10/06/2014

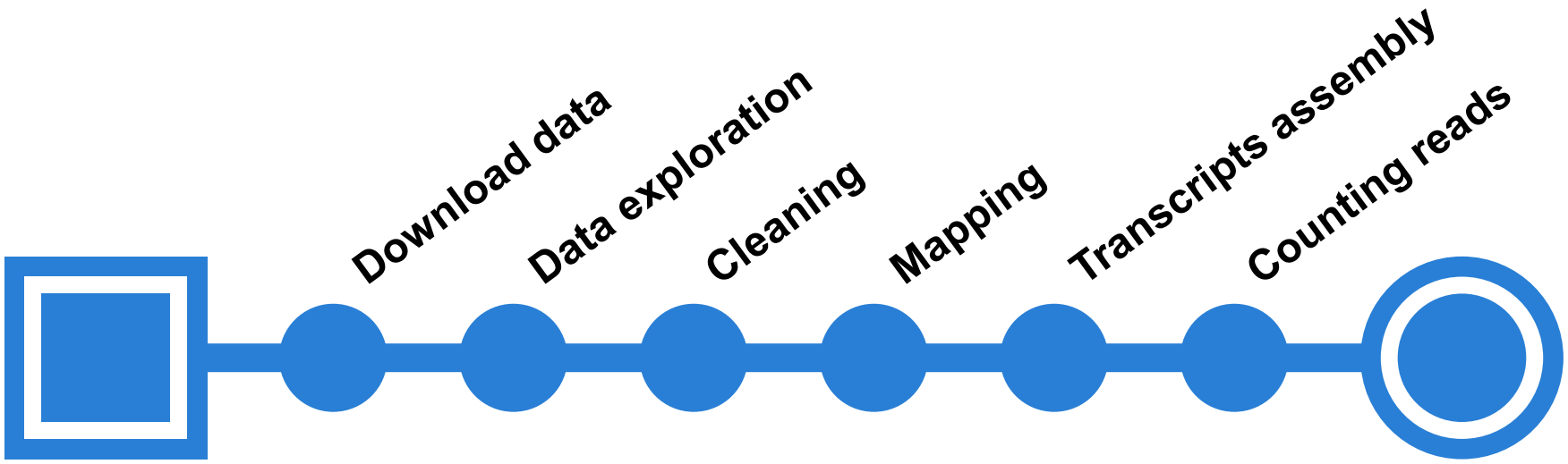
RNA-Seq analysis

With reference assembly

Cormier Alexandre, PhD student
UMR8227, Algal Genetics Group

INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Summary


Download data
Data exploration
Cleaning
Mapping
Transcripts assembly
Counting reads

Typical RNA-seq workflow

- Reference genome
- Reference transcriptome

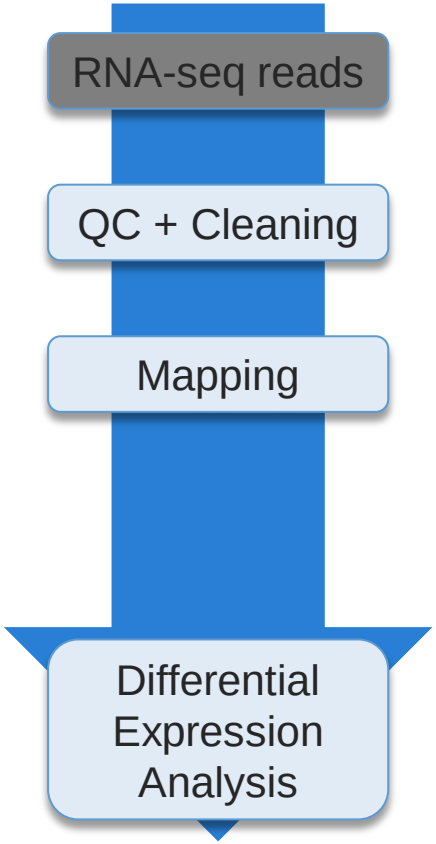
- Reference genome
- No reference transcriptome

Typical RNA-seq workflow

- Reference genome
- Reference transcriptome

- Reference genome
- No reference transcriptome

Non discovery mode

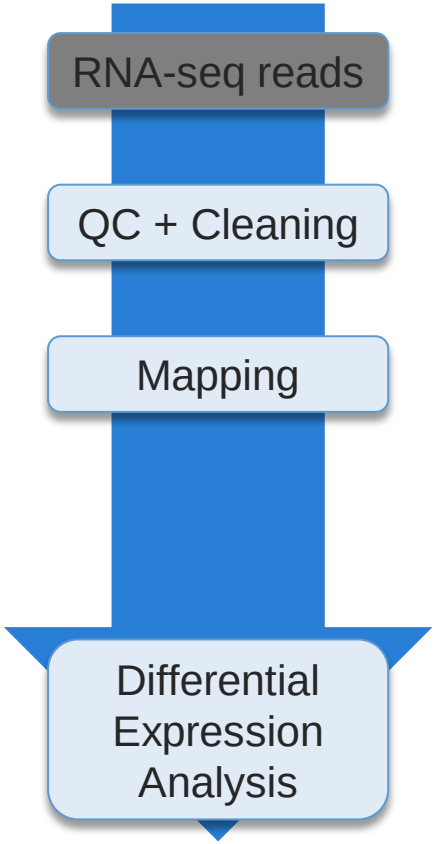


Typical RNA-seq workflow

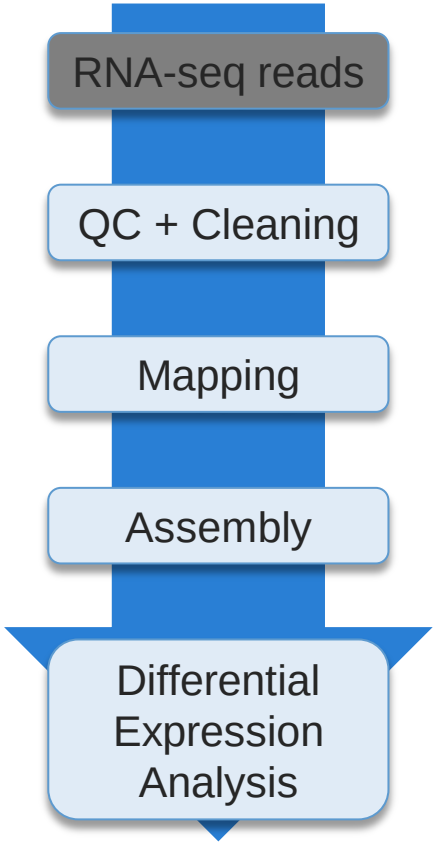
- Reference genome
- Reference transcriptome

- Reference genome
- No reference transcriptome

Non discovery mode



Discovery mode



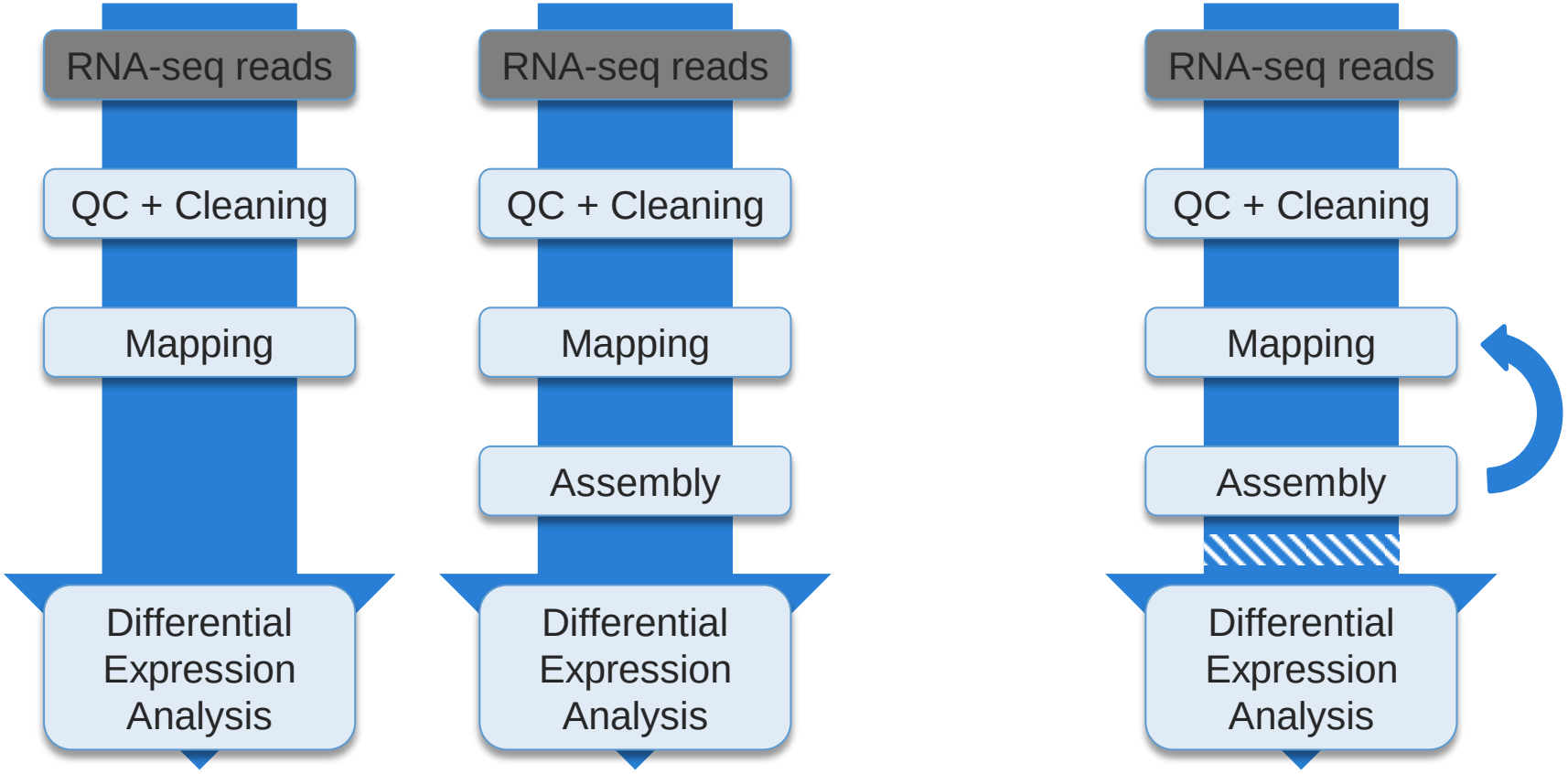
Typical RNA-seq workflow

- Reference genome
- Reference transcriptome

- Reference genome
- No reference transcriptome

Non discovery mode

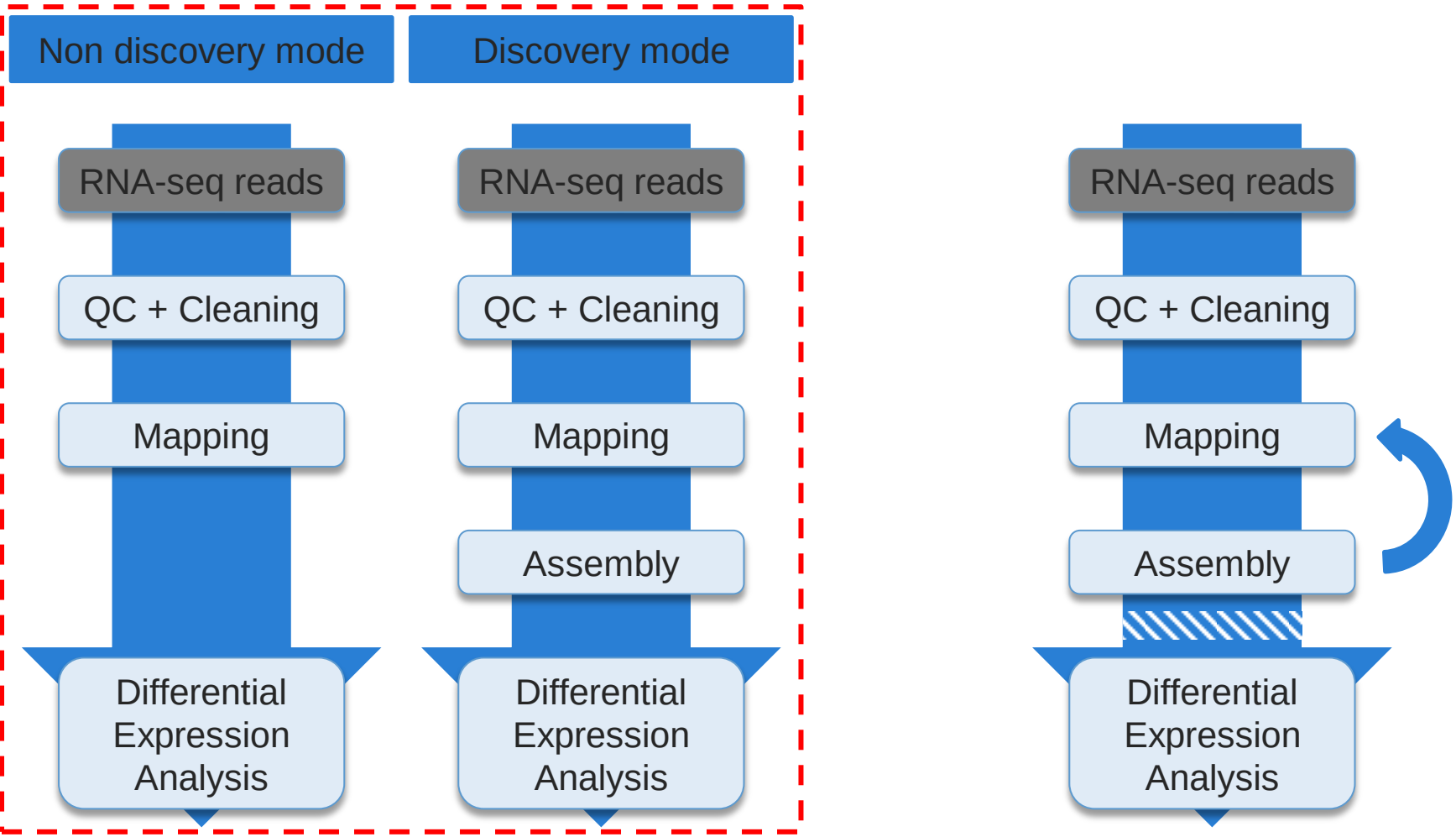
Discovery mode



Typical RNA-seq workflow

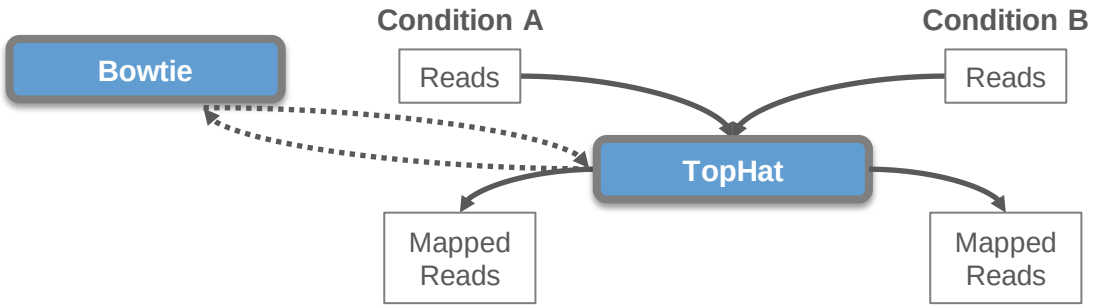
- Reference genome
- Reference transcriptome

