Using Galaxy for NGS Analysis - RNA-seq

Galaxy

- An easy-to-use, open-source, scalable framework for tool and data integration

- A platform for interactive large-scale genome analysis

Galaxy Architecture

- Galaxy Framework - a set of reusable software components, encapsulating functionality for:
  - describing generic interfaces to computational tools,
  - building concrete interfaces for users to interact with tools,
  - invoking those tools in various execution environments,
  - dealing with general and tool specific dataset formats and conversions, and
  - working with “metadata” describing datasets, tools, and their relationships.

Galaxy Architecture

- Galaxy Application - an application built using this framework that provides access to tools through a web-interface

- GALAXY Instance - a deployment of a Galaxy application with a specific set of tools. The GALAXY team maintains such an instance that integrates a large set of tools for comparative genomics.

Galaxy Framework Core Components

- *Toolbox*
  - manages all of the details of working with command-line and web-based computational tools
  - parses GALAXY tool configuration files that describe the interface to a tool – the parameters and input data it can take, their types and restrictions, and the outputs it produces – in an abstract way that is not specific to any particular user interface
  - provides support for validating inputs to a tool, and for transforming a valid set of inputs into the commands necessary to invoke that tool.

Galaxy Framework Core Components

- *Job Manager*
  - manages dependencies between jobs (invocations of tools) to ensure that required datasets have been produced without errors before a job is run
  - provides support for job queuing, to allow multiple users to each submit multiple jobs to a GALAXY instance and receive a fair execution order
  - “pluggable” underlying method for execution.

Galaxy Framework Core Components

- *Model*
  - provides an object-oriented interface for working with dataset content (stored as files on disk) and "metadata" (data about datasets, tools, and their relationships; stored in a relational database)
  - deals with support for different datatypes, datatype specific metadata, and type conversions.

Galaxy Framework Core Components

- *Web Interface*
  - provides support for interacting with a GALAXY instance through a web browser
  - generates web-based interfaces to:
    - the toolbox (for browsing and choosing tools),
    - individual tools (building forms to accept and validate user input to a tool), and
    - the model (allowing the user to work with all of the datasets they have produced).

Galaxy Implementation Details

- The GALAXY framework is implemented in the Python programming language.
- Python is not a requirement for tool authoring; GALAXY toolbox interacts with tools through command-line and web-based interfaces.
- GALAXY includes its own web server and embedded relational database (SQLite), and a GALAXY download includes all dependencies.
- The web server, the underlying relational database, and the job execution mechanism can be customized for a GALAXY instance to support higher throughput (e.g. the public GALAXY instance maintained by the GALAXY team at Penn State is integrated with Apache as the web-server, uses the enterprise class relational database PostgreSQL, and executes jobs on a computational cluster with a queue managed by Torque Portable Batch System [PBS]).

Galaxy NGS Analysis Toolsets

- an open and free web-based platform for performing accessible, reproducible, and transparent NGS analyses
  - Prepare, Quality Check and Manipulate FASTQ reads
  - Mapping
  - SAMTools
  - SNP and INDEL Analysis
  - Peak Calling/Chip-seq
  - RNA-seq Analysis

NGS Data

- Raw (Sequencing reads): FASTQ
- Derived
  - Alignments against the reference genome (SAM/BAM – sequence/binary alignment/map)
  - Annotations
    - GTF – gene transfer format
    - BED – browser extensible data
  - Genome Assemblies

RNA-seq

- a method for mapping and quantifying the transcriptome of any organism that has a genomic DNA sequence assembly.
- applications:
  - differential gene expression,
  - alternative splicing,
  - fusion gene detection,
  - coding-SNPs,
  - allele-specific expression,
  - novel transcript detection.

RNA-seq

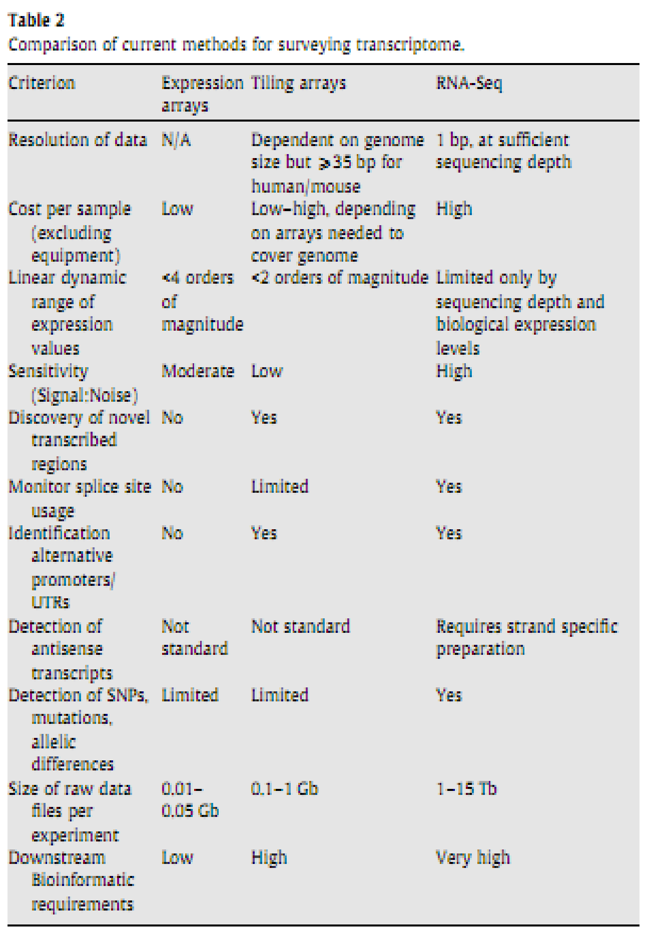


Table 2
Comparison of current methods for surveying transcriptome.

