

Differential Gene Expression

Konstantin Riege

FLI Jena

Data Source

RESEARCH ARTICLE

A complex association between DNA methylation and gene expression in human placenta at first and third trimesters

Yen Ching Lim¹, Jie Li¹, Yiyun Ni¹, Qi Liang¹, Junjiao Zhang¹, George S. H. Yeo², Jianxin Lyu^{1*}, Shengnan Jin^{1*}, Chunming Ding^{1*}

1 Key Laboratory of Laboratory Medicine, Ministry of Education of China, School of Laboratory Medicine and Life Science, Wenzhou Medical University, Wenzhou, Zhejiang, China, **2** KK Women's and Children's Hospital, Singapore, Singapore

Data Source

- 1st trimester vs. 3rd trimester human RNA-Seq samples (N=4)
 - 1T paired end N1,N2,N3,N4
 - 3T paired end N2,N3
 - 3T single end N1,N4
- ~2400 diff. expressed genes
- 21 of them with diff. methylated promoter regions (1kb upstream region)

Data Source

- Downsampling
 - HG38 toy genome composed of 21 genes of interest (GOI) and 4kb intergenic regions (2kb upstream and 2kb downstream)
 - 10% of GOI mapped reads extracted from available RAW sequencing data

https://github.com/destairdenbi/trainings/tree/master/raw_data/bs_tour

Upload GOI Data

Galaxy / Europe

Tools

search tools

[Get Data](#)

1

Download from web or upload from disk

Regular Composite Collection Rule-based

Please wait... 30 out of 33 remaining.

Name	Size	Type	Genome	Settings	Status
hg38.goi.fa	1.7 MB	Auto-detect	— Additional Speci...	⚙	100% ✓
hg38.goi.gtf	1.6 MB	Auto-detect	— Additional Speci...	⚙	100% ✓
hg38.promoter.bed	816 b	Auto-detect	— Additional Speci...	⚙	100% ✓
rbs_1T.N1.R1.fastq	14.4 MB	Auto-detect	— Additional Speci...	⚙	Adding to history... ⚙
rbs_1T.N1.R2.fastq	14.4 MB	Auto-detect	— Additional Speci...	⚙	0% ⚙
rbs_1T.N2.R1.fastq	20.9 MB	Auto-detect	— Additional Speci...	⚙	0% ⚙
rbs_1T.N2.R2.fastq	20.9 MB	Auto-detect	— Additional Speci...	⚙	0% ⚙

2

Type (set all): Auto-detect

Genome (set all): — Additional Species Are B...

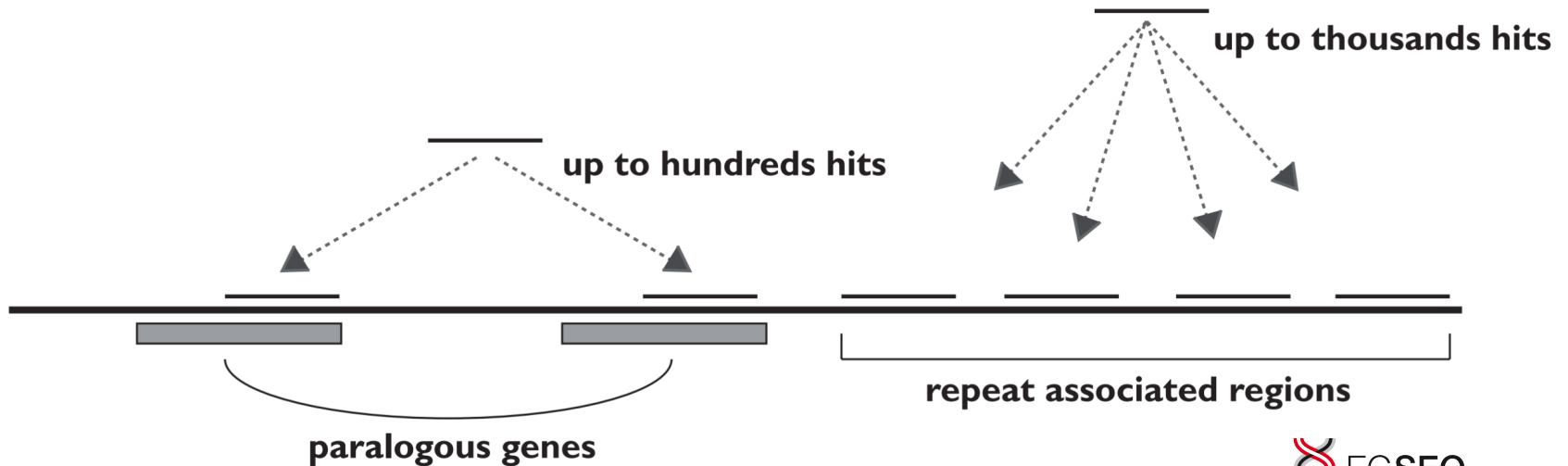
3

Choose local file Choose FTP file Paste/Fetch data Pause Reset Start Close

Learn how the NASA uses Galaxy!