- Reference genome
- No reference transcriptome

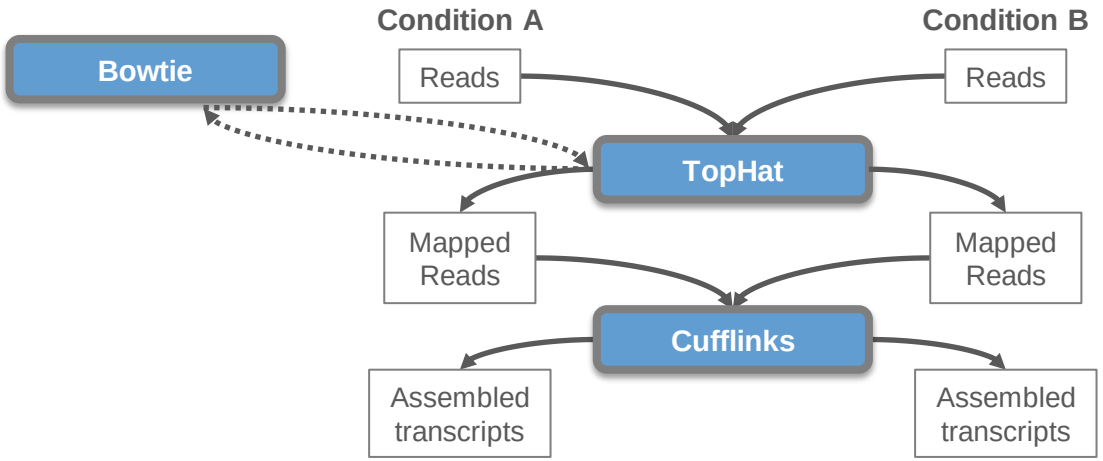


Tuxedo Workflow v2.2.0

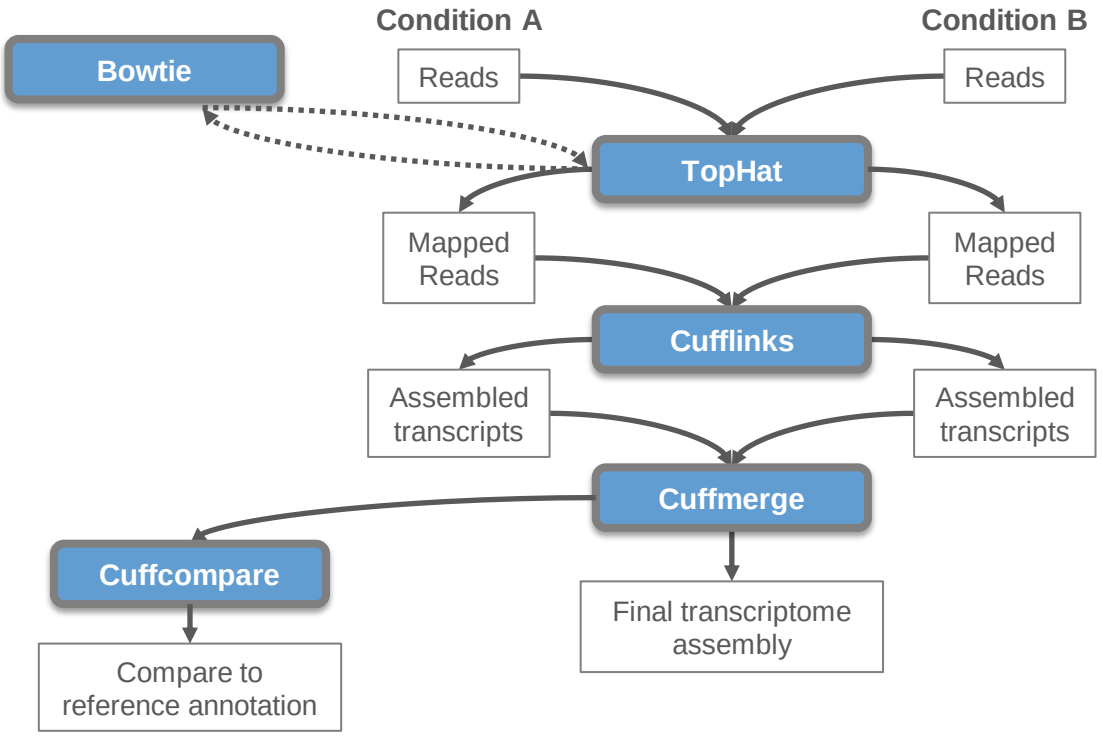
Tuxedo Workflow v2.2.0



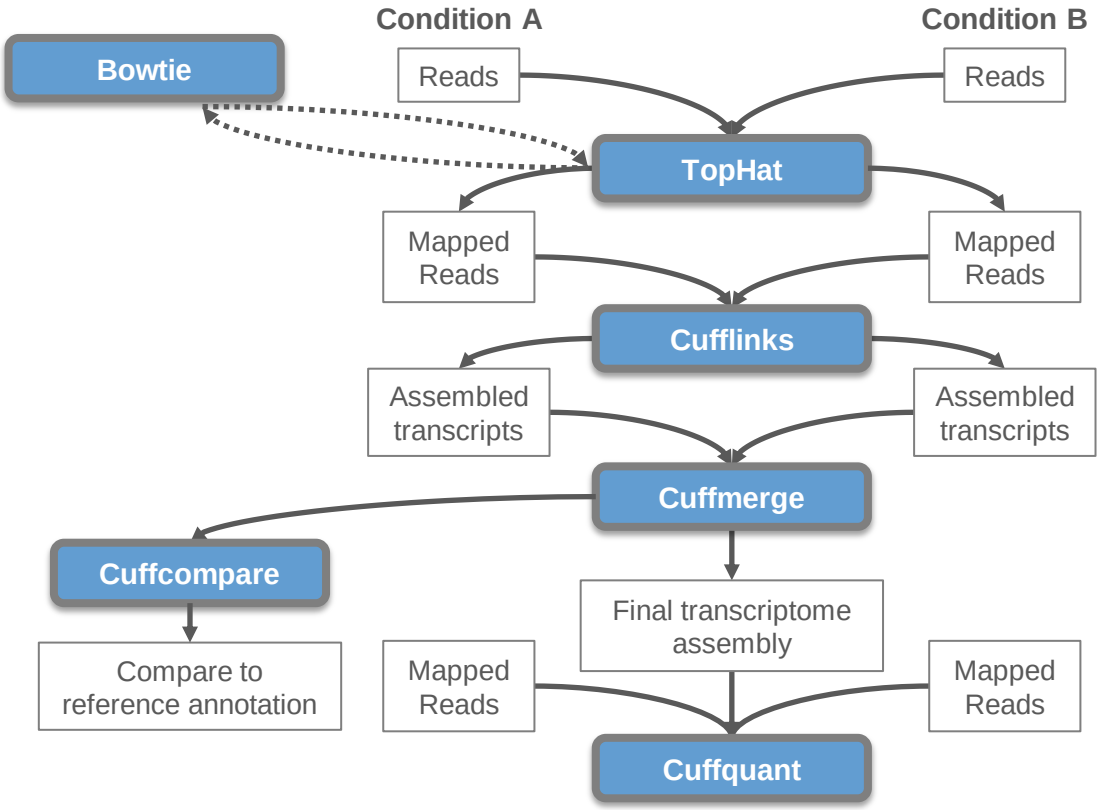
Tuxedo Workflow v2.2.0



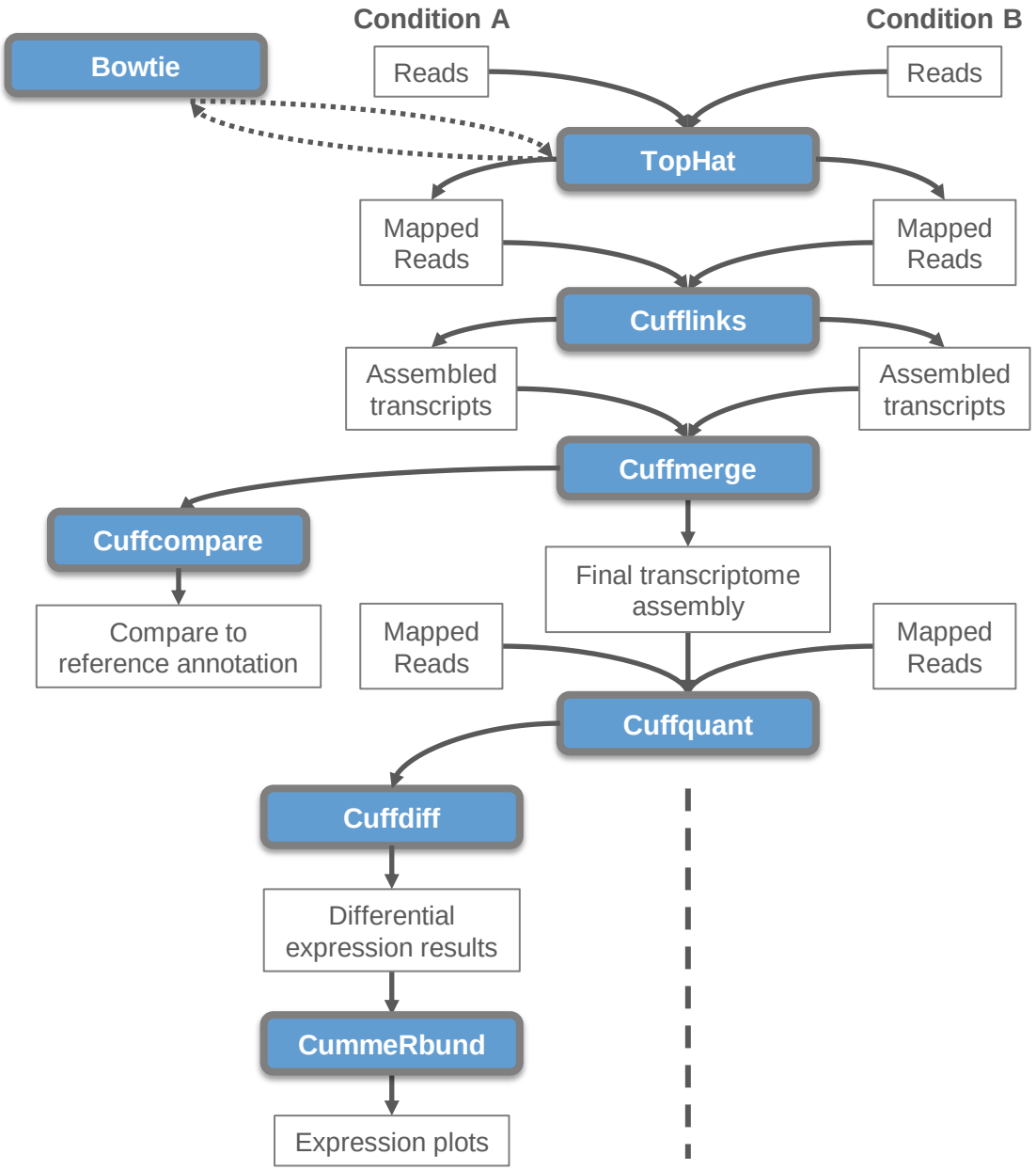
Tuxedo Workflow v2.2.0



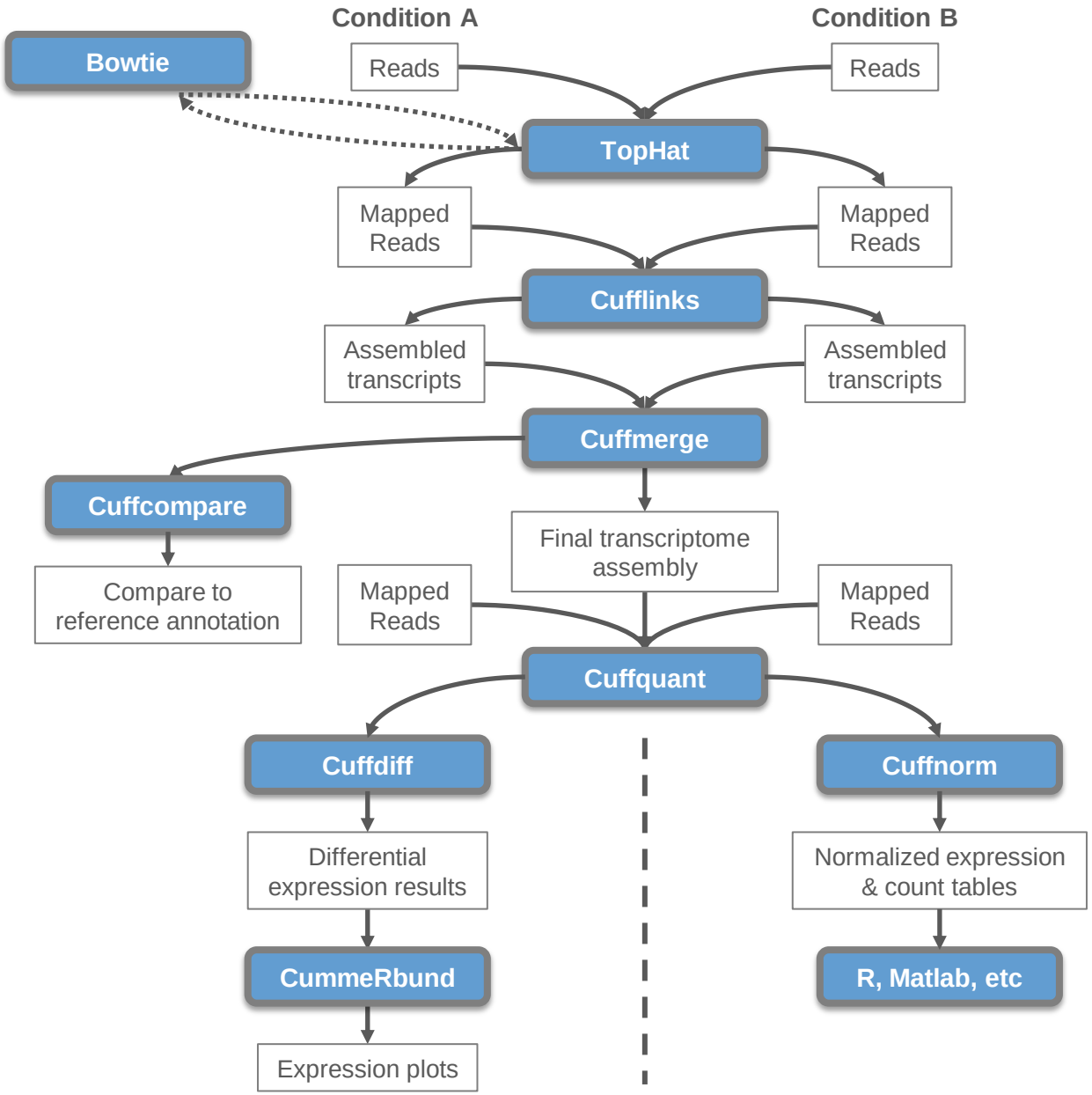
Tuxedo Workflow v2.2.0



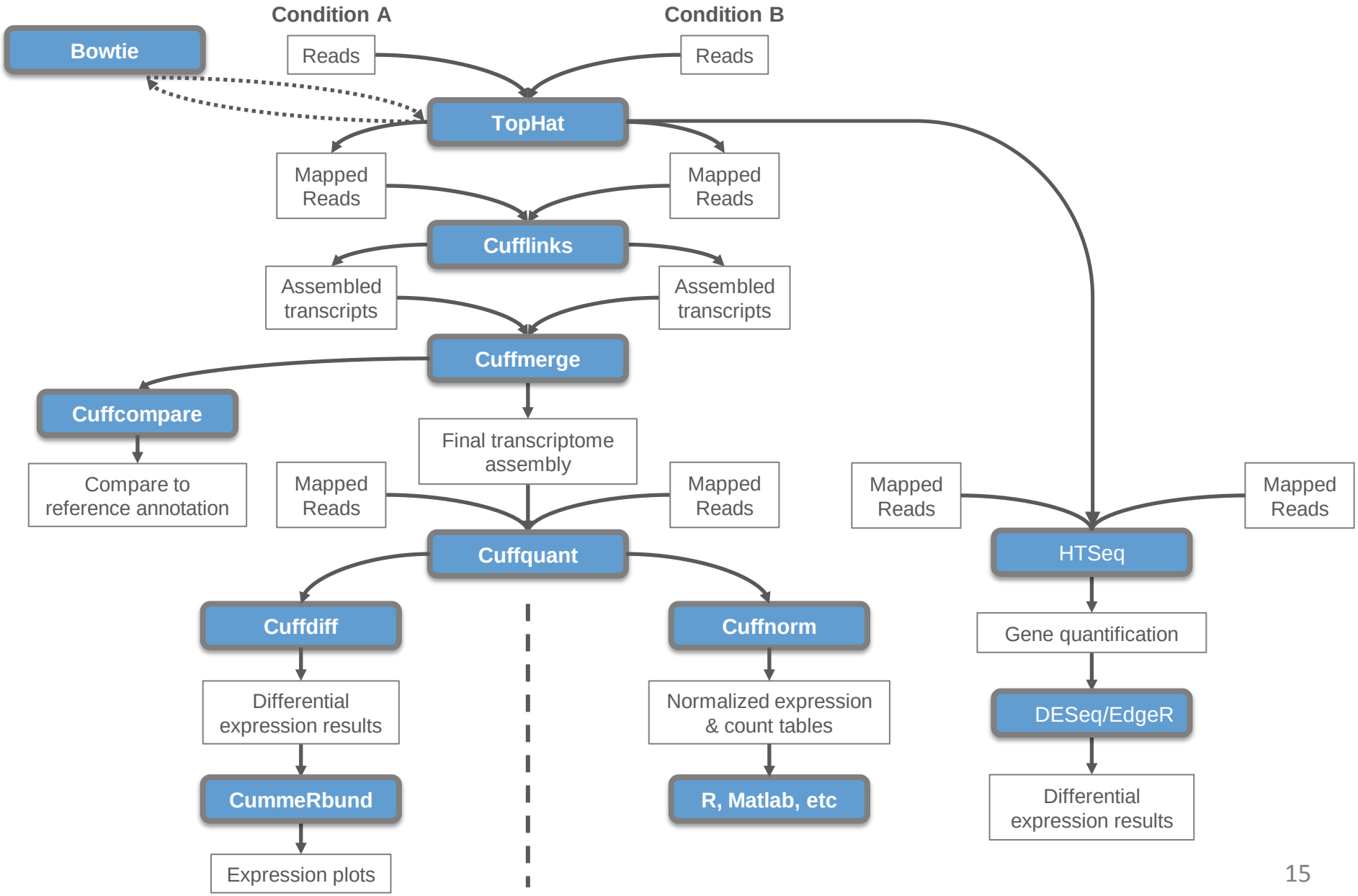
Tuxedo Workflow v2.2.0



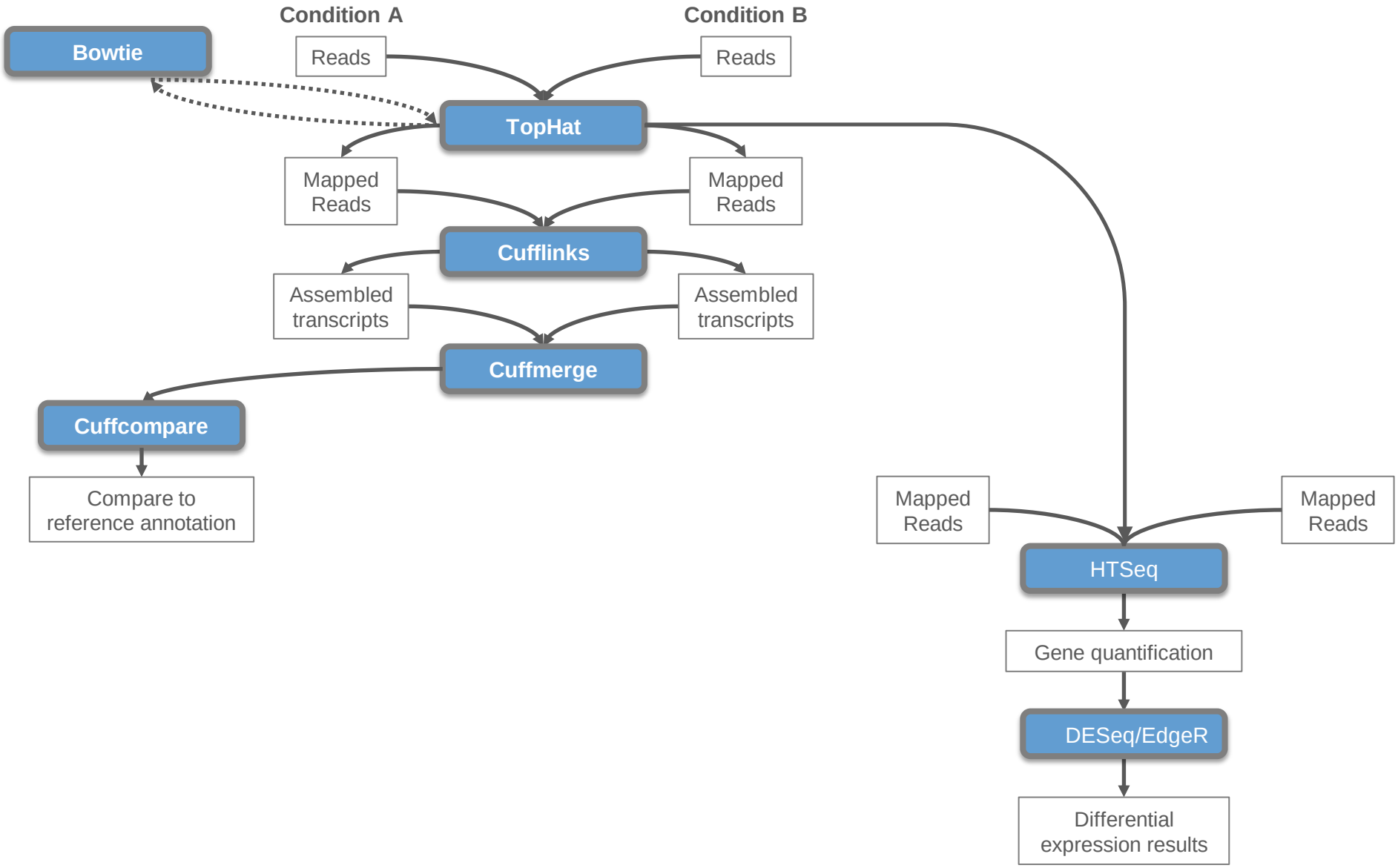
Tuxedo Workflow v2.2.0



Tuxedo Workflow v2.2.0



Tuxedo Workflow v2.2.0



Other solutions

Table 1 | Selected list of RNA-seq analysis programs

Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy[39]	Probabilistic model		
	Burrows-Wheeler transform methods	Bowtie[43]			
		BWA[44]	Incorporates quality scores		
Spliced aligners	Exon-first methods	MapSplice[52]	Works with multiple unspliced aligners	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
		SpliceMap[50]			
		TopHat[51]	Uses Bowtie alignments		
	Seed-extend methods	GSNAP[53]	Can use SNP databases		
		QPALMA[54]	Smith-Waterman for large gaps		
Transcriptome reconstruction					
Genome-guided reconstruction	Exon identification	G.Mor.Se	Assembles exons	Identifying novel transcripts using a known reference genome	Alignments to reference genome
	Genome-guided assembly	Scripture[28]	Reports all isoforms		
		Cufflinks[29]	Reports a minimal set of isoforms		
Genome-independent reconstruction	Genome-independent assembly	Velvet[61]	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads
		TransABySS[56]			

Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth* **8,** 469–477 (2011).