Criterion	Expression arrays	Tiling arrays	RNA-Seq
Resolution of data	N/A	Dependent on genome size but ⩾35 bp for human/mouse	1 bp, at sufficient sequencing depth
Cost per sample (excluding equipment)	Low	Low–high, depending on arrays needed to cover genome	High
Linear dynamic range of expression values	<4 orders of magnitude	<2 orders of magnitude	Limited only by sequencing depth and biological expression levels
Sensitivity (Signal:Noise)	Moderate	Low	High
Discovery of novel transcribed regions	No	Yes	Yes
Monitor splice site usage	No	Limited	Yes
Identification alternative promoters/UTRs	No	Yes	Yes
Detection of antisense transcripts	Not standard	Not standard	Requires strand specific preparation
Detection of SNPs, mutations, allelic differences	Limited	Limited	Yes
Size of raw data files per experiment	0.01–0.05 Gb	0.1–1 Gb	1–15 Tb
Downstream Bioinformatic requirements	Low	High	Very high

RNA-seq

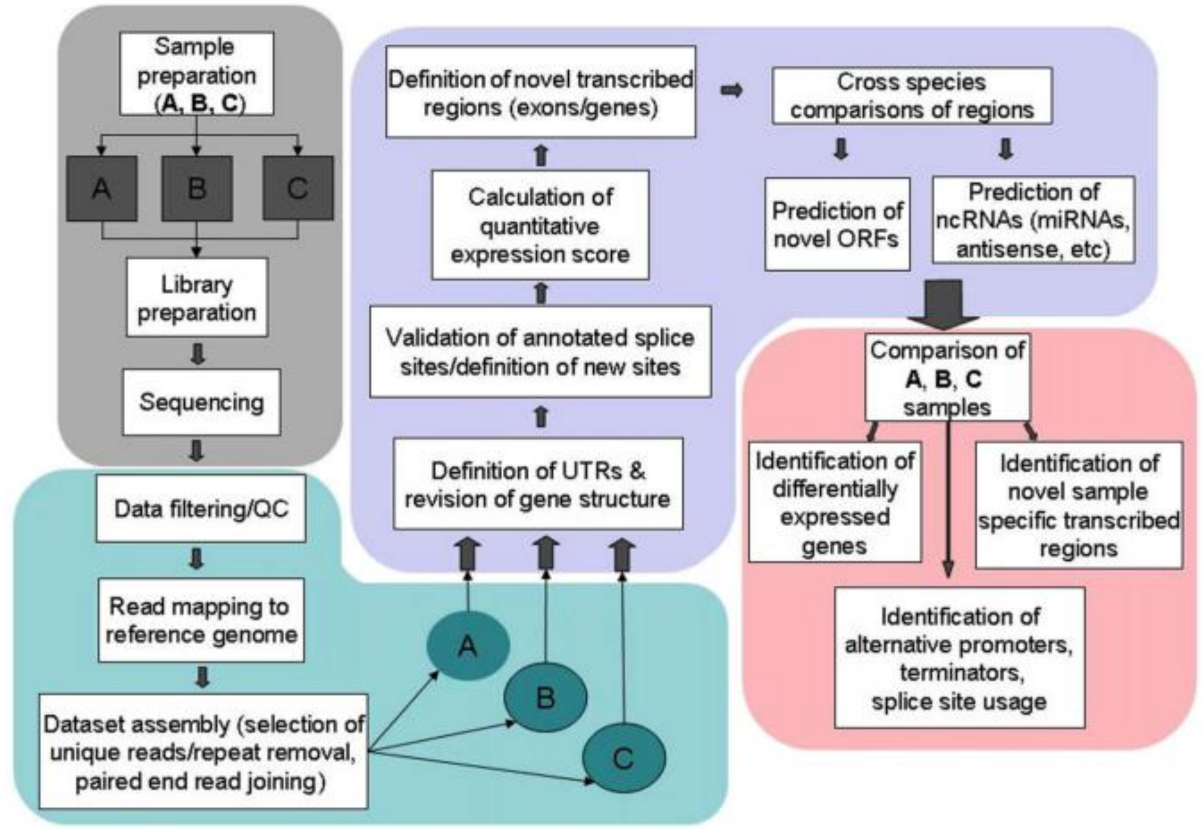
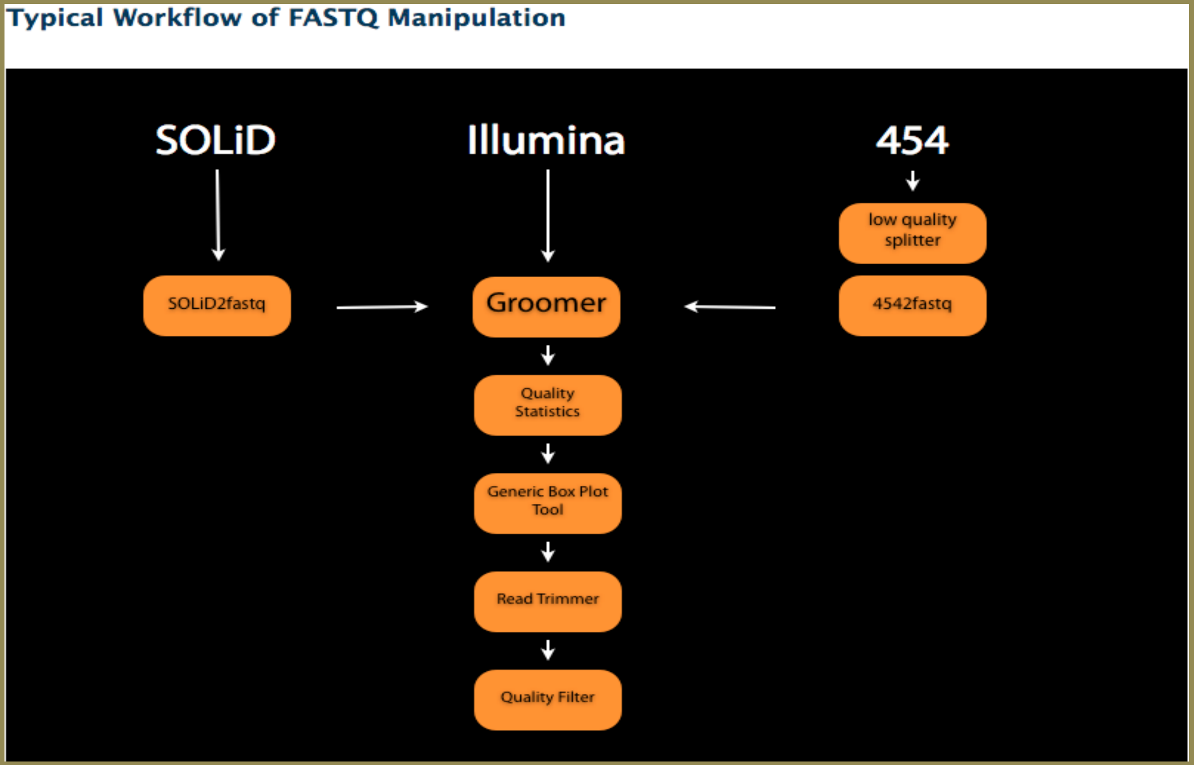