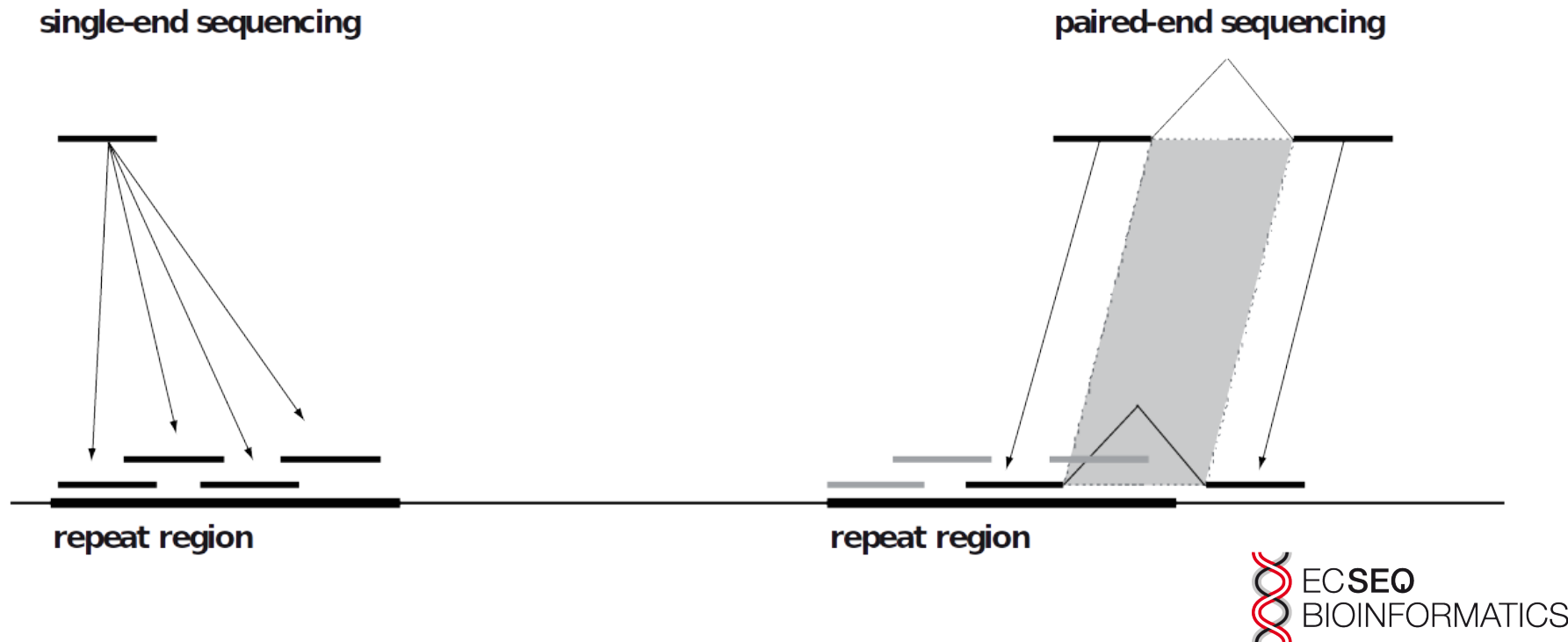
Alignment postprocessing

- Extract uniquely/unambiguously aligned reads



Alignment postprocessing

- Select only alignments with both read mates mapped properly paired



Alignment postprocessing

@HD VN:1.5 S0:coordinate											Header section
@SQ SN:ref LN:45											
r001	99	ref	7	30	8M2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*	Alignment section
r002	0	ref	9	30	3S6M1P1I4M	*	0	0	AAAAGATAAGGATA	*	
r003	0	ref	9	30	5S6M	*	0	0	GCCTAAGCTAA	* SA:Z:ref,29,-,6H5M,17,0;	
r004	0	ref	16	30	6M14N5M	*	0	0	ATAGCTTCAGC	*	
r003	2064	ref	29	17	6H5M	*	0	0	TAGGC	* SA:Z:ref,9,+,5S6M,30,1;	
r001	147	ref	37	30	9M	=	7	-39	CAGCGGCAT	* NM:i:1	

