>
<td>934.22</td>
<td>419.21</td>
<td>74.91</td>
</tr>
<tr>
<td>TCONS_00000009</td>
<td>XLOC_000004</td>
<td>2831</td>
<td>2778.37</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TCONS_00000010</td>
<td>XLOC_000004</td>
<td>3100</td>
<td>3047.37</td>
<td>17.13</td>
<td>68.72</td>
<td>30.84</td>
<td>5.51</td>
</tr>
<tr>
<td>TCONS_00000011</td>
<td>XLOC_000004</td>
<td>4414</td>
<td>4361.37</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>TCONS_00000012</td>
<td>XLOC_000005</td>
<td>6303</td>
<td>6250.37</td>
<td>333.12</td>
<td>650.31</td>
<td>291.82</td>
<td>65.46</td>
</tr>
<tr>
<td>TCONS_00000013</td>
<td>XLOC_000005</td>
<td>4305</td>
<td>4252.37</td>
<td>114.08</td>
<td>327.61</td>
<td>147.01</td>
<td>32.98</td>
</tr>
<tr>
<td>TCONS_00000095</td>
<td>XLOC_000005</td>
<td>3964</td>
<td>3911.37</td>
<td>4.99</td>
<td>15.57</td>
<td>6.99</td>
<td>1.57</td>
</tr>
<tr>
<td>TCONS_00000094</td>
<td>XLOC_000006</td>
<td>1310</td>
<td>1257.37</td>
<td>1.00</td>
<td>9.77</td>
<td>4.38</td>
<td>100.00</td>
</tr>
<tr>
<td>TCONS_00000014</td>
<td>XLOC_000007</td>
<td>1546</td>
<td>1493.37</td>
<td>193.00</td>
<td>1585.16</td>
<td>711.31</td>
<td>100.00</td>
</tr>
<tr>
<td>TCONS_00000015</td>
<td>XLOC_000008</td>
<td>1999</td>
<td>1946.37</td>
<td>54.00</td>
<td>339.76</td>
<td>152.46</td>
<td>100.00</td>
</tr>
</tbody>
</table>
TP 6th step: EBSeq IG Vector (3/3)

Parameter

- input: A count table from RSEM.

Caution: All "Isoform_abundances" tabular files have the same succession of transcripts and genes names through each line. This succession is used by the Create_IG_Vector workflow. Therefore, any "Isoform_abundance" file may be used in this step.
Isoforms differential expression: EBseq

Empirical Bayesian approach that models a number of features observed in RNA-Seq data.

2 workflows:
- Create a vector with the related group for each isoform
- Create IG Vector
- 4 RSEM outputs → 1 EBSeq input
- Create Expression Table

Runs EBSeq to find DE isoforms across two conditions:
- Isoform level DE test across two conditions
TP 7th step: RSEM Expression Table

The expression table

5 columns:

- Transcripts name
- The expected expression of F1, F2, M1 and M2

Obtained by merging the 5th column of RSEM Isoform Expression results.

Tool: ”Create_Expression_Table”, available among the shared workflows

parameters

- First Dataset, Second Dataset, Third Dataset, and Fourth Dataset: Your count tables

Execute
Differential expression, transcript level

Differential analysis

Isoforms differential expression: EBseq

Empirical Bayesian approach that models a number of features observed in RNA-Seq data.

2 workflows:

- Create a vector with the related group for each isoform
  Create IG Vector

- 4 RSEM outputs → 1 EBSeq input
  Create Expression Table

Runs EBSeq to find DE isoforms across two conditions:
Isoform level DE test across two conditions
TP 8th step: EBSeq Differential expression

Parameters

- "Isoform Expression": Our Data Matrix
- "The first row is Sample Names": Yes
- "Enter which condition each sample belongs to": M, M, F, F
- "Ig Vector": Our IG Vector