Sample preparation (A, B, C)
A
B
C
Library preparation
Sequencing
Data filtering/QC
Read mapping to reference genome
Dataset assembly (selection of unique reads/repeat removal, paired end read joining)
A
B
C
Definition of UTRs & revision of gene structure
Validation of annotated splice sites/definition of new sites
Calculation of quantitative expression score
Definition of novel transcribed regions (exons/genes)
Cross species comparisons of regions
Prediction of novel ORFs
Prediction of ncRNAs (miRNAs, antisense, etc)
Comparison of A, B, C samples
Identification of differentially expressed genes
Identification of novel sample specific transcribed regions
Identification of alternative promoters, terminators, splice site usage

Fig. 1. Example of RNA-seq work flow. A typical analysis stream for three theoretical biological samples (A, B and C) is shown with the various sections color coded as wet-lab work (grey), creation of the filtered data sets (cyan), sample (tissue/developmental stage/growth condition) specific data analysis (lilac) and cross sample analysis (pink). Much of the later analysis (lilac/pink) will be highly dependent on the experimental aims of the study and as such, only a small fraction of the possible analyses pathways are shown.

NGS QC & Manipulation

Data management and analysis issues:

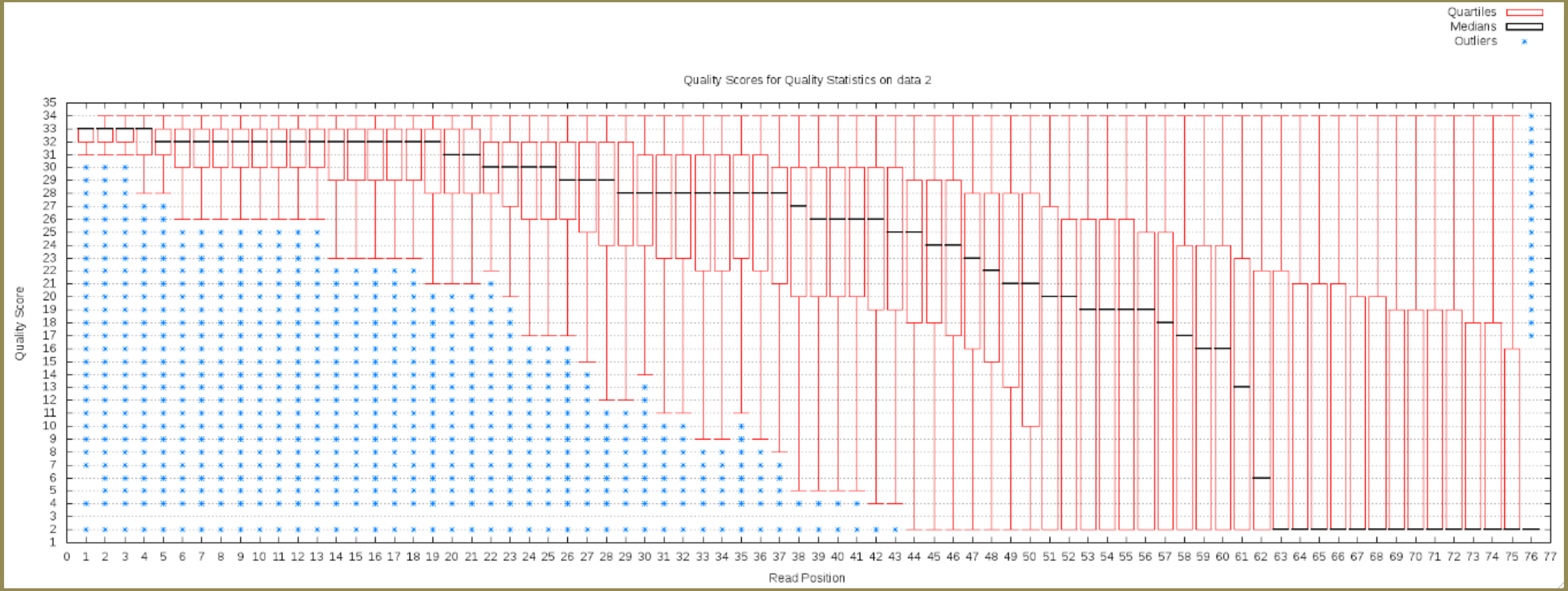
- the lack of standardized sequencer output and tools.
- the de facto standard, FASTQ, comes in a number of distinct variants.
- FASTQ contains sequence data and quality scores.