Alignment postprocessing

Header section									
<pre>@HD VN:1.5 S0:coordinate @SQ SN:ref LN:45</pre>									
<pre>r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG * r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA * r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0; r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC * r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1; r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1</pre>									
Alignment section									

Bit	Description
1	0x1 template having multiple segments in sequencing
2	0x2 each segment properly aligned according to the aligner
4	0x4 segment unmapped
8	0x8 next segment in the template unmapped
16	0x10 SEQ being reverse complemented
32	0x20 SEQ of the next segment in the template being reverse complemented
64	0x40 the first segment in the template
128	0x80 the last segment in the template
256	0x100 secondary alignment
512	0x200 not passing quality controls
1024	0x400 PCR or optical duplicate
2048	0x800 supplementary alignment

Alignment postprocessing

- Filter BAM by
 - alignment flags - pair information: 2
 - alignment tags - mapping count: NH:i:*
 - mapping quality - often an indicator for unique alignments
 - STAR: =255
 - HISAT2: >0
- Sort BAM
 1. Samtools view
 2. Samtools sort

Alignment postprocessing

1 Tools

2 [samttools view](#)

3 Filter SAM or BAM output SAM or BAM files on FLAG MAPQ RG LN or by region (Galaxy Version 1.8)

4 SAM or BAM file to filter

176: Bismark Mapper on data 123, data 124, and data 1: mapped reads (as bam)
174: Bismark Mapper on data 121, data 120, and data 1: mapped reads (as bam)
172: Bismark Mapper on data 117, data 116, and data 1: mapped reads (as bam)
170: Bismark Mapper on data 113, data 112, and data 1: mapped reads (as bam)
168: Bismark Mapper on data 109, data 108, and data 1: mapped reads (as bam)

This is a batch mode input field. Separate jobs will be triggered for each dataset

5 Header in output

Include header

6 Minimum MAPQ quality score

1

(-q)

Filter on bitwise flag

yes

7 Skip alignments with any of these flag bits set

☐ Select/Unselect all

☐ Read is paired

☐ Read is mapped in a proper pair

☒ The read is unmapped

☐ The mate is unmapped

☐ Read strand

☐ Mate strand

☐ Read is the first in a pair

☐ Read is the second in a pair

☒ The alignment or this read is not primary

☐ The read fails platform/vendor quality checks

☐ The read is a PCR or optical duplicate

☐ Supplementary alignment

(-f)

8

X

!

VCF/BCF

[bctools call](#) SNP/indel variant calling from VCF/BCF


Variant Calling

[bctools call](#) SNP/indel variant calling from VCF/BCF

Workflows

- [All workflows](#)

Alignment postprocessing

Tools 

1

samtools sort

FASTA/FASTQ

[UMI-tools group](#) Extract UMI from fastq files

SAM/BAM

[Filter pileup](#) on coverage and SNPs

[Generate pileup](#) from BAM dataset

[SAM-to-BAM](#) convert SAM to BAM

[Samtools flagstat](#) tabulate descriptive stats for BAM dataset

[samtools mpileup](#) multi-way pileup of variants

2


Samtools sort order of storing aligned sequences (Galaxy Version 2.0.2)

BAM File

3

4

203: Filter SAM or BAM, output SAM or BAM on data 176: bam
202: Filter SAM or BAM, output SAM or BAM on data 174: bam
201: Filter SAM or BAM, output SAM or BAM on data 172: bam
200: Filter SAM or BAM, output SAM or BAM on data 170: bam
199: Filter SAM or BAM, output SAM or BAM on data 168: bam
198: Filter SAM or BAM, output SAM or BAM on data 166: bam
197: Filter SAM or BAM, output SAM or BAM on data 164: bam
196: Filter SAM or BAM, output SAM or BAM on data 162: bam
176: Bismark Mapper on data 125, data 124, and data 1: mapped