Data presentation

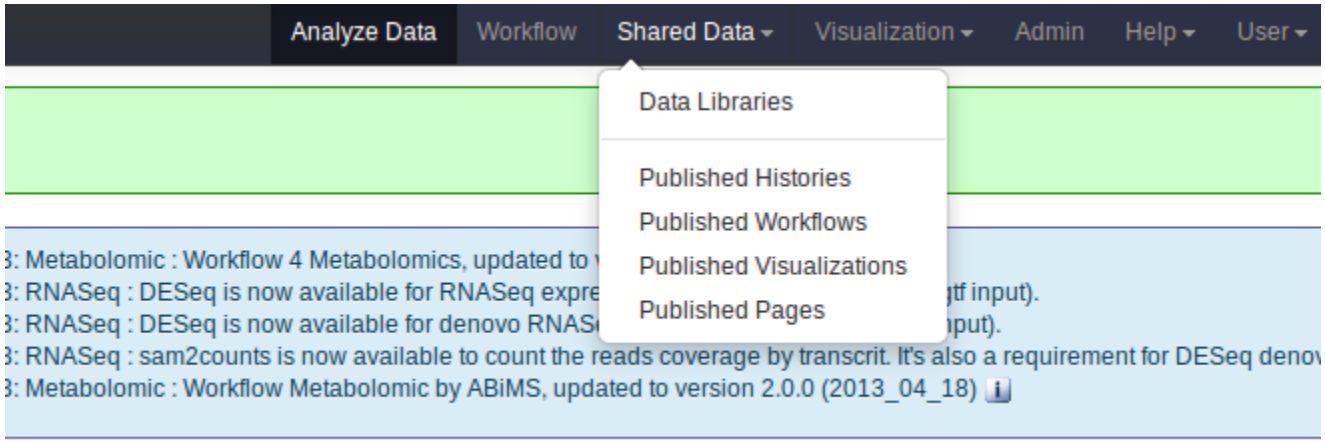
Data retrived from the ENCODE project

- 2 human cell lines :
  - Gm12878 (lymphoblastoid cell line) → 2 replicates
  - Hct116 (colorectal carcinoma cell line) → 2 replicates
- Illumina paired-end 2x75bp, insert size ~400bp
- Working only on the chromosome 22

Objective :

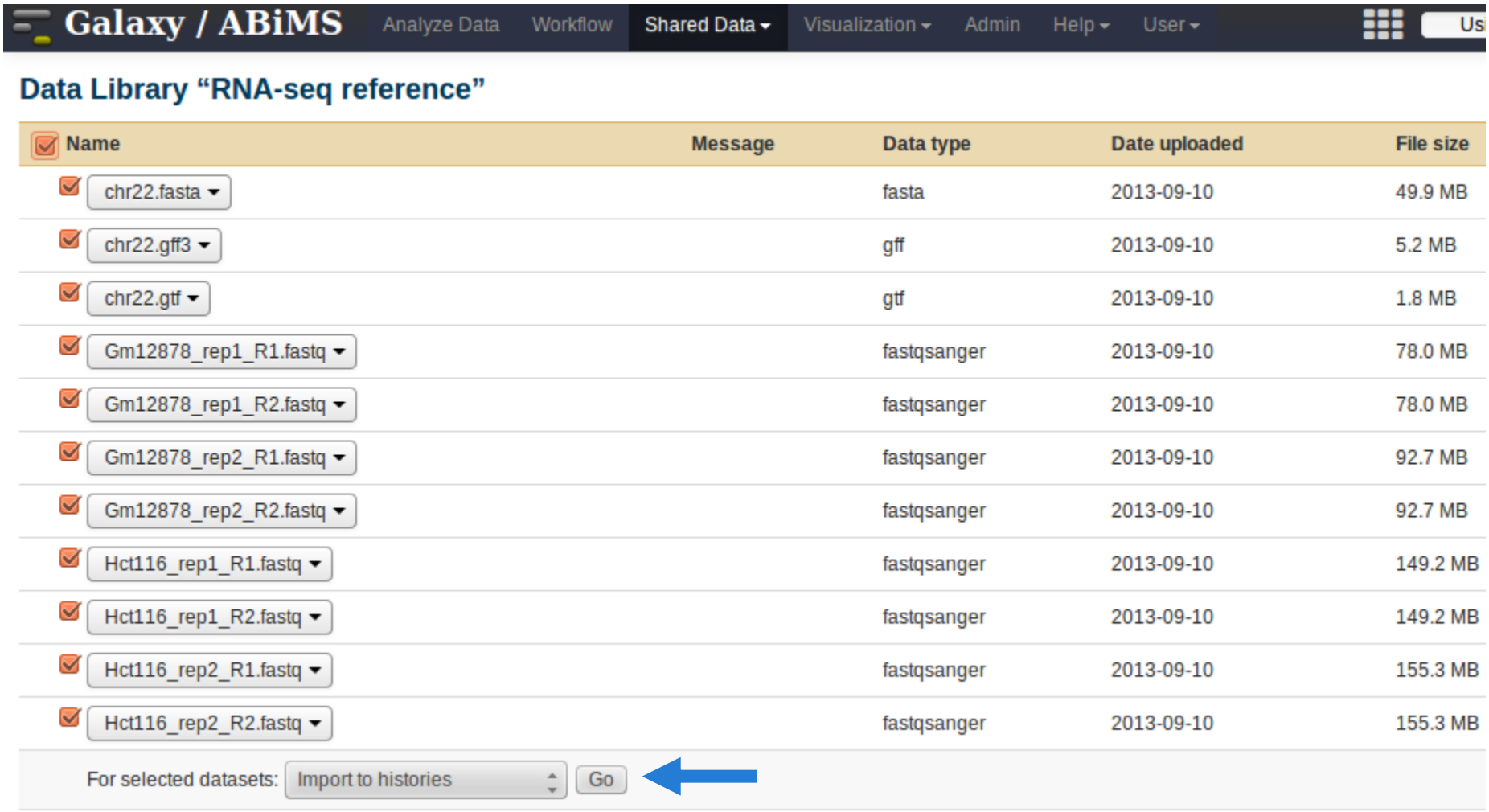
➢ **Identify differentially expressed genes in 2 human cell lines**

Get data



Data library name ↓	Data library description
RNA-seq de-novo	Dataset for RNA-seq de-novo, re-ingeneered - ppericard
RNA-seq de-novo - Assembly	Pre-cleaned and pre-processed reads for assembly, re-ingeneered - ppericard
RNA-seq de-novo - Cheat Sheet	Output datasets from time consuming tools
RNA-seq de-novo - Differential Expression	Cleaned sequences and filtered de-novo assembly for differential expression, re-ingeneered - ppericard
RNA-seq reference - Assembly	Assembled data - acormier
RNA-seq reference - Input	Dataset for RNA-seq with reference genome - acormier
RNA-seq reference - Mapping	Mapped data - acormier

Get data

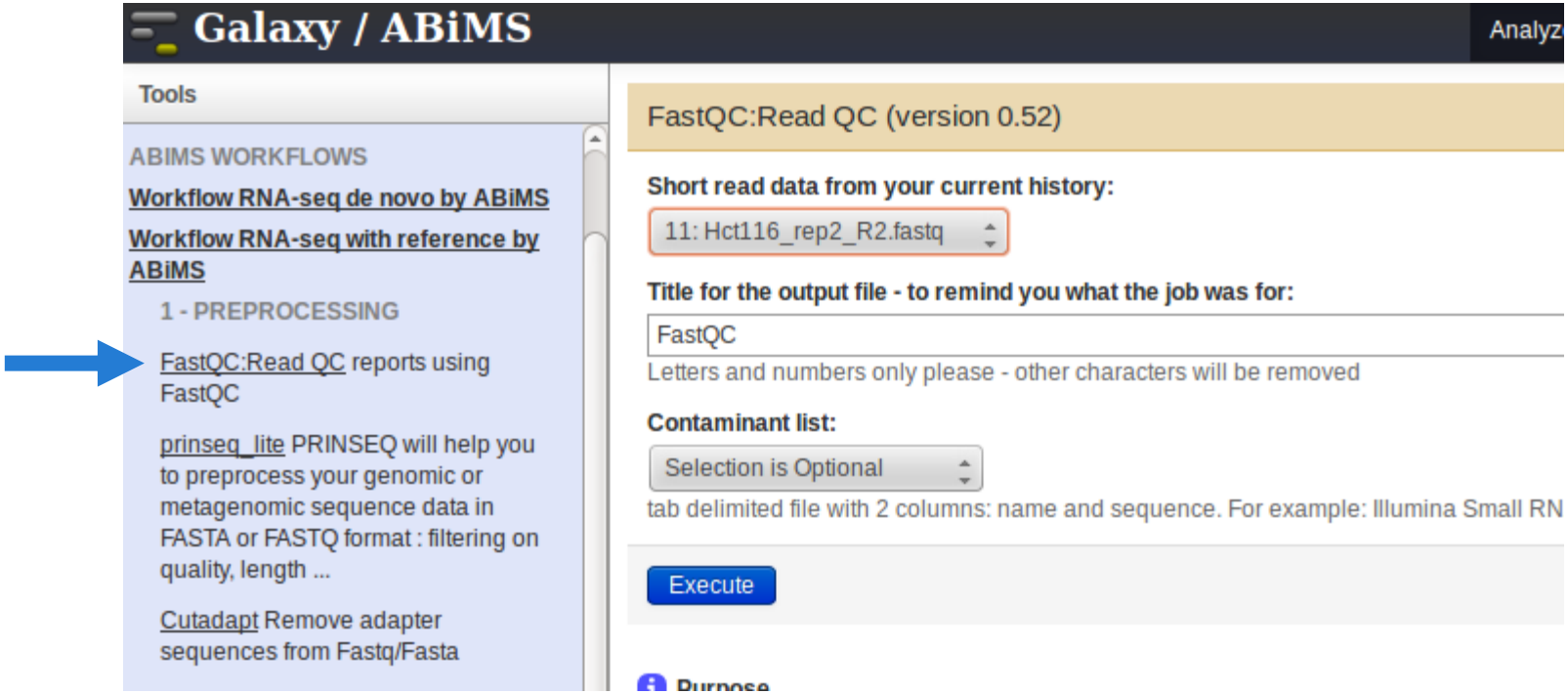


Export all data in a new history and choose a name (ex: rna-seq reference analysis)

Data exploration: FastQC

Obtain some statistics and information of a fastq file

Check the quality of the data contained in fastq file



Data exploration: FastQC

Launch FastQC analysis only on :

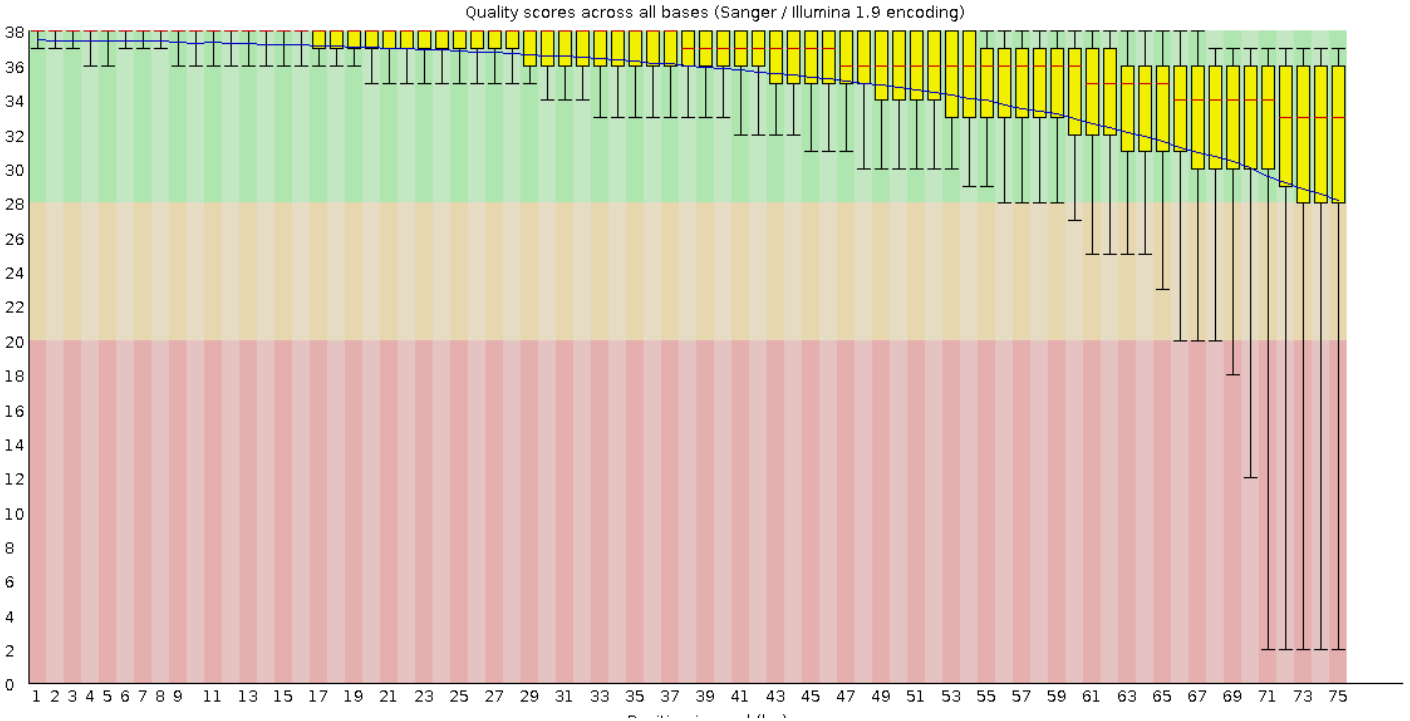
- Gm12878_rep1_R1.fastq
- Hct116_rep1_R1.fastq

Data exploration: FastQC

Basic Statistics

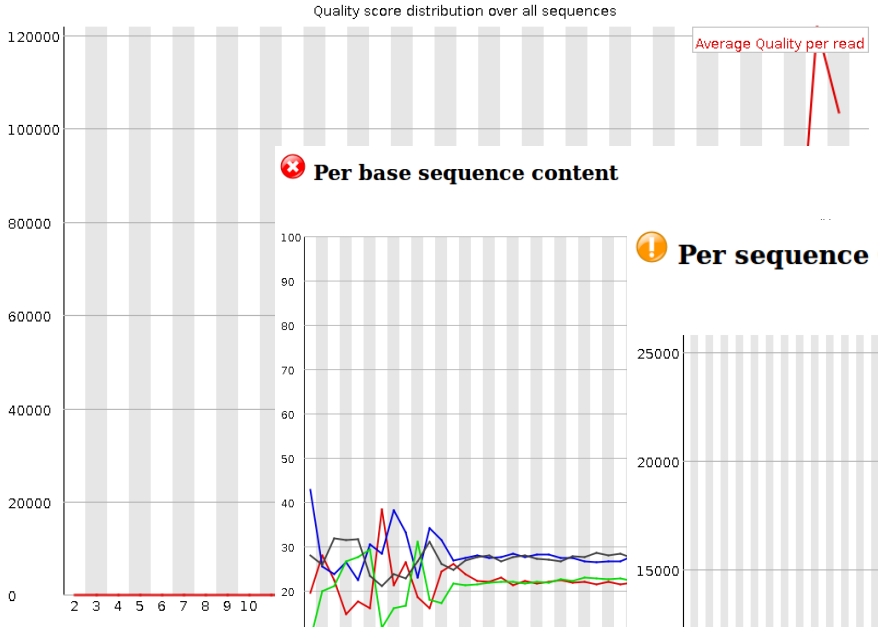
Measure	Value
Filename	Gm12878_rep1_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	406512
Filtered Sequences	0
Sequence length	75
%GC	55

Per base sequence quality



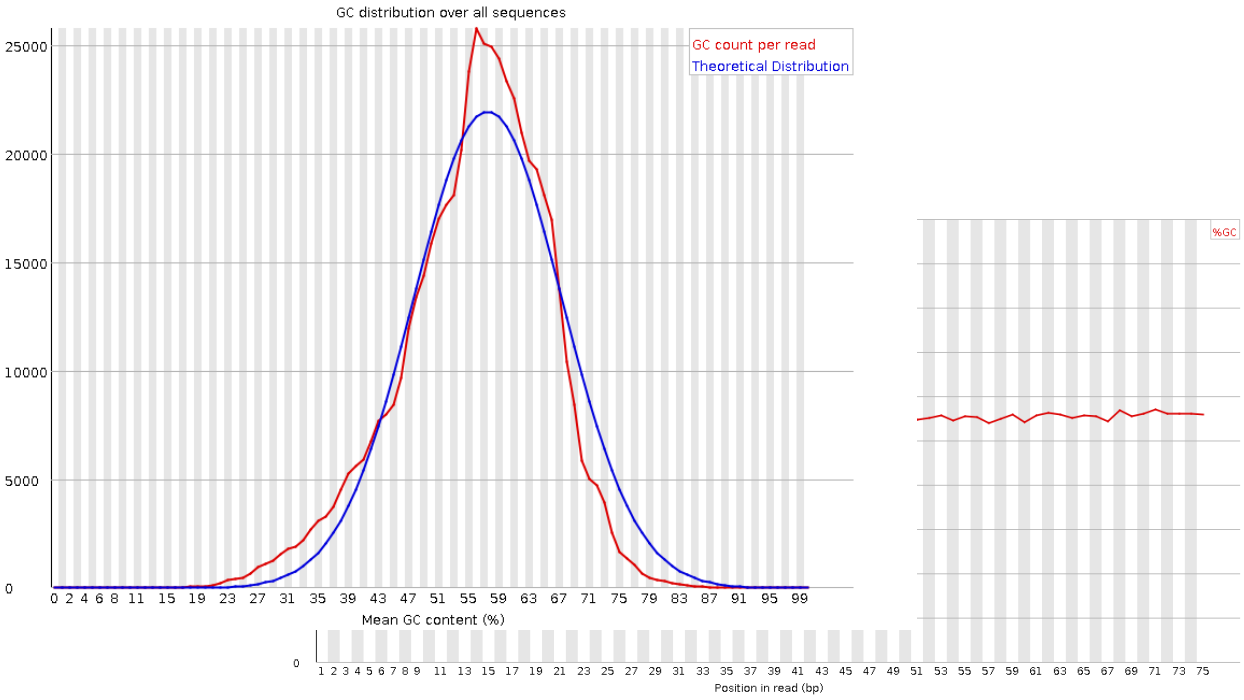
Data exploration: FastQC

Per sequence quality scores



Per base sequence content

Per sequence GC content



Data exploration: FastQC

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CAGCGCTCTCGGGACGTCTCCACCATGGCCTGGGCTCTGCTGCTCCTCACCCTCCTCACTCAGGACACAGGGTCC	1821	0.4479572558743653	No Hit
CGGGACGTCTCCACCATGGCCTGGGCTCTGCTGCTCCTCACCCTCCTCACTCAGGACACAGGGTCCTGGGCCCAG	1054	0.25927893887511316	No Hit
GGGGGCTTTGCCTGGGTGGTGTTGGTACCAGGAGACAAGGTTATAACTCCCAACATTACTGCTGGTTCCAGTGCA	973	0.2393533278230409	No Hit
GCGCTCTCGGGACGTCTCCACCATGGCCTGGGCTCTGCTGCTCCTCACCCTCCTCACTCAGGACACAGGGTCCTG	920	0.22631558231983312	No Hit
CAGAGGTCCAAGTAAACCGCTAGCTTGTTGCACCGTGGAGGCCACAGGAGCAGAAACATGGAATGCCAGACGCTG	914	0.2248396111307907	No Hit
CTTTGCCTGGGTGGTGTTGGTACCAGGAGACAAGGTTATAACTCCCAACATTACTGCTGGTTCCAGTGCAGGAGA	806	0.1982721297280277	No Hit
CACAGGTCCAGGGCAGAGGACCAACATGGGCATTTTGTTTATGAGCAAGGTGGGTCTCAGAGGTGATCGGCGATC	510	0.12545755106860315	No Hit
TCGGGACGTCTCCACCATGGCCTGGGCTCTGCTGCTCCTCACCCTCCTCACTCAGGACACAGGGTCCTGGGCCCA	494	0.12152162789782342	No Hit
CTCGGGACGTCTCCACCATGGCCTGGGCTCTGCTGCTCCTCACCCTCCTCACTCAGGACACAGGGTCCTGGGCCC	488	0.12004565670878105	No Hit
TTTGCCTGGGTGGTGTTGGTACCAGGAGACAAGGTTATAACTCCCAACATTACTGCTGGTTCCAGTGCAGGAGAT	467	0.11487975754713269	No Hit
CCAGAATGTCACAGGTCCAGGGCAGAGGACCAACATGGGCATTTTGTTTATGAGCAAGGTGGGTCTCAGAGGTGA	461	0.11340378635809029	No Hit
CCTGGGTGGTGTTGGTACCAGGAGACAAGGTTATAACTCCCAACATTACTGCTGGTTCCAGTGCAGGAGATGGTG	454	0.11168181997087416	No Hit
TTACACTCGGCCACAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGACCAACATGGGCAT	446	0.10971385838548431	No Hit
GCTTGTTGCACCGTGGAGGCCACAGGAGCAGAAACATGGAATGCCAGACGCTGGGGATGCTGGTACAAGTTGTGG	446	0.10971385838548431	No Hit
CTTGGCAGTACTTCTTCATGCTGCTGAAGTCCTTCTCCAGCTGCTTCTTGCCATCCTCATCCTGCCATTTCTTGC	443	0.10897587279096312	No Hit
CTCGGCCACAAATAAACACAGAAATAGTCCAGAATGTCACAGGTCCAGGGCAGAGGACCAACATGGGCATTTTGT	419	0.10307198803479356	No Hit

Cleaning with PRINSEQ

Cleaning: PRINSEQ

With a reference genome, the cleaning step is not necessary.