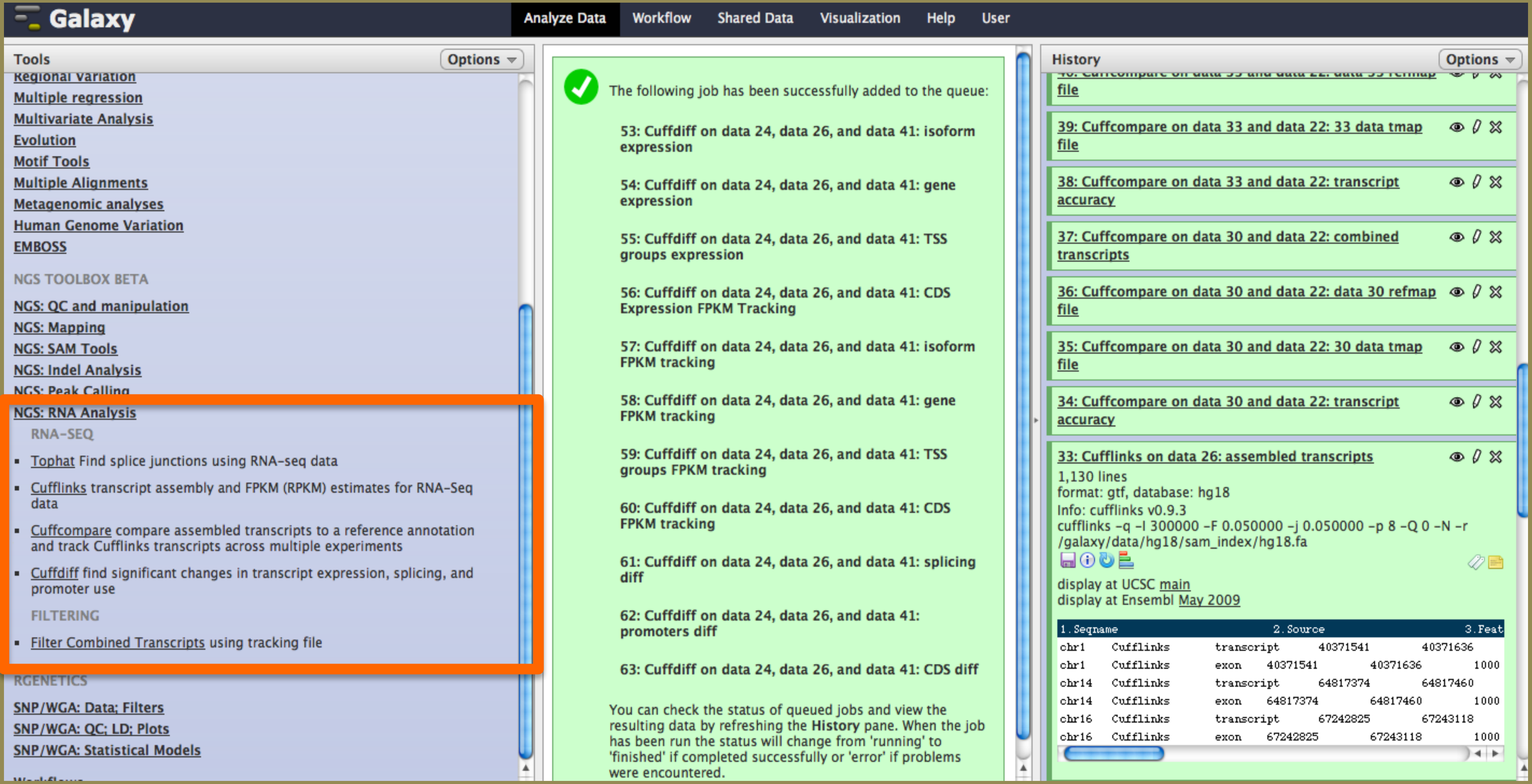
NGS QC & Manipulation

- *FASTQ Groomer* - used to verify and convert between the known FASTQ variants.
- *FASTQ Summary Statistics* – such as read counts, minimums, maximums, sums, means, quartiles with ranges of quality scores, outliers and nucleotide counts for each base position in a FASTQ file – helps determining how to trim and filter read data as quality scores can vary along the length of sequencing reads. Graphed by using the *Boxplot* tool.
- *FASTQ Trimmer* - trim bases from poor-quality ends of reads, to prevent otherwise high-quality reads from being rejected during quality filtering or from influencing mapping or assembly processes.