 This is a batch mode input field. Separate jobs will be triggered

Primary sort key

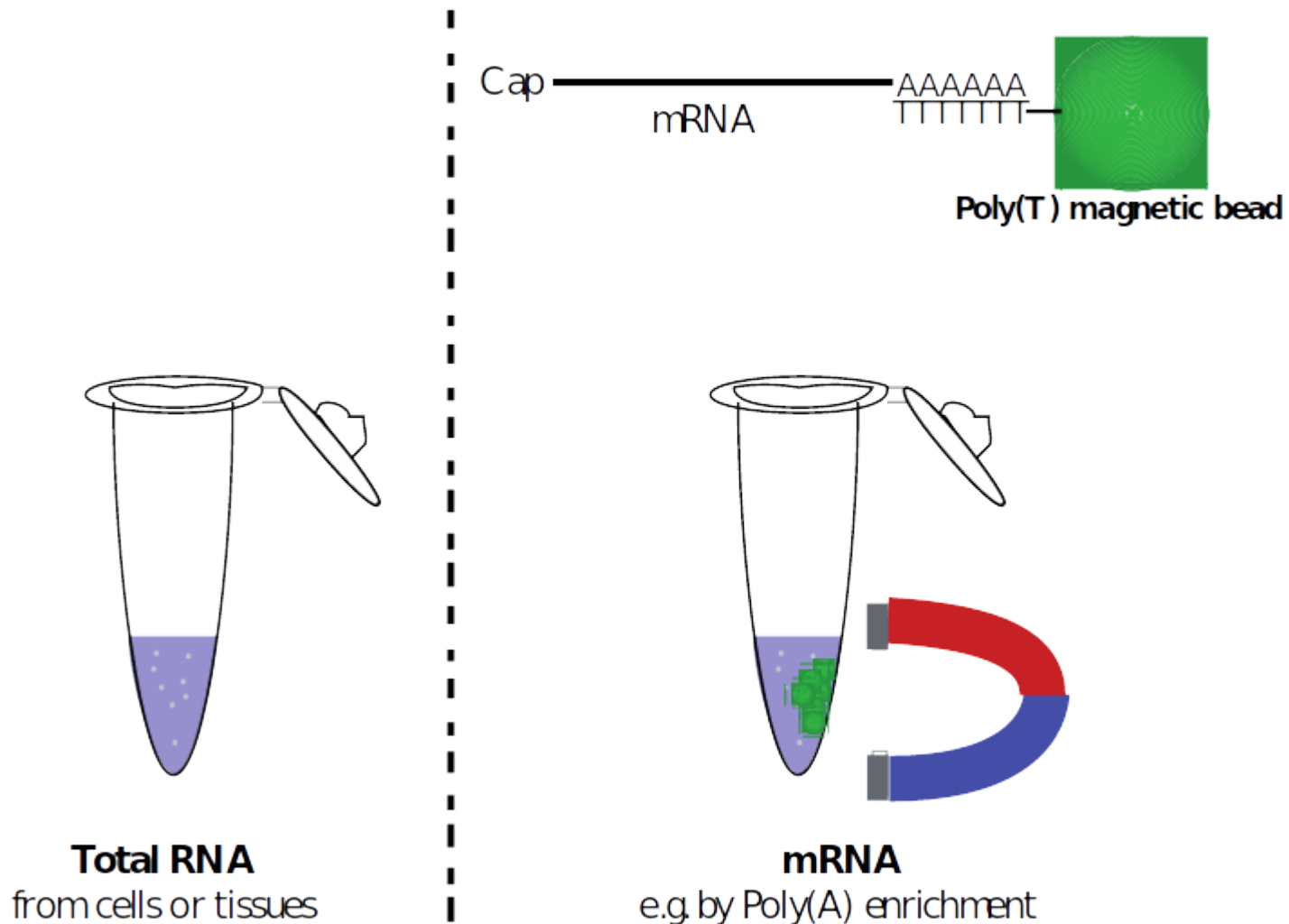
coordinate

✓ Execute

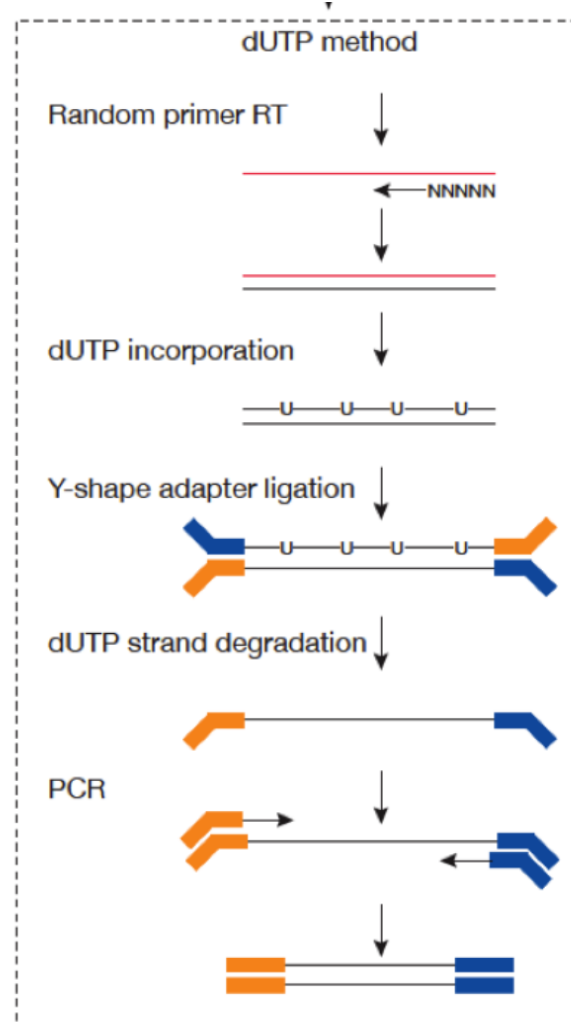
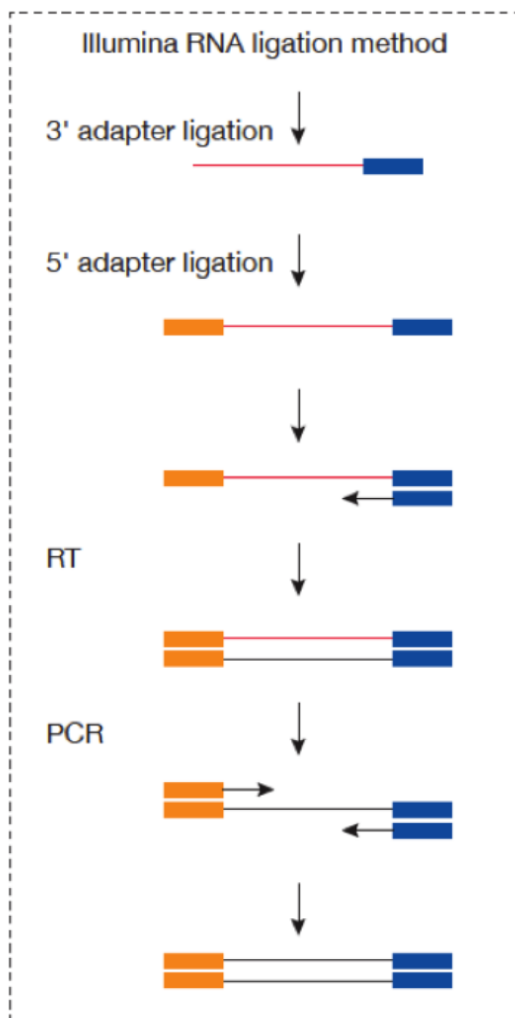
Quantification

- Requirements:
 - Annotation
 - sequencing strandness information
- **DETOUR**: infer experimental setup

Infer experiment strandness



Infer experiment strandness



fragment RNA + cDNA

RNA from cDNA using Deoxyuridine Triphosphate + cRNA

Image Credit: Zhao Zhang

Infer experiment strandness

Single index



Unique dual index



Dual index UMI



Flow cell binding sequence: Platform-specific sequences for library binding to instrument