The use of genome allows filtering reads with a poor quality and contamination.

Can be problematic with Illumina reads → diminution of the quality at the end of the sequence

Cleaning: PRINSEQ

High quality

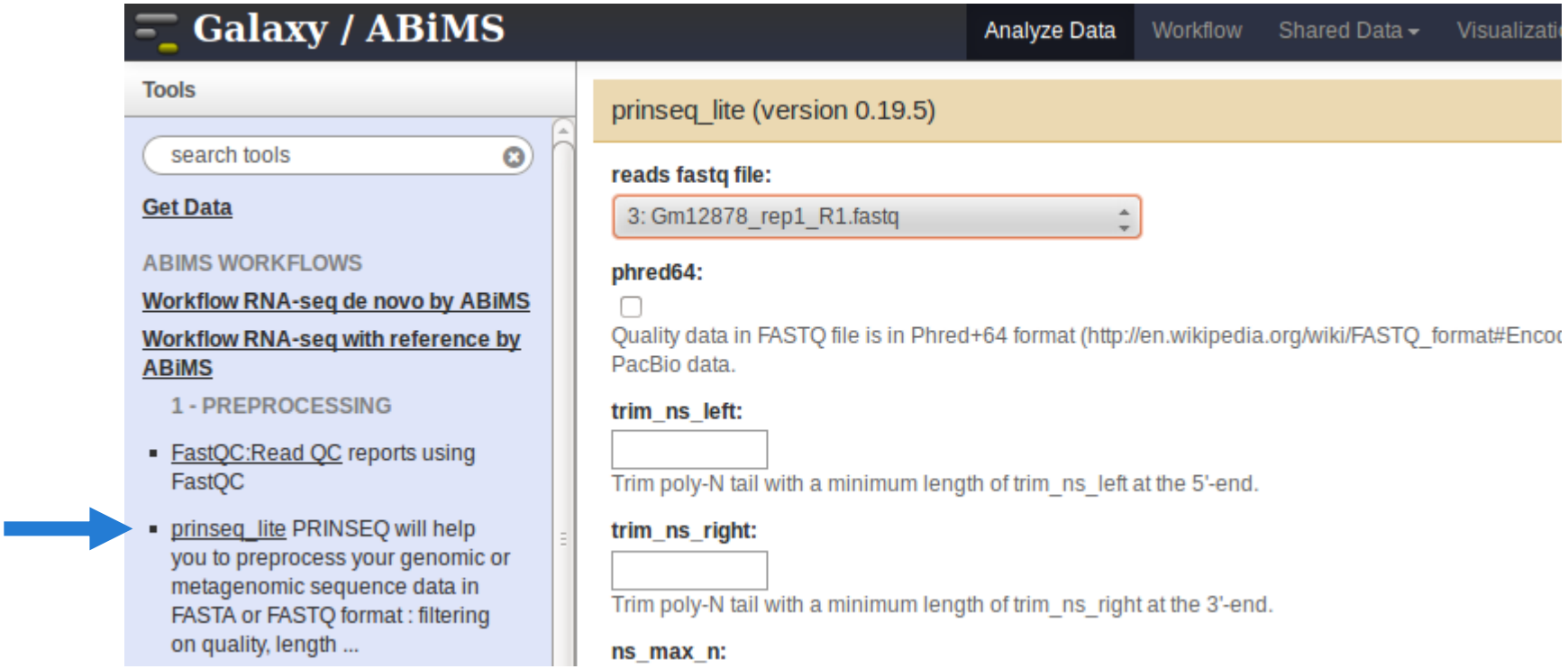
Low quality

Raw read

Mapped ?

Cleaning: PRINSEQ

Cleaning: PRINSEQ



PRINSEQ | Parameters

prinseq_lite (version 0.19.5)

reads fastq file:
3: Gm12878_rep1_R1.fastq

phred64:
☐
Quality data in FASTQ file is in Phred+64 format (http://en.wikipedia.org/v
PacBio data.

trim_ns_left:
1
Trim poly-N tail with a minimum length of trim_ns_left at the 5'-end.

trim_ns_right:
1
Trim poly-N tail with a minimum length of trim_ns_right at the 3'-end.

ns_max_n:
0
Filter sequence with more than ns_max_n Ns.

trim_qual_right:
25
Trim sequence by quality score from the 3'-end with this threshold score.

min_qual_mean:
20
Filter sequence with quality score mean below min_qual_mean.

min_len:
30
Filter sequence shorter than min_len.

noniupac:
☑
Filter sequence with characters other than A, C, G, T or N.

trim_tail_left:

Trim poly-A/T tail with a minimum length of trim_tail_left at the 5'-end.

trim_tail_right:

Trim poly-A/T tail with a minimum length of trim_tail_right at the 3'-end.

lc_method:
none
Method to filter low complexity sequences. The current options are dust an

trim_to_len:

Trim all sequence from the 3'-end to result in sequence with this length.

Launch PRINSEQ on all fastq files

Cleaning: PRINSEQ

Launch FastQC analysis only on :

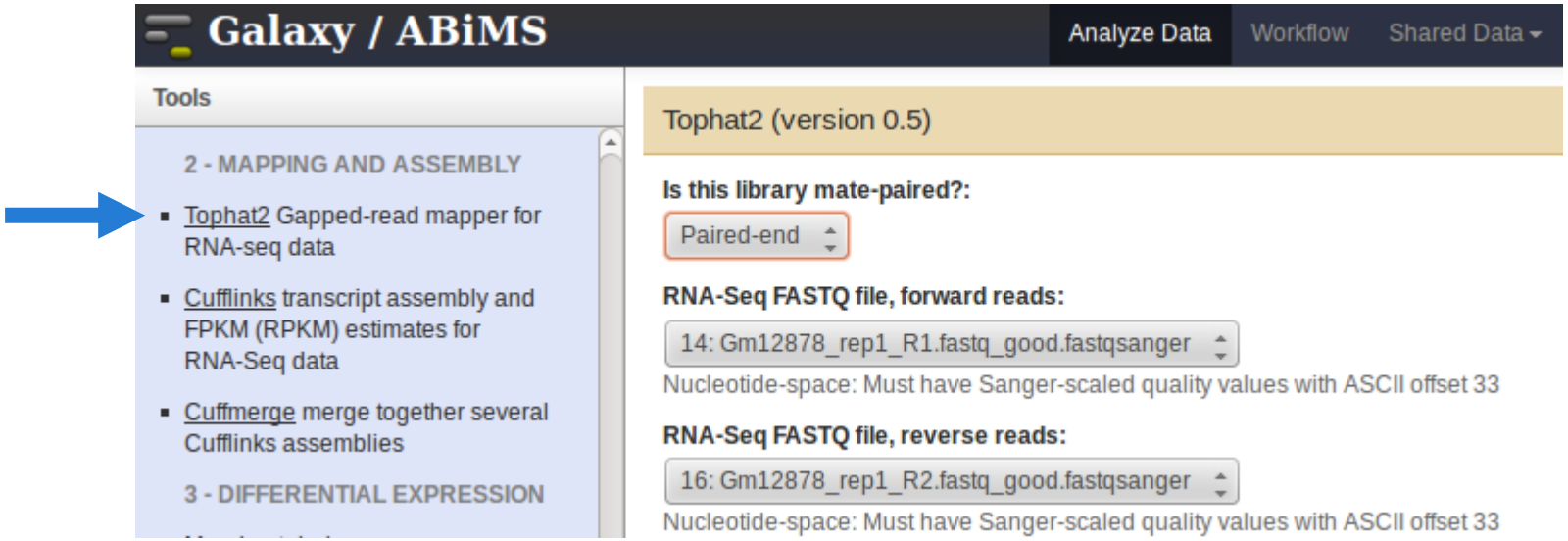
- Gm12878_rep1_R1.fastq_good.fastqsanger
- Hct116_rep1_R1.fastq_good.fastqsanger

Compare results with raw reads

Mapping with TopHat 2

Mapping: TopHat 2

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



Overview

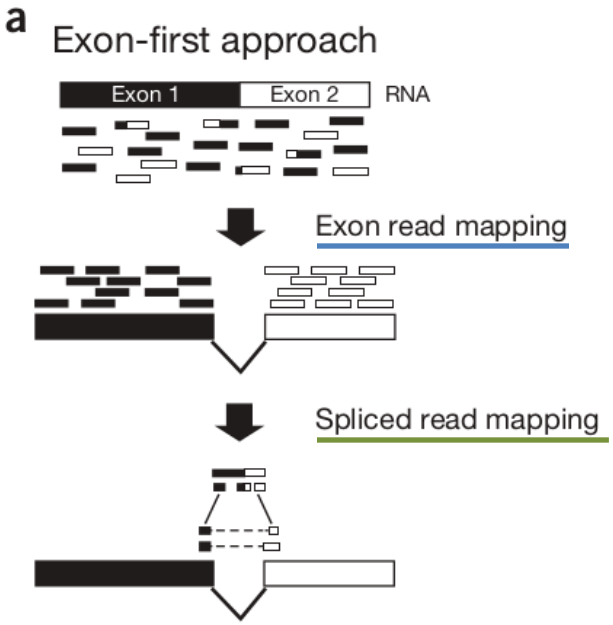
Table 1 | Selected list of RNA-seq analysis programs

Class	Category	Package
Read mapping		
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]
		Stampy[39]
	Burrows-Wheeler transform methods	Bowtie[43]
		BWA[44]
Spliced aligners	Exon-first methods	MapSplice[52]
		SpliceMap[50]
		TopHat[51]
	Seed-extend methods	GSNAP[53]
		QPALMA[54]

Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth* **8,** 469–477 (2011).

Table 1 | Selected list of RNA-seq analysis programs

Class	Category	Package
Read mapping		
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]
		Stampy[39]
	Burrows-Wheeler transform methods	Bowtie[43]
		BWA[44]
Spliced aligners	Exon-first methods	MapSplice[52]
		SpliceMap[50]
		TopHat[51]
	Seed-extend methods	GSNAP[53]
		QPALMA[54]

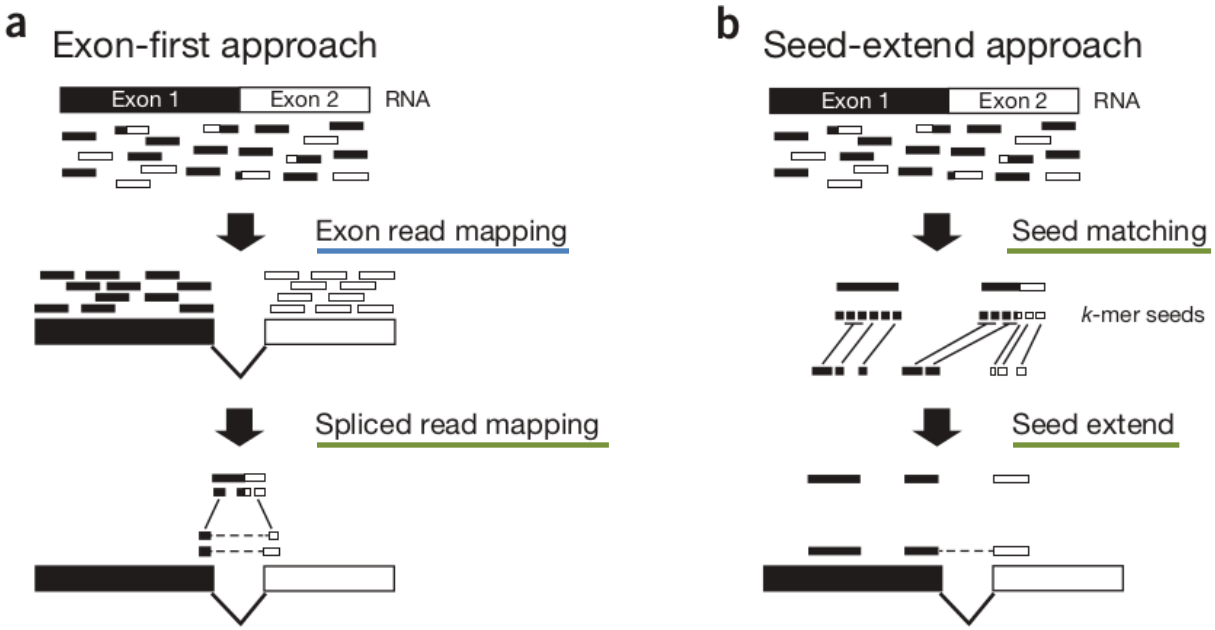


Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth* **8,** 469–477 (2011).

Overview

Table 1 | Selected list of RNA-seq analysis programs

Class	Category	Package
Read mapping		
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]
		Stampy[39]
	Burrows-Wheeler transform methods	Bowtie[43]
		BWA[44]
Spliced aligners	Exon-first methods	MapSplice[52]
		SpliceMap[50]
		TopHat[51]
	Seed-extend methods	GSNAP[53]
		QPALMA[54]

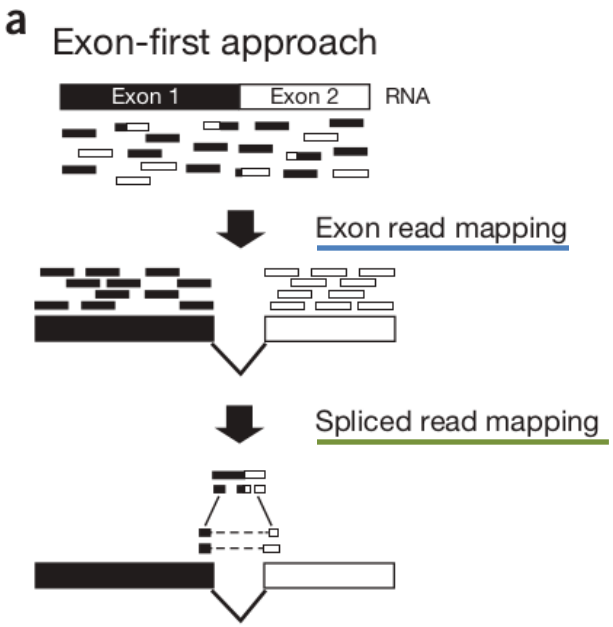


Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth* **8,** 469–477 (2011).

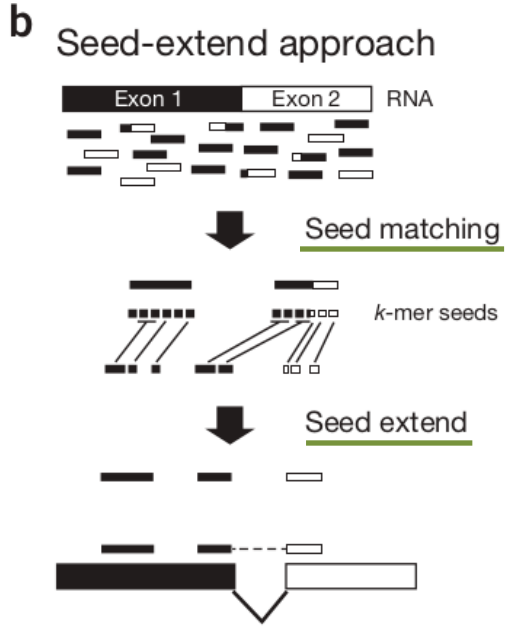
Overview

Table 1 | Selected list of RNA-seq analysis programs

Class	Category	Package
Read mapping		
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]
		Stampy[39]
	Burrows-Wheeler transform methods	Bowtie[43]
		BWA[44]
Spliced aligners	Exon-first methods	MapSplice[52]
		SpliceMap[50]
		TopHat[51]
	Seed-extend methods	GSNAP[53]
		QPALMA[54]



Faster (~x8) and less greedy



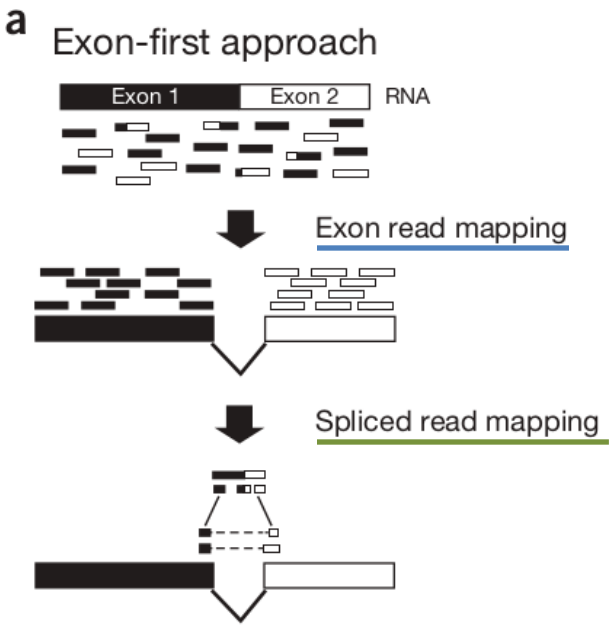
Better for polymorphic species
A little bit more exhaustive

Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth* **8,** 469–477 (2011).