BoxPlot for FASTQ summary statistics



Galaxy RNA-seq Analysis Toolset



Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Human Genome Variation
EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis

RNA-SEQ

- Tophat Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING

- Filter Combined Transcripts using tracking file

RGENETICS

SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models
Workflows

✓ The following job has been successfully added to the queue:

53: Cuffdiff on data 24, data 26, and data 41: isoform expression

54: Cuffdiff on data 24, data 26, and data 41: gene expression

55: Cuffdiff on data 24, data 26, and data 41: TSS groups expression

56: Cuffdiff on data 24, data 26, and data 41: CDS Expression FPKM Tracking

57: Cuffdiff on data 24, data 26, and data 41: isoform FPKM tracking

58: Cuffdiff on data 24, data 26, and data 41: gene FPKM tracking

59: Cuffdiff on data 24, data 26, and data 41: TSS groups FPKM tracking

60: Cuffdiff on data 24, data 26, and data 41: CDS FPKM tracking

61: Cuffdiff on data 24, data 26, and data 41: splicing diff

62: Cuffdiff on data 24, data 26, and data 41: promoters diff

63: Cuffdiff on data 24, data 26, and data 41: CDS diff

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

40: Cuffcompare on data 33 and data 22: data 33 refmap file

39: Cuffcompare on data 33 and data 22: 33 data tmap file

38: Cuffcompare on data 33 and data 22: transcript accuracy

37: Cuffcompare on data 30 and data 22: combined transcripts

36: Cuffcompare on data 30 and data 22: data 30 refmap file

35: Cuffcompare on data 30 and data 22: 30 data tmap file

34: Cuffcompare on data 30 and data 22: transcript accuracy

33: Cufflinks on data 26: assembled transcripts

1,130 lines
format: gtf, database: hg18
Info: cufflinks v0.9.3
cufflinks -q -I 300000 -F 0.050000 -j 0.050000 -p 8 -Q 0 -N -r /galaxy/data/hg18/sam_index/hg18.fa

display at UCSC main
display at Ensembl May 2009

```
1.Seqname                 2.Source                      3.Feat
chr1    Cufflinks     transcript     40371541      40371636
chr1    Cufflinks     exon     40371541      40371636       1000
chr14   Cufflinks     transcript     64817374      64817460
chr14   Cufflinks     exon     64817374      64817460       1000
chr16   Cufflinks     transcript     67242825      67243118
chr16   Cufflinks     exon     67242825      67243118       1000
```

TopHat

- a fast splice junction mapper for RNA-Seq reads.
- 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.
- outputs:
  - accepted_hits -- A list of read alignments in BAM format and,
  - splice junctions -- A UCSC BED track of junctions, each consisting of two blocks.

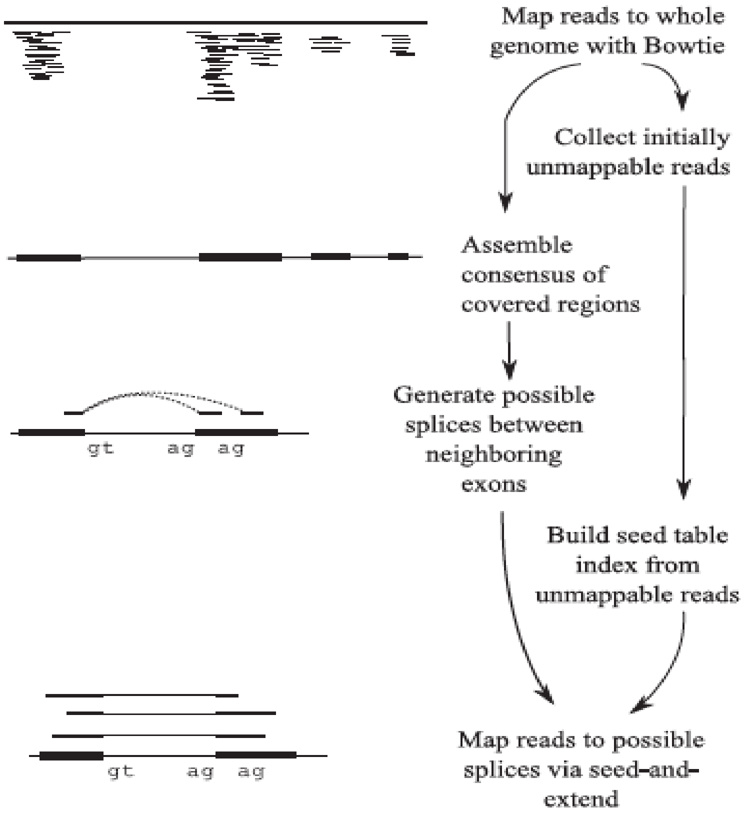
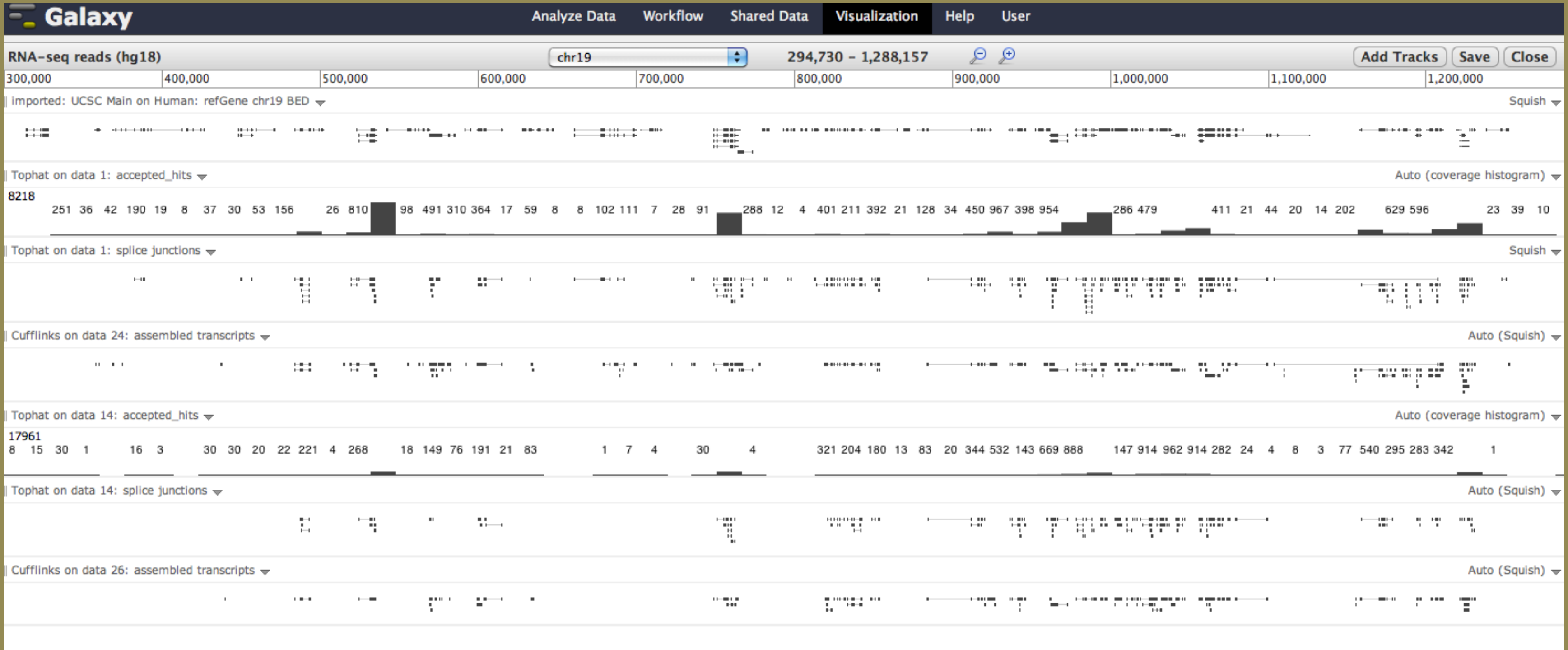