Sequencing primer sites: Binding sites for general sequencing primers



Sample indexes: Short sequences specific to a given sample library



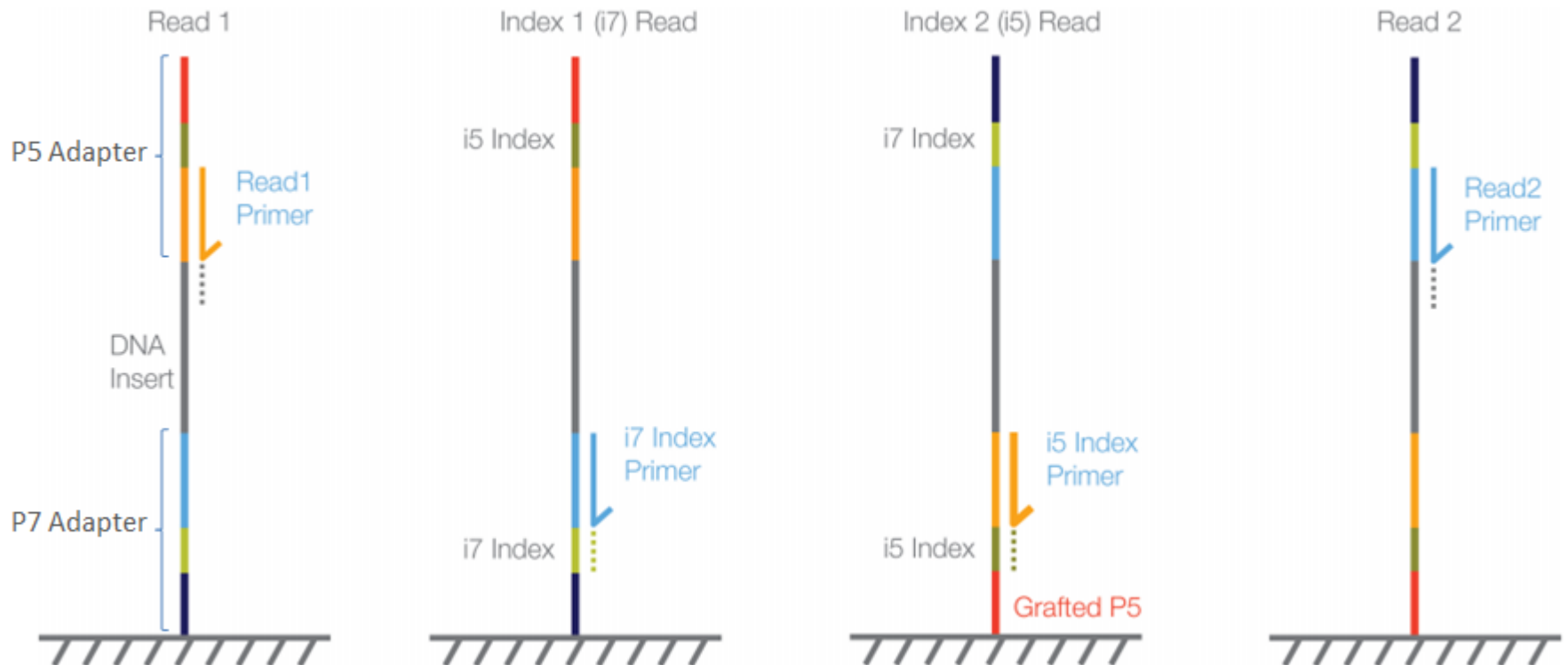
Molecular index/barcode: Short sequence used to uniquely tag each molecule in a given sample library



Insert: Target DNA or RNA fragment from a given sample library

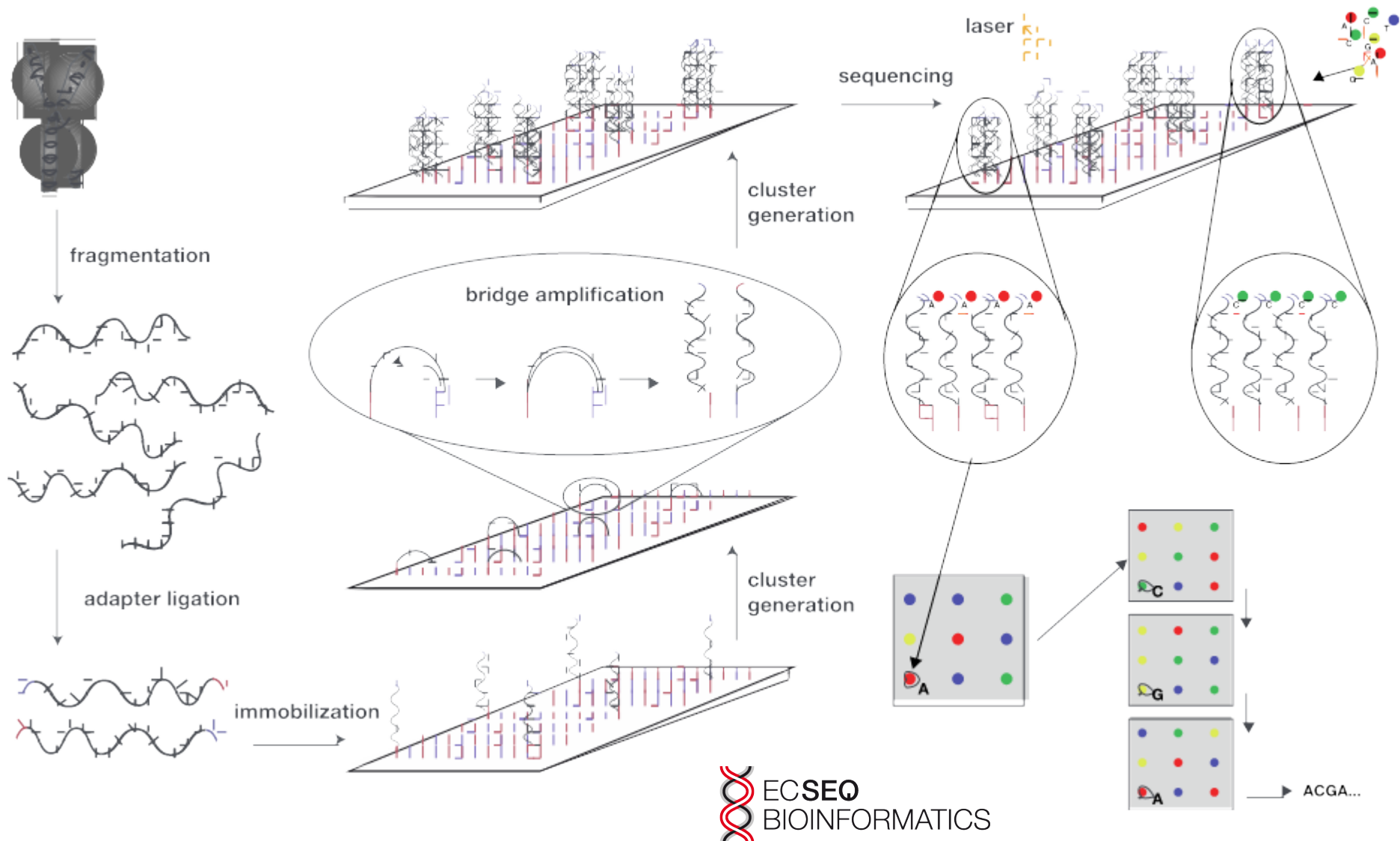
<https://eu.idtdna.com/pages/products/next-generation-sequencing/adapters>