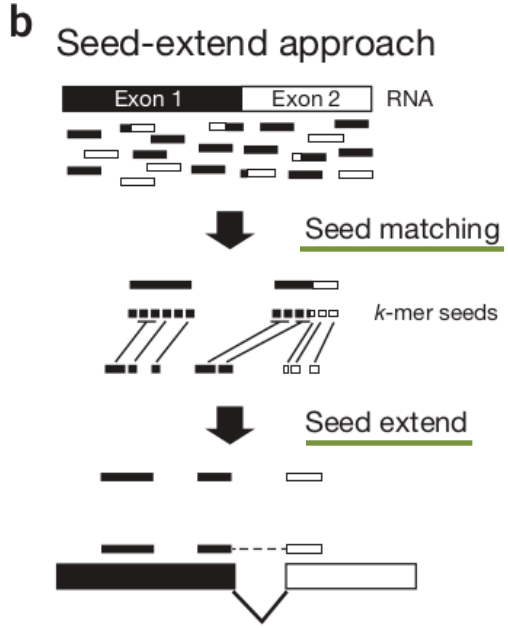
Overview

Table 1 | Selected list of RNA-seq analysis programs

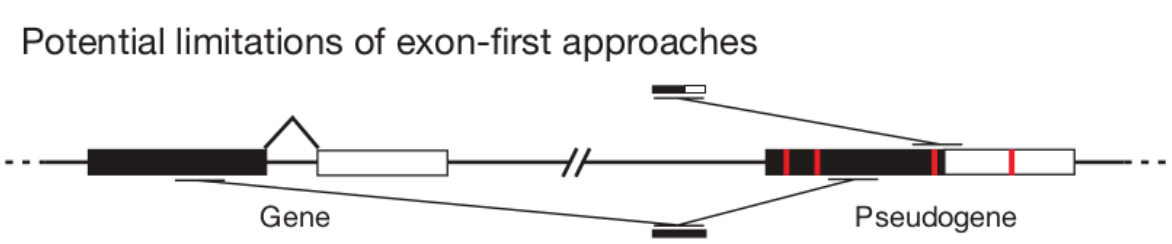
Class	Category	Package
Read mapping		
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]
		Stampy[39]
	Burrows-Wheeler transform methods	Bowtie[43]
		BWA[44]
Spliced aligners	Exon-first methods	MapSplice[52]
		SpliceMap[50]
		TopHat[51]
	Seed-extend methods	GSNAP[53]
		QPALMA[54]



Faster (~x8) and less greedy



Better for polymorphic species
A little bit more exhaustive



Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth* **8,** 469–477 (2011).

TopHat2 | Input

- Fastq file(s)
- Genome

One mapping per replicate

TopHat2 | Parameters

Tophat2 (version 0.5)

Is this library mate-paired?:

Paired-end

RNA-Seq FASTQ file, forward reads:

14: Gm12878_rep1_R1.fastq_good.fastqsanger

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

RNA-Seq FASTQ file, reverse reads:

15: Gm12878_rep1_R2.fastq_good.fastqsanger

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs:

200

Std. Dev for Distance between Mate Pairs:

20

The standard deviation for the distribution on inner distances between mate pairs.

Report discordant pair alignments?:

No

Use a built in reference genome or own from your history:

Use a genome from history

Built-ins genomes were created using default options

Select the reference genome:

1: chr22.fasta

TopHat settings to use:

Full parameter list

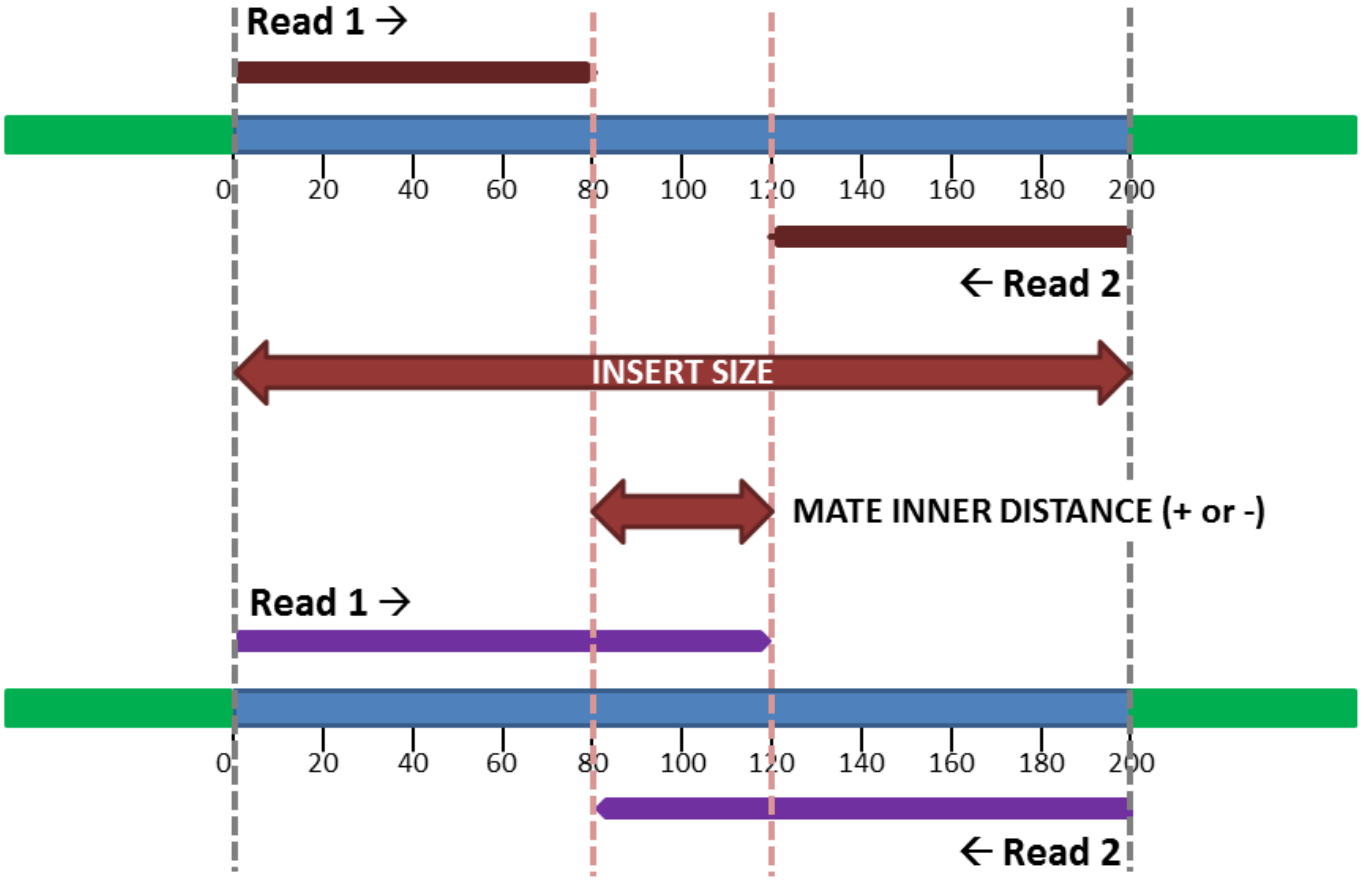
You can use the default settings or set custom values for any of Tophat's paramete

Maximum number of alignments to be allowed:

1

1: By default: 20

Insert size


Read 1 →
0
20
40
60
80
100
120
140
160
180
200
← Read 2
INSERT SIZE
MATE INNER DISTANCE (+ or -)
Read 1 →
0
20
40
60
80
100
120
140
160
180
200
← Read 2

Default parameters

TopHat 2 is optimized for:

- Human
- Mouse

If you work on these species, you can use default parameters

Else, you need to input all of the specie specifics parameters, such as intron size.

The minimum intron length:

70

TopHat will ignore donor/acceptor pairs closer than this many bases apart.

The maximum intron length:

500000

When searching for junctions ab initio, TopHat will ignore donor/acceptor pairs segment alignment of a long read.

Minimum intron length that may be found during split-segment (default) search:

50

Maximum intron length that may be found during split-segment (default) search:

500000

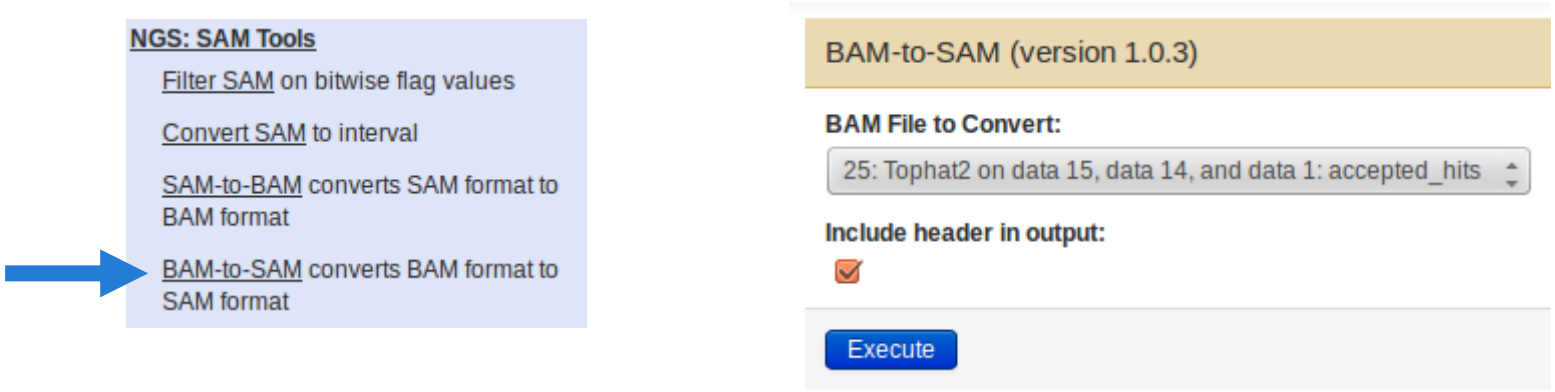
Multiple mapping reads

- Some reads will align to more than one place in the reference, because:
  - Shared exons (if reference is transcriptome)
  - Common domains, gene families
  - Paralogs, pseudogenes, etc.
- This can distort counts, leading to misleading expression levels
- If a read can't be uniquely mapped, how should it be counted or ignored?
- Should it be randomly assigned to one location among all the locations to which it aligns equally well?
- This may depend on the question you're asking...
- ...also depends on the software you use…
- …and also depends of your data (read length, quality, etc)

TopHat2 | Output

- BAM: compressed binary version of the SAM

BAM to SAM



TopHat2 | Output

acormier@n0:/projet/abims/ectocarpus/finalresult/mapping_mRNA ✖ acormier@n0:/scratch/umr8227/ga ✖

SAM file

```
@SQ     SN:sctg_985     LN:8100
@SQ     SN:sctg_986     LN:7184
@SQ     SN:sctg_987     LN:9975
@SQ     SN:sctg_988     LN:7227
@SQ     SN:sctg_99      LN:541954
@SQ     SN:sctg_990     LN:7105
@SQ     SN:sctg_991     LN:7074
@SQ     SN:sctg_993     LN:13104
@SQ     SN:sctg_994     LN:6914
@SQ     SN:sctg_995     LN:6895
@SQ     SN:sctg_996     LN:10769
@SQ     SN:sctg_997     LN:6840
@SQ     SN:sctg_998     LN:6818
@SQ     SN:sctg_999     LN:6999
@PG     ID:TopHat       VN:2.0.9        CL:/usr/local/genome2/tophat-2.0.9/bin/tophat -p 10 -I 26000 -g 40 -o /projet/umr7139/genetique/acormier/tmp/assembly/mature/transcriptome_multihit_40/tophat/gpo1 -G /projet/umr7139/genetique/acormier/finalresult/annotations/all_24112012_FeV.gff3 /projet/abims/ectocarpus/finalresults/index/genome_hybride/male_female_genome_13032013 /projet/abims/ectocarpus/sequence_data_archive/input/sequence_data_clean/RNA-seq_data/Ec32_matGAfm_RNA-seq/110215_SN132_B_s_3_seq_GPO-1.fastq
HWI-ST132_0435:4:64:7500:126200#CTTGTT  256     sctg_285        482     1       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACGCCCACTAG
        gggggggggggggggegggggggggggggggggggggggggggegegeddecgedbde_ebeddZebe_cee__ea`edabe_PZYQK[ZZY^[``[_]    AS:i:-12        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:95C3T0     YT:Z:UU NH:i:3  CC:Z:sctg_297       CP:i:41043      HI:i:1
HWI-ST365:344:D272UACXX:5:1213:16716:24795      256     sctg_285        482     3       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACG
CCCCAGAT        BC@DDFFFFHHHHHJJGHGIJJJJJIJEIJJJJIJJJJJJJJIJJJJIJHHHAEFFFFFDEEEEDEDDDDDDDDDBDDDDDDDDDDDCCDDDDDDDDD      AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:96C0T2     YT:Z:UU NH:i:2  HI:i:1
HWI-ST365:344:D272UACXX:5:2111:11374:47914      256     sctg_285        482     3       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACG
CCCCAGAT        <@:BDDDDDHADFGHBAEDHHCEIGGHIGHEGGBHIIFHHEEHI<DGHH5AEHFDFDCE@C>@@@A@CCA>?B@88<BBCBB?ACCAA:@>@BCBCCCBC      AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:96C0T2     YT:Z:UU NH:i:2  HI:i:1
HWI-ST365:344:D272UACXX:5:2302:18156:7792       256     sctg_285        482     3       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACG
CCCCAGAT        @C@DFFFFHHHHHJJFHIJJJJJJJJHIJJJJJJIGIJJJJJJJJIIJJHHH?DFFFFFCCEEDEECDDDDDDBB?BDDDDDDDCDDD@CCDDDDD@BBC      AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:96C0T2     YT:Z:UU NH:i:2  HI:i:1
HWI-ST365:344:D272UACXX:4:1109:1955:88530       256     sctg_285        482     3       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACG
CCCCAGAT        CCCFFFFFHHHHHJJFHIIJJJJJJJJGIJJJJJJJIIJGIJJJIJIJIJJHHEFEFFFFFECEEEEDDDDDDDDD@DDBDDDDDDDDDCCDDDDDDDDD      AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:96C0T2     YT:Z:UU NH:i:2  HI:i:1
HWI-ST365:344:D272UACXX:4:1201:12971:89917      0       sctg_285        482     3       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACG
CCCCAGAT        @@@DDEFFHHHFHIICGFGIIIJJIJEIIJGJJIJJIJIJJEHJGGGIG@EE6@DDCFDBECEECACDDCDDBDD@?BDDBDDDCDCC>CCBDDDD@?A@      AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:96C0T2     YT:Z:UU NH:i:2  HI:i:1
HWI-ST365:344:D272UACXX:4:1206:1426:62608       256     sctg_285        482     3       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACG
CCCCAGAT        CCBFFFFFHHHHHJJGHHJJJIJJJJGIJJJJJIJJJJJJJIJJJJIJJHHHFFFFFFEEEEDECDDCDDDDDDBDDDDDDDDDDCCDDDDDDDDD      AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:96C0T2     YT:Z:UU NH:i:2  HI:i:1
HWI-ST365:344:D272UACXX:4:1307:16528:39263      256     sctg_285        482     3       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACG
CCCCAGAT        BCBFFFFFHHHHHHIIAEGIJJIJJJJHIJJJJJJJJJJJJIJIJJJJJHHHADFFFFFEEEEEEDDDDDDDDDDDDCDDDDDDDCCCBDDDDDDDD      AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:96C0T2     YT:Z:UU NH:i:2  HI:i:1
HWI-ST365:344:D272UACXX:4:2106:8182:82314       256     sctg_285        482     3       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACG
CCCCAGAT        B@@FFFFFHHHHHJJCGHIGIGIJJJFHIJJJJJIJJJJJJJJIJJIJJGHHACFFEFFDEDEEEEDDDDDDDDDDBDDDDDBDDDDCCDDDDDDDBDD      AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:96C0T2     YT:Z:UU NH:i:2  HI:i:1
HWI-ST365:344:D272UACXX:4:2316:1898:65958       256     sctg_285        482     3       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACG
CCCCAGAT        BCBDFFFFHGHFDGI@FHGIJJJJJJHIIJJGIJIJIJIJJIJJJIJIHHHHFDFCEFCEDDEDEDDCDDDDDDDDDDDDDDDDCDD@CCBBDDDDDDD      AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:96C0T2     YT:Z:UU NH:i:2  HI:i:1
HWI-ST365:344:D272UACXX:4:1309:18012:38728      256     sctg_285        482     3       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACG
CCCCAGAT        @@@DDFFFHGFDHIJEFFEHIJJJJIFGIGGIJDHIIJIJIIIIJGIIHI@EH=BDCDFF@ACEEDECDDDDDDDDBBBBCDDDDDDDCDDBDDDDDDBD      AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:96C0T2     YT:Z:UU NH:i:2  HI:i:1
HWI-ST365:344:D272UACXX:4:2112:3562:91936       0       sctg_285        482     3       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACG
CCCCAGAT        =??DDFFFHHHHHIIGIHIJJHJJJJHIJJJJJJJJJJIJJJJJIJJJHHHHFFFFFFEEEEDEEDDDDDDDDDDBDDDDDDDDDDCDDDDDDDDDD      AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:96C0T2     YT:Z:UU NH:i:2  HI:i:1
HWI-ST365:344:D272UACXX:4:2303:10772:3449       0       sctg_285        482     3       100M    *       0       0       GGTGTCTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAAACAAATGAGTACG
CCCCAGAT        @@BDDFFFHHHHFHJJEHHIJIIJGIIFHIJJJIGIIJJIJJIJIGGGGIBEEEDFFFFFBCEEECECDDDDDDDBB@BDDDDDDDDDCDDDDBBDB?BC      AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:96C0T2     YT:Z:UU NH:i:2  HI:i:1
HWI-ST132_0435:3:5:11248:111084#GATCAG  0       sctg_285        487     50      94M     *       0       0       CTCTAGTAGGTGTAGGAACGGTGCTCCGTCTCGAAAGGGACGCGTGGTCAATGGTCAAATGACTACGGCAAGCAACAAATGAGTACGCCTACTA  ddddda^abbcd_eefdffffefdffedbfeea\ed[cdaffad`fea]Lcc`c^eeedefbfffeddSdZ^^``Z\`[NZ]_Y^bbcd_Ycc`      AS:i:-12        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:89C0C3     YT:Z:UU NH:i:1
HWI-ST132_0435:3:4:13165:48628#GATCAG   0       sctg_285        585     50      94M     *       0       0       GTTTTTCCCACAAGCACACACGCACAACAGCACACACCTAGTTTCTCATGAACGGTCCATAGGTTCGGGCGGGGATATGTGTGGAGCTATGATG  gdggggeeeegdggcgggggggggegggggggggedgggggdeeggeegggdd\daccce^`bXa\\aaaabUbUb^c_`bb_cZRO\UW`V__      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:94 YT:Z:UU NH:i:1
HWI-ST365:344:D272UACXX:5:2112:2304:90892       0       sctg_285        603     50      100M    *       0       0       CACGCACAACAGCACACACCTAGTTTCTCATGAACGGTCCATAGGTTCGGGCGGGGATATGTGTGGAGCTATGATGGTCGTGTACATCTGTG
TTGCTGTT        C@CFFFFFFHHHGIIIIIJIJJJHIIIJGGIJIIJJJHIJJIFIJGGGHIJJHDDD8@DDDCDCBDDDDDDDDCDDDDDDDDDDEEEDDDEDDDDDDDDD      AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:100        YT:Z:UU NH:i:1
HWI-ST365:344:D272UACXX:5:1103:4944:62465       16      sctg_285        607     50      100M    *       0       0       CACAACAGCACACACCTAGTTTCTCATGAACGGTCCATAGGTTCGGGCGGGGATATGTGTGGAGCTATGATGGTCGTGTACATCTGTGTTGC
```

SAM (Sequence Alignment/Map)

@SQ SN:sctg_997 LN:6840 @SQ SN:sctg_998 LN:6818 @SQ SN:sctg_999 LN:6999		Reference sequence dictionary
@PG ID:TopHat VN:2.0.3		Program
HWI-ST132_0435:3:63:3889:100528#GATCAG		Sequence ID
0		Flag
sctg_0		Reference sequence name
476		Leftmost mapping position
3		Mapping quality
73M		CIGAR string
*		Reference of the mate/next read
0		Mapping position of mate/next read
0		Insert size
CCCGCCGCTCCATGATCTCCAAGAGGCGCAGCTCTCGCAAGGCTTCCGCCAAGGTGGTGGCTT		Sequence ID
ggggggggggggggggggggggggggggeeeggeeggYb^ce^bbbc_cac[ddacaa_c		Quality
AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:73 YT:Z:UU XS:A:+ NH:i:2 CC:Z:sctg_861 CP:i:9032 HI:i:0		Mapper tag

http://samtools.sourceforge.net/SAMv1.pdf

SAM flags

http://picard.sourceforge.net/explain-flags.html

This utility explains SAM flags in plain English.