Map reads to whole genome with Bowtie
Collect initially unmappable reads
Assemble consensus of covered regions
Generate possible splices between neighboring exons
gt
ag
ag
Build seed table index from unmappable reads
Map reads to possible splices via seed-and-extend
gt
ag
ag

Fig. 1. The TopHat pipeline. RNA-Seq reads are mapped against the whole reference genome, and those reads that do not map are set aside. An initial consensus of mapped regions is computed by Maq. Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form potential splice junctions. The IUM reads are indexed and aligned to these splice junction sequences.

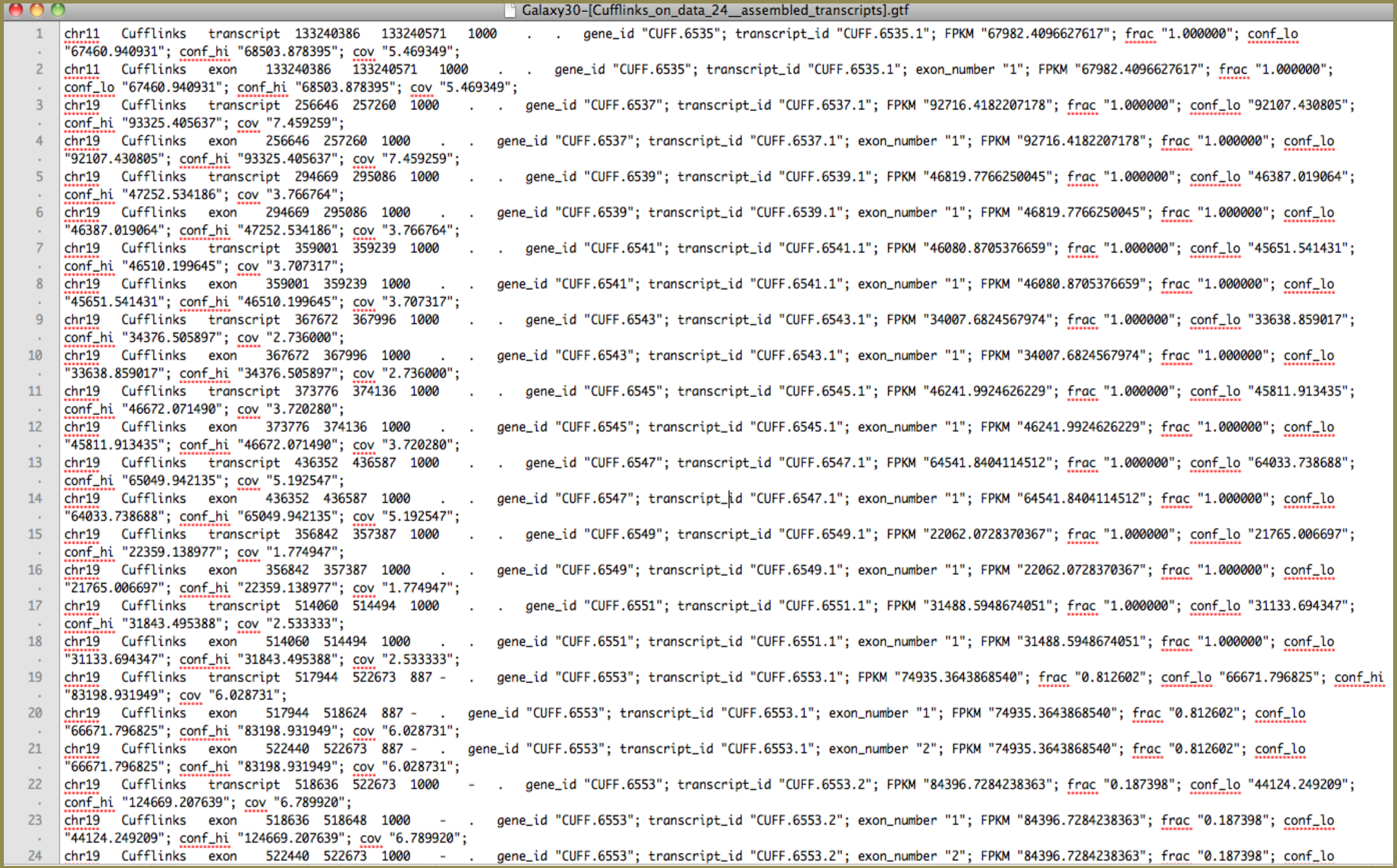
TopHat and Cufflinks Output in Track Browser