✓ Execute
### List of DE isoforms

<table>
<thead>
<tr>
<th>Gene/IsoformID</th>
<th>PPDE</th>
<th>RealFC</th>
<th>PosteriorFC</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCONS_00000013</td>
<td>1</td>
<td>0.02</td>
<td>0.03</td>
<td>2.3</td>
<td>2.37</td>
<td>99.36</td>
<td>110.07</td>
</tr>
<tr>
<td>TCONS_00000015</td>
<td>1</td>
<td>1.88</td>
<td>1.88</td>
<td>443.99</td>
<td>517.53</td>
<td>263.93</td>
<td>247.78</td>
</tr>
<tr>
<td>TCONS_00000075</td>
<td>0.96</td>
<td>4.06</td>
<td>3.99</td>
<td>76.38</td>
<td>131.07</td>
<td>0</td>
<td>51.04</td>
</tr>
<tr>
<td>TCONS_00000074</td>
<td>0.91</td>
<td>2.18</td>
<td>2.17</td>
<td>183.58</td>
<td>105.47</td>
<td>88.49</td>
<td>43.85</td>
</tr>
</tbody>
</table>

...
"Classical" RNA-Seq analysis method. Many methods (and tools):

Expression estimation:
- Bayesian estimation of parameters of a model: BitSeq, Cufflinks, eXpress
- Expectation-maximization approach to inferring isoform abundances: RSEM-EBseq, Sailfish/Salmon, Kallisto

Mapping to:
- the genome: Cuffdiff2, BitSeq, FluxCapacitor
- the transcriptome: eXpress, RSEM-EBseq
- Mapping-free: Sailfish/Salmon, Kallisto
Mapping-free?

Kallisto example: De Bruijn Graph ("Alignment" talk, day1) on transcriptome

Transcriptome (3 isoformes):

- Red
- Blue
- Green
Mapping-free?

Kallisto example: De Bruijn Graph ("Alignment" talk, day1) on transcriptome

Transcriptome (3 isoformes):

Transcriptome de Bruijn Graph:
Mapping-free?

Kallisto example: De Bruijn Graph ("Alignment" talk, day1) on transcriptome

Transcriptome (3 isoformes):

1 mapped read (non mapping-free method):

Transcriptome de Bruijn Graph:
Mapping-free?

Kallisto example: De Bruijn Graph ("Alignment" talk, day1) on transcriptome

Transcriptome (3 isoformes):

Transcriptome de Bruijn Graph:

1 mapped read (non mapping-free method):
Mapping-free?

Kallisto example: De Bruijn Graph ("Alignment" talk, day1) on transcriptome

Transcriptome (3 isoformes):

1 mapped read (non mapping-free method):

Transcriptome de Bruijn Graph:

Choice of the equivalence class:

EM step change: replace "transcript" by "equivalence class" in the abundance estimation
Mapping-free?

Kallisto example: De Bruijn Graph ("Alignment" talk, day1) on transcriptome

- Stand for "multimap" reads
- Need to adapt algorithm to use stranding RNAseq
- No mapping = no visualization

Isoforms with RNA-Seq: not yet

Isoforms discovery and quantification from RNA-Seq: not yet a well-established measure

▷ Methods based on transcriptome are generally better (for quantification but not for discovery)
▷ EM methods are better than count-based methods (many EM methods are available but differ little in accuracy)
▷ the more abundant is the isoform, the more accurately it is inferred
▷ major bottleneck: small size of read (comparing to 2.2 kb for mammals transcripts), multimap reads

Evaluate the accuracy of isoform abundance computational methods: difficult

▷ too few number of isoform with experimental validation strategies (ex. qRT-PCR)
▷ synthetically generated datasets may not capture adequately the complexities of RNA-Seq experiments
Conclusion

**Improve?**

- don’t forget the micro-arrays designed for isoform detection (not for discovery of new isoform, model organism)
- gain statistical power with ”spike” measurements
- make protocols like ribodepletion but for highly expressed ”housekeeping genes” (to enrich with ”interesting” transcripts)
- complete isoform definitions by other NGS studies?
  - ChIPSeq with a protein from the spliceosome as target
  - capturing the 5’ ends of RNAs ...
- full-length cDNAs technology (Pacific Biosciences)?
  a too low throughput ($10^4$ transcripts, summer 2015)