Flag: 16 Explain

Explanation:
- [ ] read paired
- [ ] read mapped in proper pair
- [ ] read unmapped
- [ ] mate unmapped
- [x] read reverse strand
- [ ] mate reverse strand
- [ ] first in pair
- [ ] second in pair
- [ ] not primary alignment
- [ ] read fails platform/vendor quality checks
- [ ] read is PCR or optical duplicate

Summary:
read reverse strand

This utility explains SAM flags in plain English.

Flag: 99 Explain

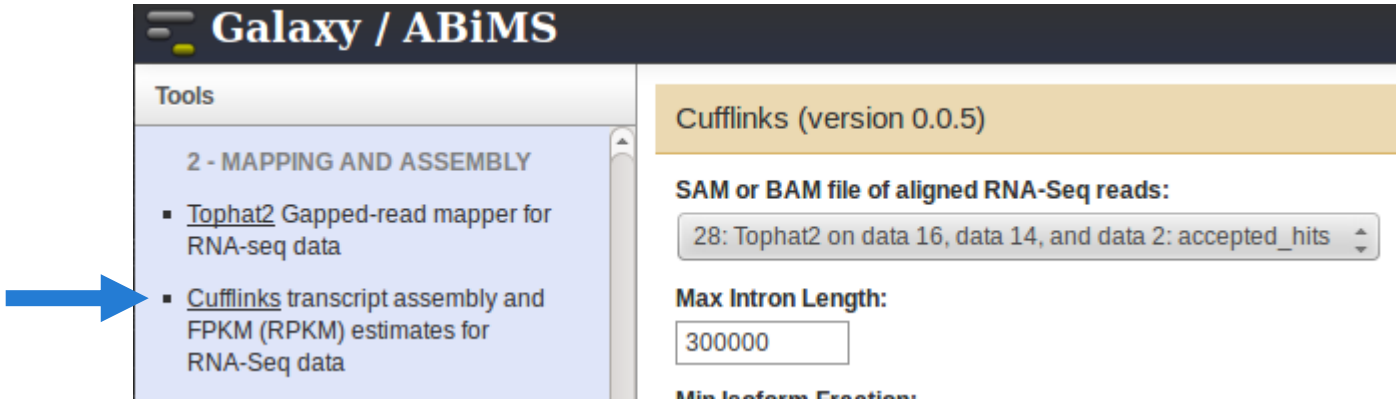
Explanation:
- [x] read paired
- [x] read mapped in proper pair
- [ ] read unmapped
- [ ] mate unmapped
- [ ] read reverse strand
- [x] mate reverse strand
- [x] first in pair
- [ ] second in pair
- [ ] not primary alignment
- [ ] read fails platform/vendor quality checks
- [ ] read is PCR or optical duplicate

Summary:
read paired
read mapped in proper pair
mate reverse strand
first in pair

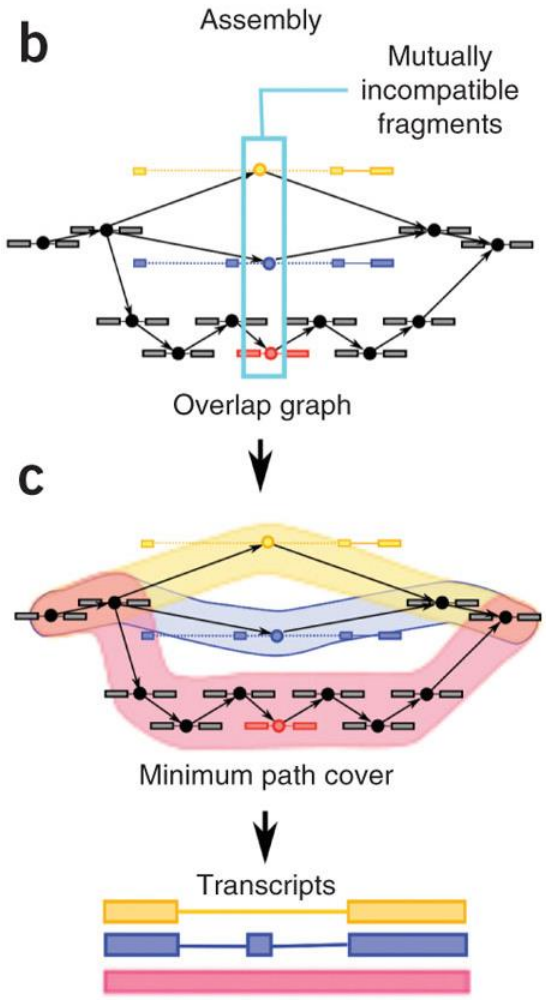
Transcripts assembly with Cufflinks 2

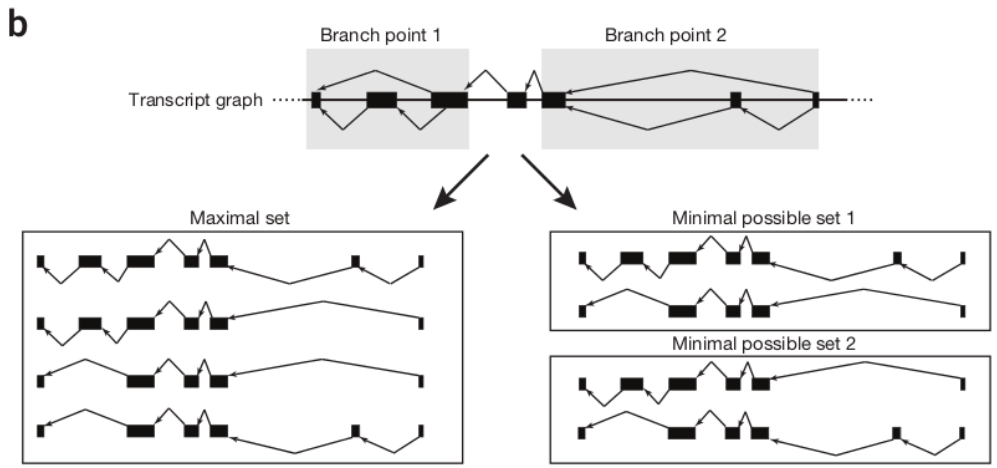
Assembly: Cufflinks

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples.



Assembly: Cufflinks





Scripture
Sensitivity ↗

Cufflinks
Precision ↗

Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech* **28,** 511–515 (2010).
Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth* **8,** 469–477 (2011).

Cufflinks | Input

- BAM
- Genome
- Annotations

One assembly per replicate in case of DE analysis

Cufflinks | Parameters

Cufflinks (version 0.0.5)

SAM or BAM file of aligned RNA-Seq reads:

28: Tophat2 on data 16, data 14, and data 2: 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 diffe

Use Reference Annotation:

Use reference annotation as guide

Reference Annotation:

11: hg19_chr22.gtf

Gene annotation dataset in GTF or GFF3 format.

Perform Bias Correction:

Yes

Bias detection and correction can significantly improve accuracy

Reference sequence data:

History

Using reference file:

2: chr22.fa

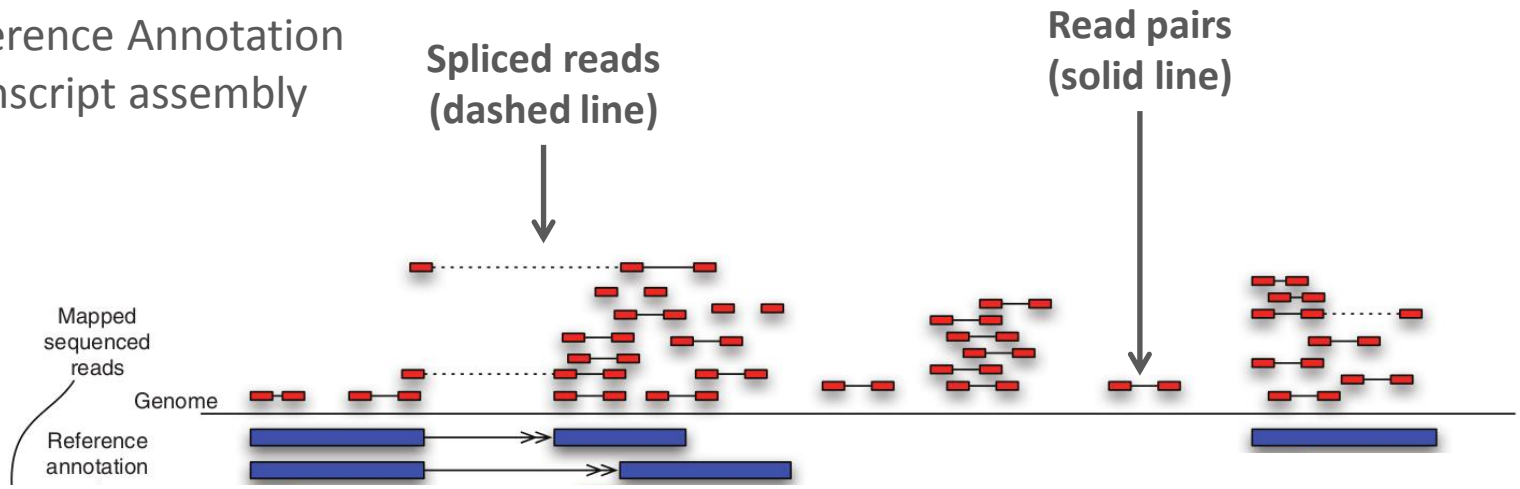
Use multi-read correct:

No

Tells Cufflinks to do an initial estimation procedure to more accur

Why use annotation?

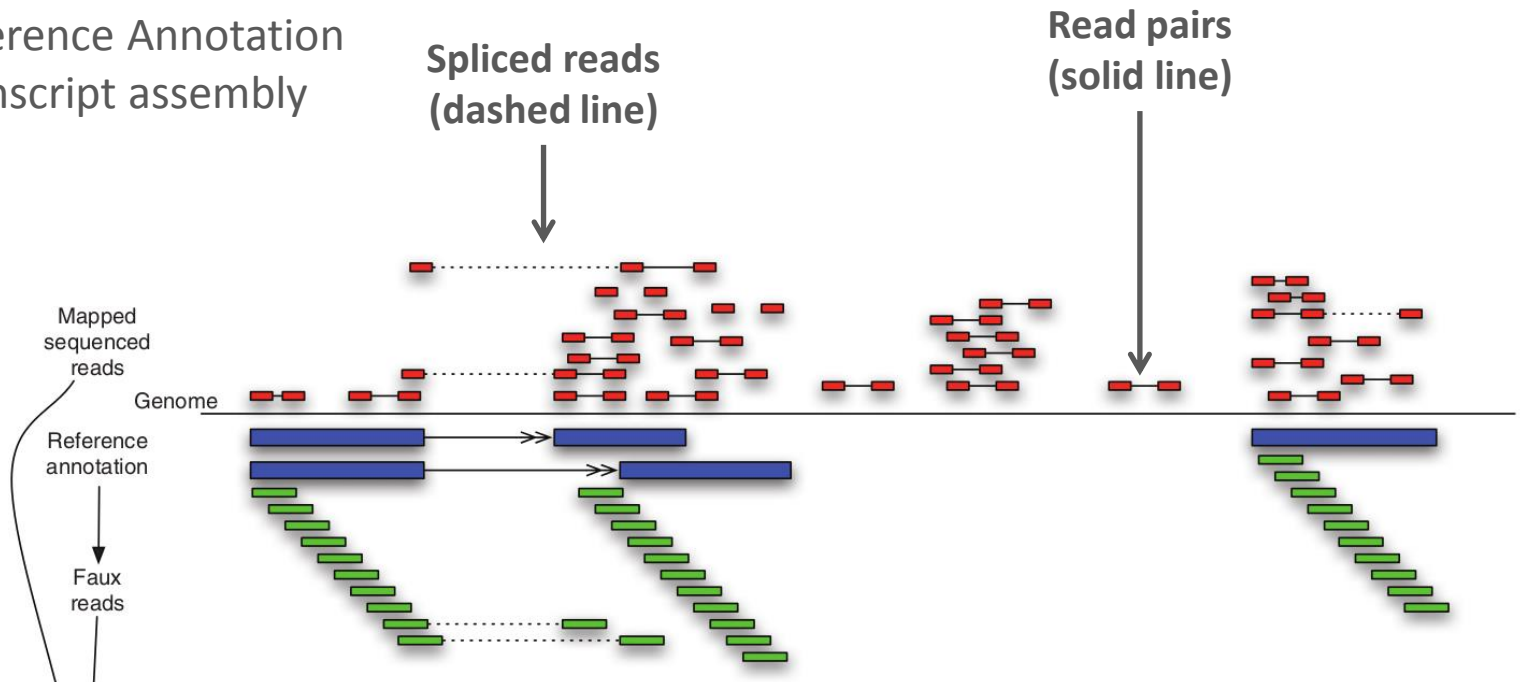
RABT: Reference Annotation Based Transcript assembly



Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* btr355 (2011). doi:10.1093/bioinformatics/btr355

Why use annotation?

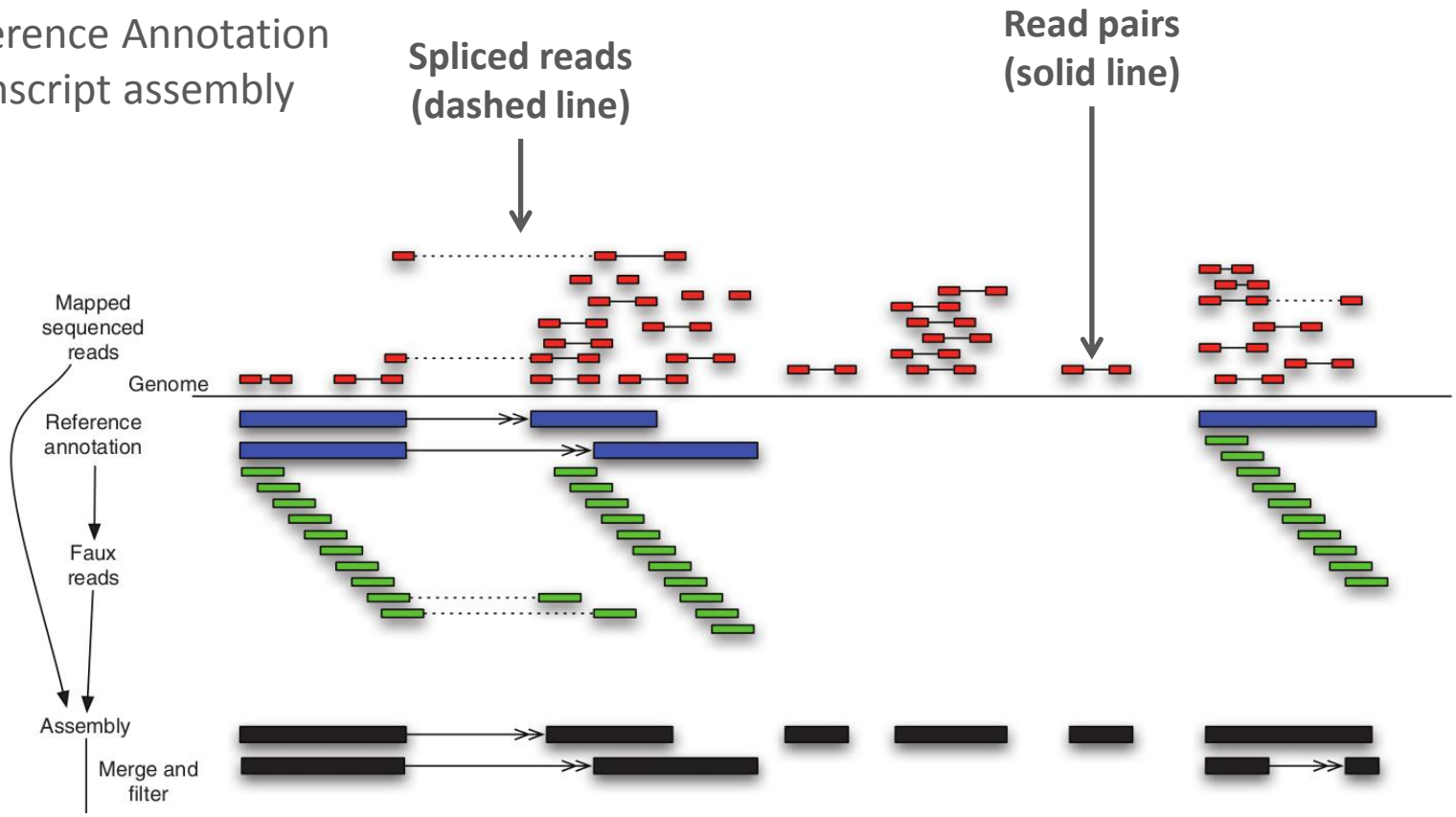
RABT: Reference Annotation Based Transcript assembly



Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* btr355 (2011). doi:10.1093/bioinformatics/btr355

Why use annotation?

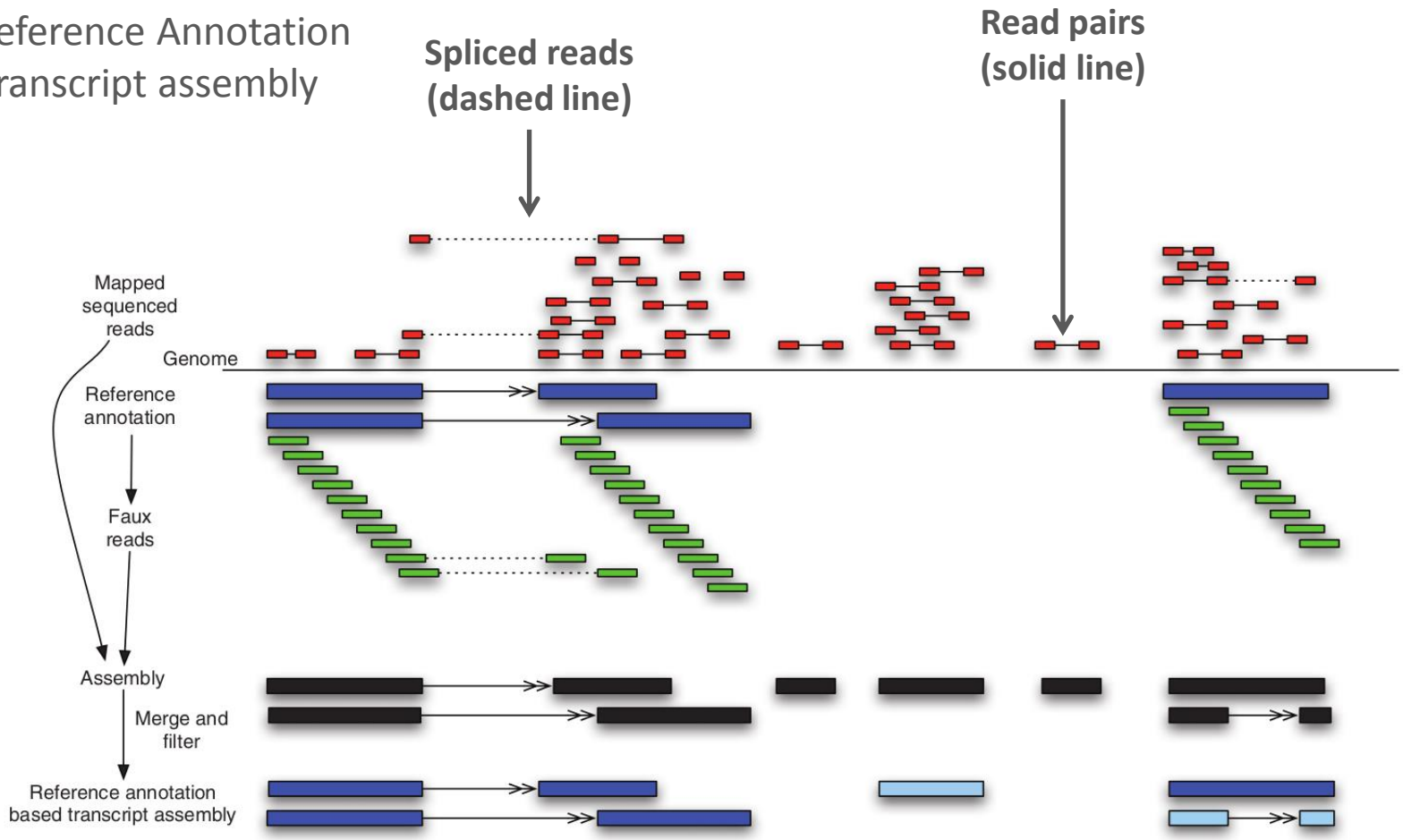
RABT: Reference Annotation Based Transcript assembly



Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* btr355 (2011). doi:10.1093/bioinformatics/btr355

Why use annotation?