Cufflinks

- assembles transcripts (de novo or reference-guided),
- estimates the relative abundances of these transcripts based on how many reads support each one (FPKM - Fragments Per Kilobase of exon per Million fragments mapped estimates), and
- tests for differential expression and regulation in RNA-Seq samples.
- input: aligned RNA-seq reads (SAM/BAM; e.g from TopHat)
- outputs:
  - GTF file containing assembled isoforms
  - Tabular files, with coordinates and expression values
    - Transcripts
    - Genes

Cufflinks GTF assembled transcripts and their relative abundances



Galaxy30-[Cufflinks_on_data_24__assembled_transcripts].gtf

```
chr11   Cufflinks   transcript  133240386   133240571   1000    .   .   gene_id "CUFF.6535"; transcript_id "CUFF.6535.1"; FPKM "67982.4096627617"; frac "1.000000"; conf_lo
"67460.940931"; conf_hi "68503.878395"; cov "5.469349";
chr11   Cufflinks   exon    133240386   133240571   1000    .   .   gene_id "CUFF.6535"; transcript_id "CUFF.6535.1"; exon_number "1"; FPKM "67982.4096627617"; frac "1.000000";
conf_lo "67460.940931"; conf_hi "68503.878395"; cov "5.469349";
chr19   Cufflinks   transcript  256646  257260  1000    .   .   gene_id "CUFF.6537"; transcript_id "CUFF.6537.1"; FPKM "92716.4182207178"; frac "1.000000"; conf_lo "92107.430805";
conf_hi "93325.405637"; cov "7.459259";
chr19   Cufflinks   exon    256646  257260  1000    .   .   gene_id "CUFF.6537"; transcript_id "CUFF.6537.1"; exon_number "1"; FPKM "92716.4182207178"; frac "1.000000"; conf_lo
"92107.430805"; conf_hi "93325.405637"; cov "7.459259";
chr19   Cufflinks   transcript  294669  295086  1000    .   .   gene_id "CUFF.6539"; transcript_id "CUFF.6539.1"; FPKM "46819.7766250045"; frac "1.000000"; conf_lo "46387.019064";
conf_hi "47252.534186"; cov "3.766764";
chr19   Cufflinks   exon    294669  295086  1000    .   .   gene_id "CUFF.6539"; transcript_id "CUFF.6539.1"; exon_number "1"; FPKM "46819.7766250045"; frac "1.000000"; conf_lo
"46387.019064"; conf_hi "47252.534186"; cov "3.766764";
chr19   Cufflinks   transcript  359001  359239  1000    .   .   gene_id "CUFF.6541"; transcript_id "CUFF.6541.1"; FPKM "46080.8705376659"; frac "1.000000"; conf_lo "45651.541431";
conf_hi "46510.199645"; cov "3.707317";
chr19   Cufflinks   exon    359001  359239  1000    .   .   gene_id "CUFF.6541"; transcript_id "CUFF.6541.1"; exon_number "1"; FPKM "46080.8705376659"; frac "1.000000"; conf_lo
"45651.541431"; conf_hi "46510.199645"; cov "3.707317";
chr19   Cufflinks   transcript  367672  367996  1000    .   .   gene_id "CUFF.6543"; transcript_id "CUFF.6543.1"; FPKM "34007.6824567974"; frac "1.000000"; conf_lo "33638.859017";
conf_hi "34376.505897"; cov "2.736000";
chr19   Cufflinks   exon    367672  367996  1000    .   .   gene_id "CUFF.6543"; transcript_id "CUFF.6543.1"; exon_number "1"; FPKM "34007.6824567974"; frac "1.000000"; conf_lo
"33638.859017"; conf_hi "34376.505897"; cov "2.736000";
chr19   Cufflinks   transcript  373776  374136  1000    .   .   gene_id "CUFF.6545"; transcript_id "CUFF.6545.1"; FPKM "46241.9924626229"; frac "1.000000"; conf_lo "45811.913435";
conf_hi "46672.071490"; cov "3.720280";
chr19   Cufflinks   exon    373776  374136  1000    .   .   gene_id "CUFF.6545"; transcript_id "CUFF.6545.1"; exon_number "1"; FPKM "46241.9924626229"; frac "1.000000"; conf_lo
"45811.913435"; conf_hi "46672.071490"; cov "3.720280";
chr19   Cufflinks   transcript  436352  436587  1000    .   .   gene_id "CUFF.6547"; transcript_id "CUFF.6547.1"; FPKM "64541.8404114512"; frac "1.000000"; conf_lo "64033.738688";
conf_hi "65049.942135"; cov "5.192547";
chr19   Cufflinks   exon    436352  436587  1000    .   .   gene_id "CUFF.6547"; transcript_id "CUFF.6547.1"; exon_number "1"; FPKM "64541.8404114512"; frac "1.000000"; conf_lo
"64033.738688"; conf_hi "65049.942135"; cov "5.192547";
chr19   Cufflinks   transcript  356842  357387  1000    .   .   gene_id "CUFF.6549"; transcript_id "CUFF.6549.1"; FPKM "22062.0728370367"; frac "1.000000"; conf_lo "21765.006697";
conf_hi "22359.138977"; cov "1.774947";
chr19   Cufflinks   exon    356842  357387  1000    .   .   gene_id "CUFF.6549"; transcript_id "CUFF.6549.1"; exon_number "1"; FPKM "22062.0728370367"; frac "1.000000"; conf_lo
"21765.006697"; conf_hi "22359.138977"; cov "1.774947";
chr19   Cufflinks   transcript  514060  514494  1000    .   .   gene_id "CUFF.6551"; transcript_id "CUFF.6551.1"; FPKM "31488.5948674051"; frac "1.000000"; conf_lo "31133.694347";
conf_hi "31843.495388"; cov "2.533333";
chr19   Cufflinks   exon    514060  514494  1000    .   .   gene_id "CUFF.6551"; transcript_id "CUFF.6551.1"; exon_number "1"; FPKM "31488.5948674051"; frac "1.000000"; conf_lo
"31133.694347"; conf_hi "31843.495388"; cov "2.533333";
chr19   Cufflinks   transcript  517944  522673  887 -   .   gene_id "CUFF.6553"; transcript_id "CUFF.6553.1"; FPKM "74935.3643868540"; frac "0.812602"; conf_lo "66671.796825"; conf_hi
"83198.931949"; cov "6.028731";
chr19   Cufflinks   exon    517944  518624  887 -   .   gene_id "CUFF.6553"; transcript_id "CUFF.6553.1"; exon_number "1"; FPKM "74935.3643868540"; frac "0.812602"; conf_lo
"66671.796825"; conf_hi "83198.931949"; cov "6.028731";
chr19   Cufflinks   exon    522440  522673  887 -   .   gene_id "CUFF.6553"; transcript_id "CUFF.6553.1"; exon_number "2"; FPKM "74935.3643868540"; frac "0.812602"; conf_lo
"66671.796825"; conf_hi "83198.931949"; cov "6.028731";
chr19   Cufflinks   transcript  518636  522673  1000    -   .   gene_id "CUFF.6553"; transcript_id "CUFF.6553.2"; FPKM "84396.7284238363"; frac "0.187398"; conf_lo "44124.249209";
conf_hi "124669.207639"; cov "6.789920";
chr19   Cufflinks   exon    518636  518648  1000    -   .   gene_id "CUFF.6553"; transcript_id "CUFF.6553.2"; exon_number "1"; FPKM "84396.7284238363"; frac "0.187398"; conf_lo
"44124.249209"; conf_hi "124669.207639"; cov "6.789920";
chr19   Cufflinks   exon    522440  522673  1000    -   .   gene_id "CUFF.6553"; transcript_id "CUFF.6553.2"; exon_number "2"; FPKM "84396.7284238363"; frac "0.187398"; conf_lo
```