**Adapt! biological query + organism + data**

- Parameters, softwares, sequencing protocols (single or paired-end, stranded or not)

RNA-Seq is just an unique and sampled RNA capture in a given position, at a given time, of one biological experiment

... a poor quality photo comparing to real life
### Start a new workflow

#### Your workflows

<table>
<thead>
<tr>
<th>Name</th>
<th># of Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create_ALL_GenomeVector_EBA</td>
<td>3</td>
</tr>
<tr>
<td>Create Data Matrix</td>
<td>11</td>
</tr>
<tr>
<td>Fastq_2_FastqSanger</td>
<td>2</td>
</tr>
<tr>
<td>Extract Counts</td>
<td>2</td>
</tr>
<tr>
<td>RSeM</td>
<td>19</td>
</tr>
</tbody>
</table>
Add some details

Both name and annotation are important for your own workflows management

Create New Workflow

**Workflow Name:**
Create_IG_Vector_EBA

**Workflow Annotation:**
EBSeq input
A description of the workflow;

Create
Add some details

The workflow is created empty, let us add some tags before diving through tools

Edit Workflow Attributes

Name:
Create_IG_Vector_EBA

Tags:
EBSeq  x  EBA  x  IG_Vector  x

Apply tags to make it easy to search for and find items with the same tag.

Annotation / Notes:
EBSeq input
Add an annotation or notes to a workflow; annotations are available when a workflow is viewed.
**Cut**

### Tools
- **Filter and Sort**
  - Cut columns from a table
- **QC and manipulation**
  - FASTQC:
    - FASTQ/SAM/BAM
  - FastQC: Read QC (v0.10.1) reports using FastQC
- **Pre-processing**
  - Sickle (v1.33)
  - Windowed Adaptive

### Workflow Canvas | Create_IG_Vector_EBA

- **Cut columns from a table**
  - Cut columns
    - **Out file**
      - **out_file1 (tabular)**
  - **Delimited by**
    - **Tab**
  - **From**
    - Data input 'input' (txt)
  - **Annotation / Notes**
    - Keeps columns 1 & 2
Add some actions

**Details**

**Configure Output: 'out_file1'**

**Label**

This will provide a short name to describe the output - this must be unique across workflows.

**Rename dataset**

Transcript to Gene Table

This action will rename the output dataset. Click [here](#) for more information. Valid inputs are: **input**.

**Change datatype**
Remove Beginning

Text manipulation
- Remove beginning of a file
- Column Regex Find
  And Replace

Pre-processing
- cutadapt (v1.3) A tool that removes adapter sequences from DNA sequencing reads, only for fastq format for this version.

Details
- Remove beginning of a file
(Galaxy Version 1.0.0)

Remove first
1 lines
from
Data input 'input' (txt)

Annotation / Notes
Removes the header from the output file: out_file1

Galaxy
Analyze Data  Workflow  Shared Data  Visualization  Help
EBSeq IG Vector

**Tools**
- EBSeq Isoform level DE test across two conditions
  - EBSeq to find DE isoforms across two conditions
- EBSeq Get Ig vector from gene-isoform mapping for isoform level DE analysis
  - Get Ig vector from gene-isoform mapping for isoform level DE analysis

**Workflow Canvas**
- Remove beginning
- from
  - out_file1
- EBSeq Get Ig vector from gene-isoform mapping for isoform level DE analysis
  - Output (tabular)

**Details**
- EBSeq Get Ig vector from gene-isoform mapping for isoform level DE analysis (Galaxy Version 1.0.1)

**Input**
- Data input 'input' (data)
  - Input should be no-header tab-delimited file, first column is isoform names, second column is gene names. See below for more info...
Custom output

- **EBSeq Get Ig vector from gene-isofrom mapping for isoform level DE analysis**

**Label**
This will provide a short name to describe the output - this must be unique across workflows.

**Rename dataset**
IG_Vector
This action will rename the output dataset. Click here for more information. Valid inputs are: input.
End

Do not forget to save!