Infer experiment strandness



<https://support.illumina.com/bulletins/2016/04/adapter-trimming-why-are-adapter-sequences-trimmed-from-only-the--ends-of-reads.html>

Infer experiment strandness



Infer experiment strandness

- Tool box: RSeQC
 - Requirements:
 - Annotation in BED format
1. GTF to BED conversion
 2. infer_experiment.py

Infer experiment strandness

The image shows a screenshot of the Galaxy web interface. On the left, the 'Tools' panel is visible, with a search bar containing 'gtf to bed'. Below the search bar, under the 'Convert Formats' section, three tools are listed: 'Convert gffGTF to annotated GTF to BED for SplicingTie results', 'Convert GTF to BED12', and 'GTF-to-BEDGraph converter'. The 'Workflows' section is partially visible at the bottom. On the right, the configuration page for the 'Convert GTF to BED12 (Galaxy Version 1.7)' tool is shown. The 'GTF File to convert' section has three file selection buttons and a text box containing '2: hg38.goi.gtf'. The 'Advanced options' section has a 'Use default options' button and a text box containing 'Advanced options for gtfToGenePred.'. At the bottom of the configuration page is a blue 'Execute' button with a checkmark icon. Three red arrows with black numbers are overlaid on the image: arrow 1 points to the 'Tools' header, arrow 2 points to the 'Convert GTF to BED12' tool in the list, and arrow 3 points to the 'GTF File to convert' section.

Tools

gtf to bed

Convert Formats

- [Convert gffGTF to annotated GTF to BED for SplicingTie results](#)
- [Convert GTF to BED12](#)
- [GTF-to-BEDGraph converter](#)

Workflows

Convert GTF to BED12 (Galaxy Version 1.7)

GTF File to convert

2: hg38.goi.gtf

Advanced options

Use default options

Advanced options for gtfToGenePred.

✓ Execute

Infer experiment strandness: SE

Tools

infer_experiment

RNA Analysis

[Infer Experiment](#) speculates how RNA-seq were configured

[FPKM Count](#) calculates raw read count, FPM, and FPKM for each gene

[BAM to Wiggle](#) converts all types of RNA-seq data from .bam to .wig

[RPKM Count](#) calculates raw count and RPKM values for transcript at exon, intron, and mRNA level

Workflows

- [All workflows](#)

Infer Experiment speculates how RNA-seq were configured (Galaxy Version 2.6.4.1)

Input .bam file

153: HISAT2 on data 94 and data 1: aligned reads (BAM)

(--input-file)

Reference gene model

177: Convert GTF to BED12 on data 2

(--refgene)

Number of reads sampled from SAM/BAM file (default = 200000)

200000

(--sample-size)

Minimum mapping quality

0

Minimum mapping quality for an alignment to be considered as "uniquely mapped" (--mapq)