Cuffcompare

- compares assembled transcripts to a reference annotation
- tracks Cufflinks transcripts across multiple experiments (e.g. across a time course)
- input: Cufflinks' GTF assembled transcripts (>2 samples), and optionally a reference annotation
- outputs:
    - transcripts accuracy file - TXT
        - accuracy of the transcripts in each sample when compared to the reference annotation data.
    - transcripts combined file - GTF
        - union of all transcribed fragments in each sample.
    - transcripts tracking file – tabular (tmap)
        - matches transcripts up between samples. Each row contains a transcript structure that is present in one or more input GTF files.

Cuffdiff

- 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.
- used to find differentially expressed genes and transcripts, as well as genes that are being differentially regulated at the transcriptional and post-transcriptional level.

Cuffdiff

- input: Cufflinks or Cuffcompare assembled or combined GTF files along with two SAM/BAM files containing the fragment alignments for two or more sample (only 2 in Galaxy).

Cuffdiff

- Transcript FPKM expression tracking.
- Gene FPKM expression tracking; tracks the summed FPKM of transcripts sharing each gene_id.
- Primary transcript FPKM tracking; tracks the summed FPKM of transcripts sharing each tss_id.
- Coding sequence FPKM tracking; tracks the summed FPKM of transcripts sharing each p_id, independent of tss_id.

Cuffdiff

- Transcript differential FPKM.
- Gene differential FPKM. Tests differences in the summed FPKM of transcripts sharing each gene_id.
- Primary transcript differential FPKM. Tests differences in the summed FPKM of transcripts sharing each tss_id.
- Coding sequence differential FPKM. Tests differences in the summed FPKM of transcripts sharing each p_id independent of tss_id.

Cuffdiff

- Differential splicing tests: this tabular file lists, for each primary transcript, the amount of overloading detected among its isoforms, i.e. how much differential splicing exists between isoforms processed from a single primary transcript. Only primary transcripts from which two or more isoforms are spliced are listed in this file.
- Differential promoter tests: this tabular file lists, for each gene, the amount of overloading detected among its primary transcripts, i.e. how much differential promoter use exists between samples. Only genes producing two or more distinct primary transcripts (i.e. multi-promoter genes) are listed here.
- Differential CDS tests: this tabular file lists, for each gene, the amount of overloading detected among its coding sequences, i.e. how much differential CDS output exists between samples. Only genes producing two or more distinct CDS (i.e. multi-protein genes) are listed here.

References

1. https://bitbucket.org/galaxy/galaxy-central/wiki/
2. Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010 April; 38(6): 1767–1771
3. Daniel Blankenberg, Assaf Gordon, Gregory Von Kuster, Nathan Coraor , James Taylor, Anton Nekrutenko and the Galaxy Team. Manipulation of FASTQ data with Galaxy. BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 14 2010, pages 1783–1785 doi:10.1093/bioinformatics/btq281
4. B.T. Wilhelm, J.-R. Landry. RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48 (2009) 249–257
5. Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105-1111 (2009)
6. Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nature Biotechnology doi:10.1038/nbt.1621
7. http://cufflinks.cbcb.umd.edu/manual.html
8. http://cufflinks.cbcb.umd.edu/howitworks.html