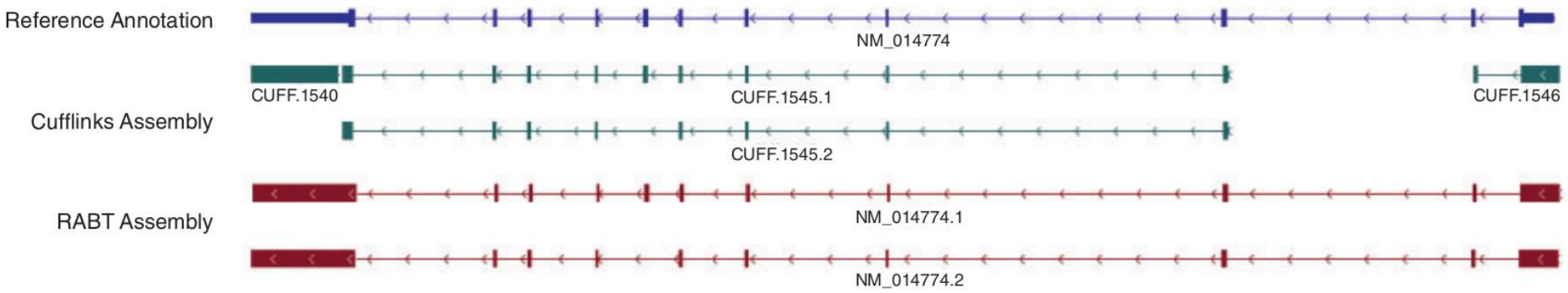
RABT: Reference Annotation Based Transcript assembly



Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* btr355 (2011). doi:10.1093/bioinformatics/btr355

Why use annotation?

RABT: Reference Annotation Based Transcript assembly



Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* btr355 (2011). doi:10.1093/bioinformatics/btr355

gtf/gff format

GFF (general feature format) is a file format used for describing genes and other features of DNA, RNA and protein sequences.

gff3

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attribute
chr22	protein_coding	gene	19701987	19712295	.	+	.	ID=ENSG00000184702;Name=SEPT5
chr22	protein_coding	mRNA	19707711	19708397	.	+	.	ID=ENST00000413258;Name=SEPT5-016;Parent=ENSG00000184702
chr22	protein_coding	protein	19707711	19708397	.	+	.	ID=ENSP00000404673;Name=SEPT5-016;Parent=ENST00000413258
chr22	protein_coding	CDS	19707711	19707761	.	+	1	Name=CDS:SEPT5;Parent=ENST00000413258
chr22	protein_coding	CDS	19707843	19707977	.	+	1	Name=CDS:SEPT5;Parent=ENST00000413258
chr22	protein_coding	CDS	19708165	19708189	.	+	1	Name=CDS:SEPT5;Parent=ENST00000413258
chr22	protein_coding	CDS	19708291	19708397	.	+	0	Name=CDS:SEPT5;Parent=ENST00000413258
chr22	protein_coding	exon	19707711	19707761	.	+	.	Parent=ENST00000413258
chr22	protein_coding	exon	19707843	19707977	.	+	.	Parent=ENST00000413258
chr22	protein_coding	exon	19708165	19708189	.	+	.	Parent=ENST00000413258
chr22	protein_coding	exon	19708291	19708397	.	+	.	Parent=ENST00000413258

Cufflinks | Output

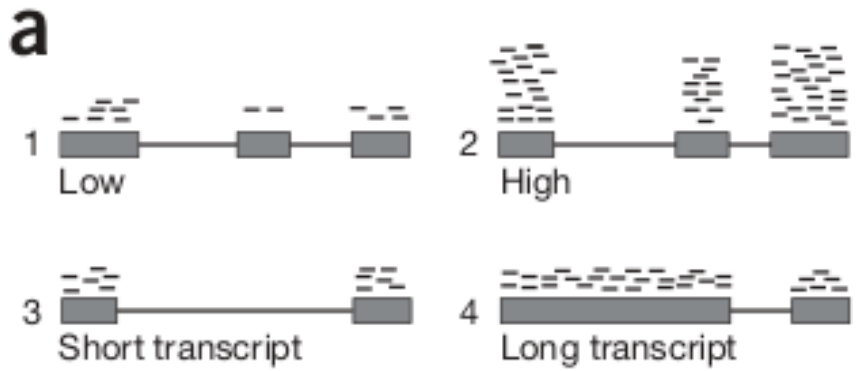
- GTF file (x4)

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr22	**Cufflinks**	**transcript**	**16122830**	**16124132**	**1000**	**.**	**.**	**gene_id "CUFF.1"; transcript_id "CUFF.1.1"; FPKM "148.5475880585"; frac "1.000000"; conf_lo "130.187774"; conf_hi "166.351044"; cov "16.147352"; full_read_support "yes";**
chr22	Cufflinks	exon	16122830	16124132	1000	.	.	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; exon_number "1"; FPKM "148.5475880585"; frac "1.000000"; conf_lo "130.187774"; conf_hi "166.351044"; cov "16.147352";
chr22	Cufflinks	transcript	16256332	16287937	1	-	.	gene_id "NM_001136213"; transcript_id "NM_001136213"; FPKM "0.0000000000"; frac "0.000000"; conf_lo "0.000000"; conf_hi "0.000000"; cov "0.000000"; full_read_support "no";
chr22	Cufflinks	exon	16256332	16256677	1	-	.	**gene_id "NM_001136213";** transcript_id "NM_001136213"; exon_number "1"; FPKM "0.0000000000"; frac "0.000000"; conf_lo "0.000000"; conf_hi "0.000000"; cov "0.000000";
chr22	Cufflinks	exon	16258185	16258303	1	-	.	gene_id "NM_001136213"; transcript_id "NM_001136213"; exon_number "2"; FPKM "0.0000000000"; frac "0.000000"; conf_lo "0.000000"; conf_hi "0.000000"; cov "0.000000";
chr22	Cufflinks	exon	16266929	16267095	1	-	.	gene_id "NM_001136213"; transcript_id "NM_001136213"; exon_number "3"; FPKM "0.0000000000"; frac "0.000000"; conf_lo "0.000000"; conf_hi "0.000000"; cov "0.000000";
chr22	Cufflinks	exon	16268137	16268181	1	-	.	gene_id "NM_001136213"; transcript_id "NM_001136213"; exon_number "4"; FPKM "0.0000000000"; frac "0.000000"; conf_lo "0.000000"; conf_hi "0.000000"; cov "0.000000";
chr22	Cufflinks	exon	16269873	16269943	1	-	.	gene_id "NM_001136213"; transcript_id "NM_001136213"; exon_number "5"; FPKM "0.0000000000"; frac "0.000000"; conf_lo "0.000000"; conf_hi "0.000000"; cov "0.000000";

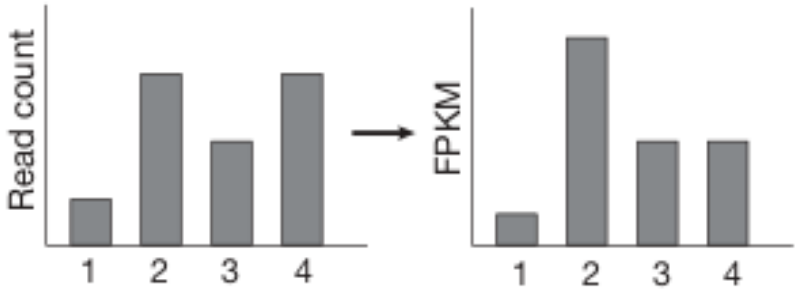
FPKM (RPKM)

- **F**ragments ~~Reads~~ **P**er **K**ilobase of exon model per **M**illion mapped fragments

$$FPKM = 10^9 \times \frac{C}{NL}$$



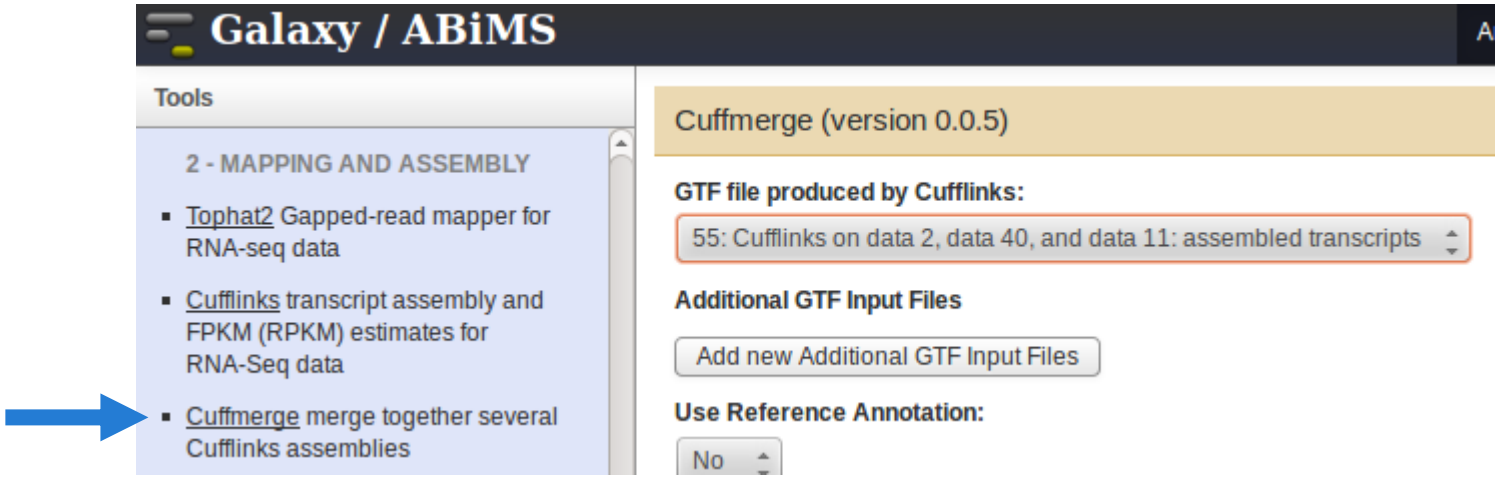
C= the number of reads mapped onto the gene's exons
N= total number of mapped reads
L= the sum of the exons in base pairs (transcript length)



Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth* **5,** 621–628 (2008).
Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech* **28,** 511–515 (2010).

Assembly: Cuffmerge

Cuffmerge is used to merge together several Cufflinks assemblies. It also handles running Cuffcompare for you, and automatically filters a number of transfrags that are probably artifacts.



Cuffmerge | Input

- gtf from Cufflinks
- Genome
- Annotation

Cuffmerge | Parameters

Cuffmerge (version 0.0.5)

GTF file produced by Cufflinks:

43: Cufflinks on data 2, data 28, and data 11: assembled transcripts

Additional GTF Input Files

Additional GTF Input Files 1

GTF file produced by Cufflinks:

47: Cufflinks on data 2, data 32, and data 11: assembled transcripts

Remove Additional GTF Input Files 1

Additional GTF Input Files 2

GTF file produced by Cufflinks:

51: Cufflinks on data 2, data 36, and data 11: assembled transcripts

Remove Additional GTF Input Files 2

Additional GTF Input Files 3

GTF file produced by Cufflinks:

55: Cufflinks on data 2, data 40, and data 11: assembled transcripts

Remove Additional GTF Input Files 3

Add new Additional GTF Input Files

Use Reference Annotation:

Yes

Reference Annotation:

11: hg19_chr22.gtf

Requires an annotation file in GFF3 or GTF format.

Use Sequence Data:

Yes

Use sequence data for some optional classification functions, including the ad

Choose the source for the reference list:

History

Using reference file:

2: chr22.fa

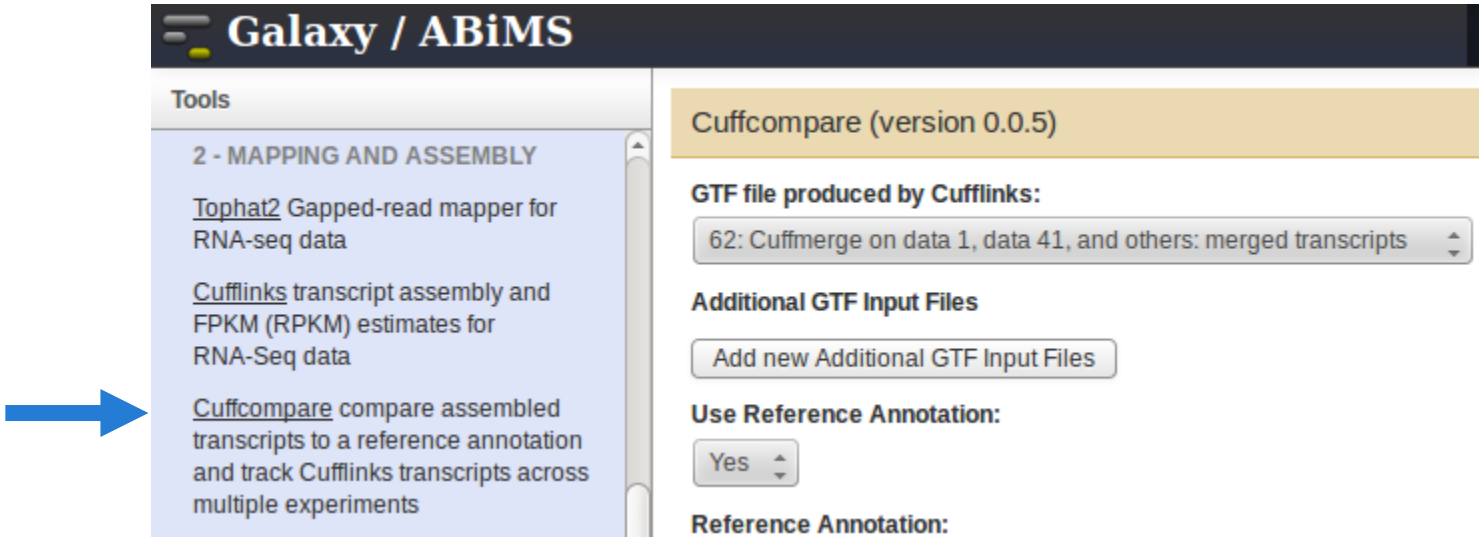
Cuffmerge | Output

- gtf (x1)

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr22	**Cufflinks**	**exon**	**16162066**	**16162388**	**.**	**+**	**.**	**gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "1"; gene_name "NR_073460"; oId "NR_073459"; nearest_ref "NR_073460"; class_code "="; tss_id "TSS1";**
chr22	Cufflinks	exon	16164482	16164569	.	+	.	gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "2"; gene_name "NR_073460"; oId "NR_073459"; nearest_ref "NR_073460"; class_code "="; tss_id "TSS1";
chr22	Cufflinks	exon	16171952	16172265	.	+	.	gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "3"; gene_name "NR_073460"; oId "NR_073459"; nearest_ref "NR_073460"; class_code "="; tss_id "TSS1";
chr22	Cufflinks	exon	16414985	16415982	.	+	.	gene_id "XLOC_000002"; transcript_id "TCONS_00000002"; exon_number "1"; oId "CUFF.3.1"; class_code "u"; tss_id "TSS2";
chr22	Cufflinks	exon	16414987	16415562	.	+	.	gene_id "XLOC_000002"; transcript_id "TCONS_00000003"; exon_number "1"; oId "CUFF.4.1"; class_code "u"; tss_id "TSS2";
chr22	Cufflinks	exon	16415764	16415930	.	+	.	gene_id "XLOC_000002"; transcript_id "TCONS_00000003"; exon_number "2"; oId "CUFF.4.1"; class_code "u"; tss_id "TSS2";
chr22	Cufflinks	exon	17082801	17083105	.	+	.	gene_id "XLOC_000003"; transcript_id "TCONS_00000004"; exon_number "1"; gene_name "NR_001591"; oId "NR_001591"; nearest_ref "NR_001591"; class_code "="; tss_id "TSS3";
chr22	Cufflinks	exon	17092548	17092783	.	+	.	gene_id "XLOC_000003"; transcript_id "TCONS_00000004"; exon_number "2"; gene_name "NR_001591"; oId "NR_001591"; nearest_ref "NR_001591"; class_code "="; tss_id "TSS3";
chr22	Cufflinks	exon	17094967	17095068	.	+	.	gene_id "XLOC_000003"; transcript_id "TCONS_00000004"; exon_number "3"; gene_name "NR_001591"; oId "NR_001591"; nearest_ref "NR_001591"; class_code "="; tss_id "TSS3";

Assembly: Cuffcompare

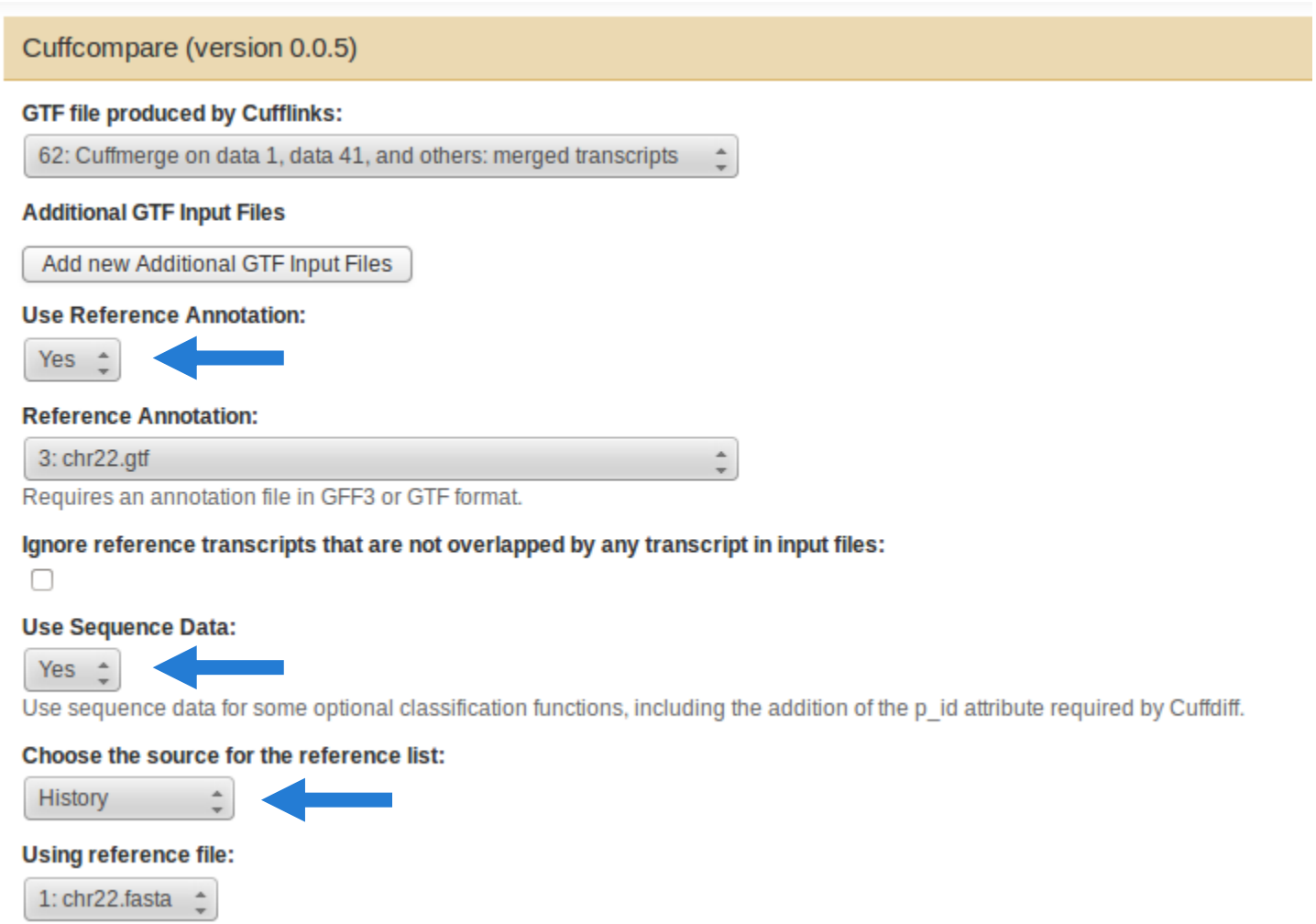
Cuffcompare is used to compare assembled transcripts to a reference annotation.



Cuffcompare | Input

- gtf from Cufflinks / Cuffmerge
- Reference annotation
- Genome

Cuffcompare | Parameters



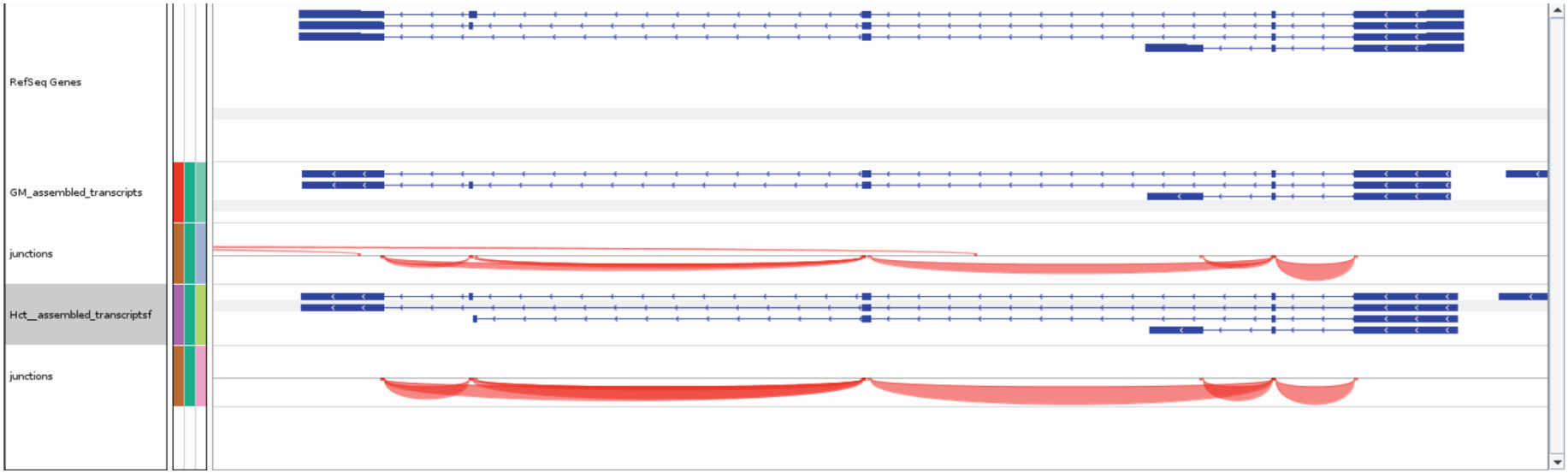
Cuffcompare | Output

The following table shows the code used by Cufflinks to classify the transcripts in comparison with the reference annotation

Priority	Code	Description
1	**=**	**Complete match of intron chain**
2	c	Contained
3	**j**	**Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript**
4	e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment
5	i	A transfrag falling entirely within a reference intron
6	o	Generic exonic overlap with a reference transcript
7	p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
8	r	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
9	**u**	**Unknown, intergenic transcript**
10	**x**	**Exonic overlap with reference on the opposite strand**
11	s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)
12	.	(.tracking file only, indicates multiple classifications)

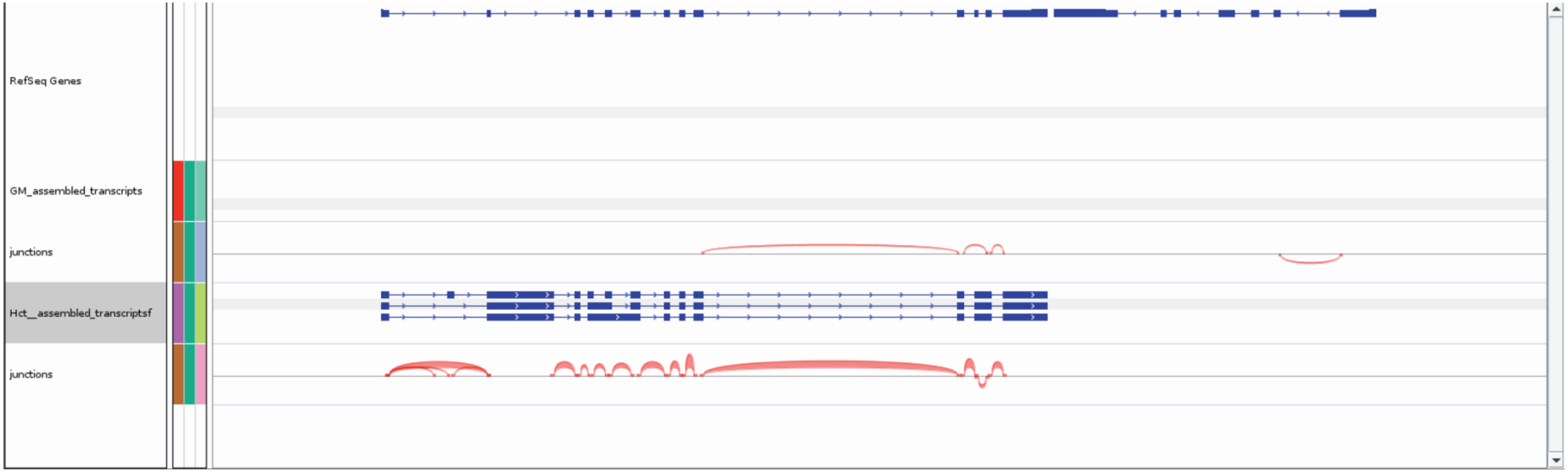
Examples

=


RefSeq Genes
GM_assembled_transcripts
junctions
Hct__assembled_transcriptsf
junctions

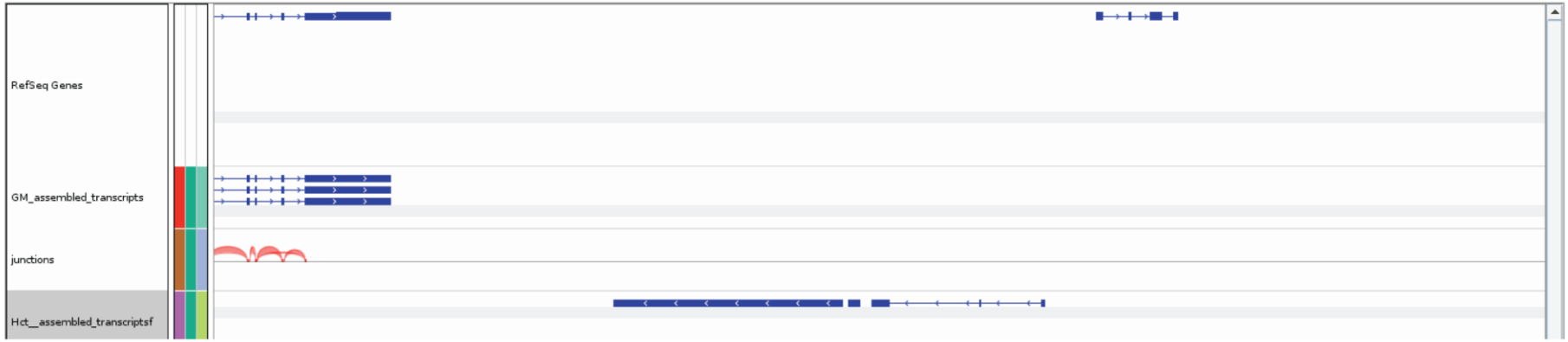
Examples

J


RefSeq Genes
GM_assembled_transcripts
junctions
Hct__assembled_transcriptsf
junctions

Examples

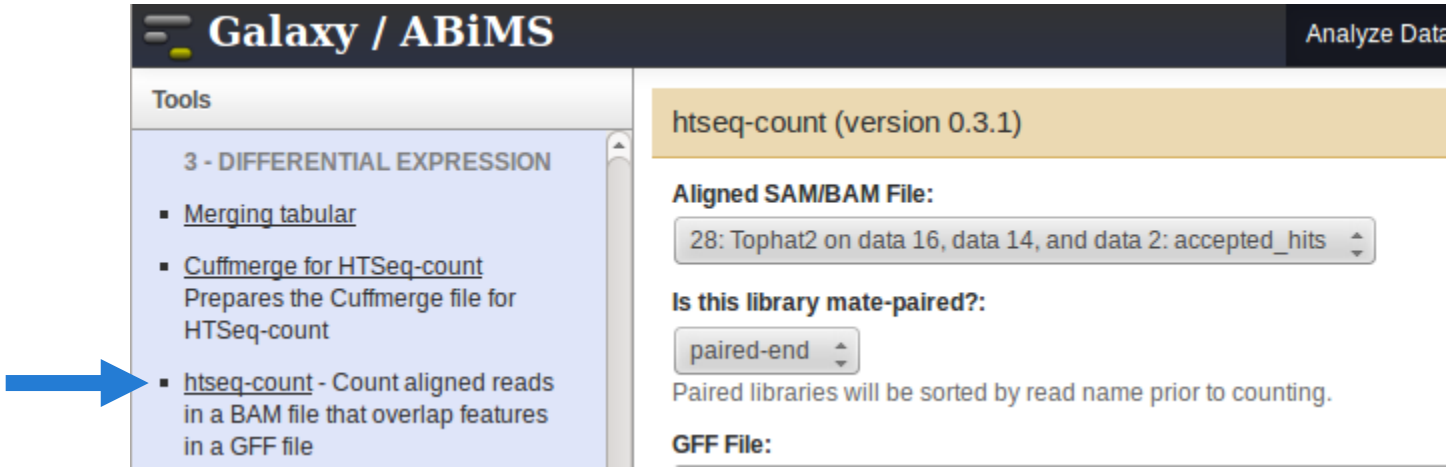
U


RefSeq Genes
GM_assembled_transcripts
junctions
Hct__assembled_transcriptsf

Read counting per gene with HTSeq-count

Counting: HTSeq

HTSeq is a Python package that provides infrastructure to process data from high-throughput sequencing assays.



HTSeq | Input

- BAM
- gtf/gtf annotation file

One counting per replicate

HTSeq | Parameters

htseq-count (version 0.3.1)

Aligned SAM/BAM File:

25: Tophat2 on data 15, data 14, and data 1: accepted_hits

Is this library mate-paired?:

paired-end

Paired libraries will be sorted by read name prior to counting.

GFF File:

54: Cuffmerge on data 1, data 40, and others: merged transcripts

Mode:

Intersection (nonempty)

Mode to handle reads overlapping more than one feature.

Stranded:

No

Specify whether the data is from a strand-specific assay. 'Reverse' means ye

Minimum alignment quality:

0

Skip all reads with alignment quality lower than the given minimum value

Feature type:

exon

Feature type (3rd column in GFF file) to be used. All features of other types a

ID Attribute:

gene_id

GFF attribute to be used as feature ID. Several GFF lines with the same featu
specified type MUST have a value for this attribute. The default, suitable for F

Additional BAM Output:

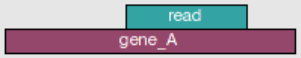
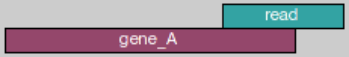
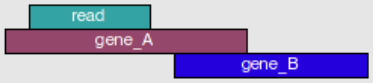
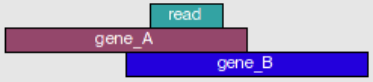
Write out all SAM alignment records into an output BAM file, annotating each

HTSeq | Mode

Mode:

Intersection (nonempty)

Mode to handle reads overlapping more than one feature.

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

HTSeq | Attribute

chr22	hg19_refGene	CDS	17443626	17443766	0.000000	-	0	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	exon	17442827	17443766	0.000000	-	.	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	CDS	17444615	17444719	0.000000	-	0	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	exon	17444615	17444719	0.000000	-	.	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	CDS	17445656	17445752	0.000000	-	1	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	exon	17445656	17445752	0.000000	-	.	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	CDS	17446068	17446158	0.000000	-	2	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	exon	17446068	17446158	0.000000	-	.	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	CDS	17446990	17447254	0.000000	-	0	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	exon	17446990	17447254	0.000000	-	.	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	CDS	17449188	17449273	0.000000	-	2	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	exon	17449188	17449273	0.000000	-	.	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	CDS	17450833	17451083	0.000000	-	1	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	exon	17450833	17451083	0.000000	-	.	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	CDS	17468850	17469057	0.000000	-	2	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	exon	17468850	17469057	0.000000	-	.	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	CDS	17472763	17473066	0.000000	-	0	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	exon	17472763	17473066	0.000000	-	.	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	CDS	17488831	17489004	0.000000	-	0	gene_id "NM_001037814"; transcript_id "NM_001037814";
chr22	hg19_refGene	exon	17488831	17489112	0.000000	-	.	gene_id "NM_001037814"; transcript_id "NM_001037814";

Feature

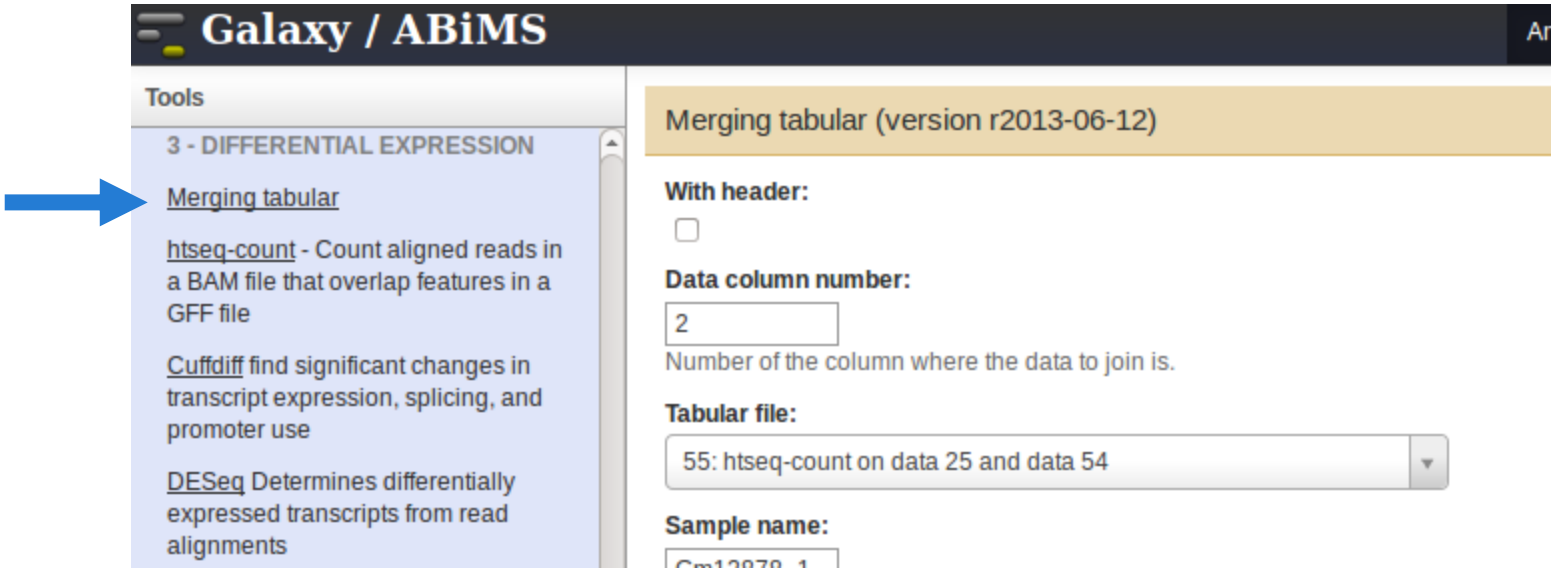
Attribute

HTSeq | Output

- Tabular file (x4)

gene ID	Read count
NM_000026	1256
NM_000106	0
NM_000185	2
NM_000262	3164
NM_000268	0
NM_000343	4
NM_000355	16
NM_000362	181
NM_000395	0
NM_000398	450
NM_000407	0
NM_000487	0
NM_000496	38
NM_000631	0
NM_000675	262
NM_000714	247
NM_000754	149

Merging tabular



Merging tabular | Parameters

Merging tabular (version r2013-06-12)

With header:

☐

Data column number:

2

Number of the column where the data to join is.

Tabular file:

55: htseq-count on data 25 and data 54

Sample name:

Gm12878_1

Write the name corresponding to your sample.

Tabular file:

57: htseq-count on data 29 and data 54

Sample name:

Gm12878_2

Write the name corresponding to your sample.

Tabular files

Tabular file 1

Tabular outputs:

59: htseq-count on data 33 and data 54

Sample name:

Hct116_1

Write the name corresponding to your sample.

Remove Tabular file 1

Tabular file 2

Tabular outputs:

61: htseq-count on data 37 and data 54

Sample name:

Hct116_2

Write the name corresponding to your sample.

Remove Tabular file 2

Add new Tabular file

Merging tabular | Output

- A matrix

	Gm12878_1	Gm12878_2	Hct116_1	Hct116_2
NM_001003891	86	98	140	139
NM_033200	1379	1639	3499	3583
NM_152513	523	589	36	33
NM_015330	7	8	17	19
NR_046423	0	0	2	1
NR_026815	58	73	256	238
NR_001283	22	44	20	31
NM_001198726	0	0	0	0
NM_032050	0	0	0	0
NR_037611	0	4	19	10
NM_177405	1	2	0	0
NM_019008	2433	2789	4233	4494
NM_014292	1927	1874	5186	5120
NM_024821	157	178	278	298
NM_018943	11	13	0	0
NM_033070	48	63	180	182
NR_038949	2	4	0	0
NM_001130921	0	0	0	0
NM_001130919	0	0	0	0
NR_024448	448	519	893	947
NR_002727	4	2	1	0
NM_019106	24	42	707	765
NM_001164501	1	1	0	0
NM_004810	869	1039	4	4

End