This is SingleEnd Data

Fraction of reads failed to determine: 0.0000

Fraction of reads explained by "++/--": 0.5033

Fraction of reads explained by "+-/-+": 0.4967

Infer experiment strandness: PE

Tools

infer_experiment

RNA Analysis

Infer Experiment speculates how RNA-seq were configured

FPKM Count calculates raw read count, FPM, and FPKM for each gene

BAM to Wiggle converts all types of RNA-seq data from .bam to .wig

RPKM Count calculates raw count and RPKM values for transcript at exon, intron, and mRNA level

Workflows

- All workflows

Infer Experiment speculates how RNA-seq were configured (Galaxy Version 2.6.4.1)

Input bam file

155: HISAT2 on data 129, data 128, and data 1: aligned reads (BAM)

(--input-file)

Reference gene model

177: Convert GTF to BED12 on data 2

(--refgene)

Number of reads sampled from SAM/BAM file (default = 200000)

200000

(--sample-size)

Minimum mapping quality

0

Minimum mapping quality for an alignment to be considered as "uniquely mapped" (--mapq)

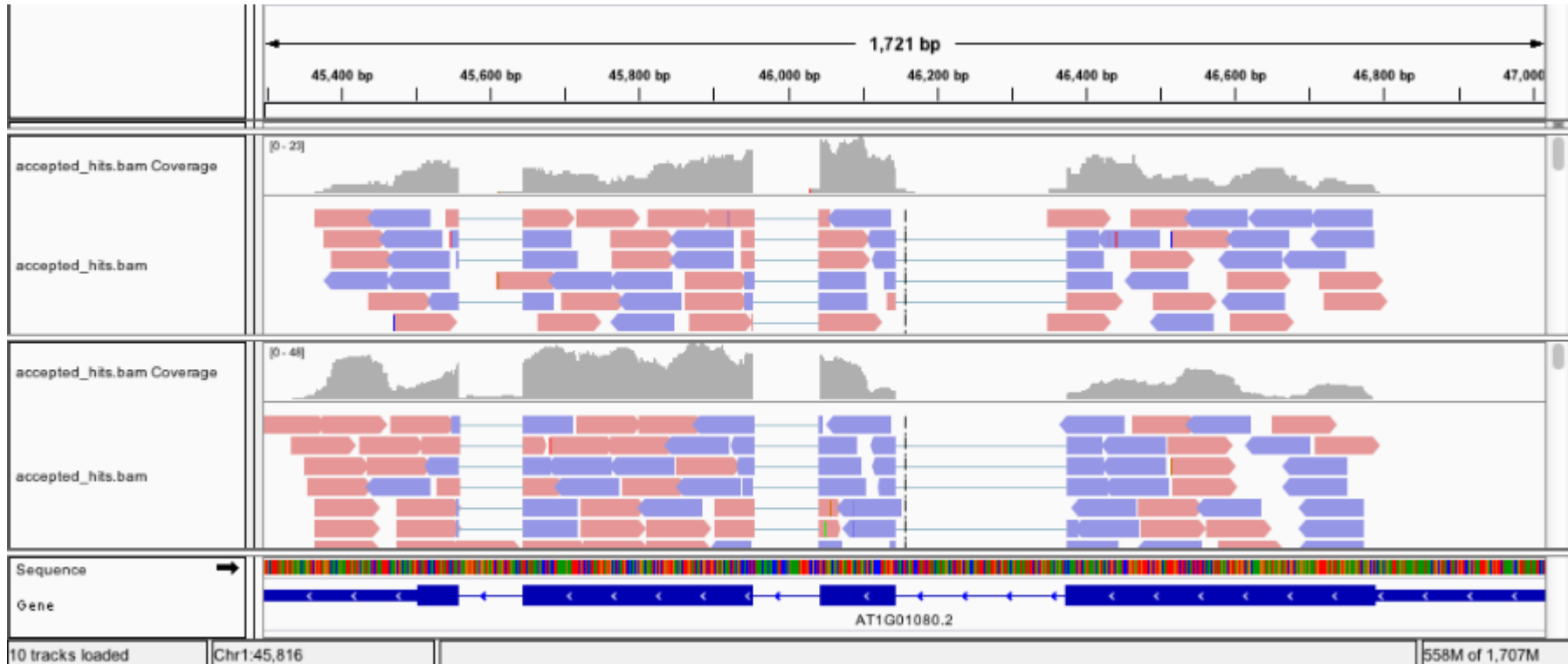
This is PairEnd Data

Fraction of reads failed to determine: 0.0000

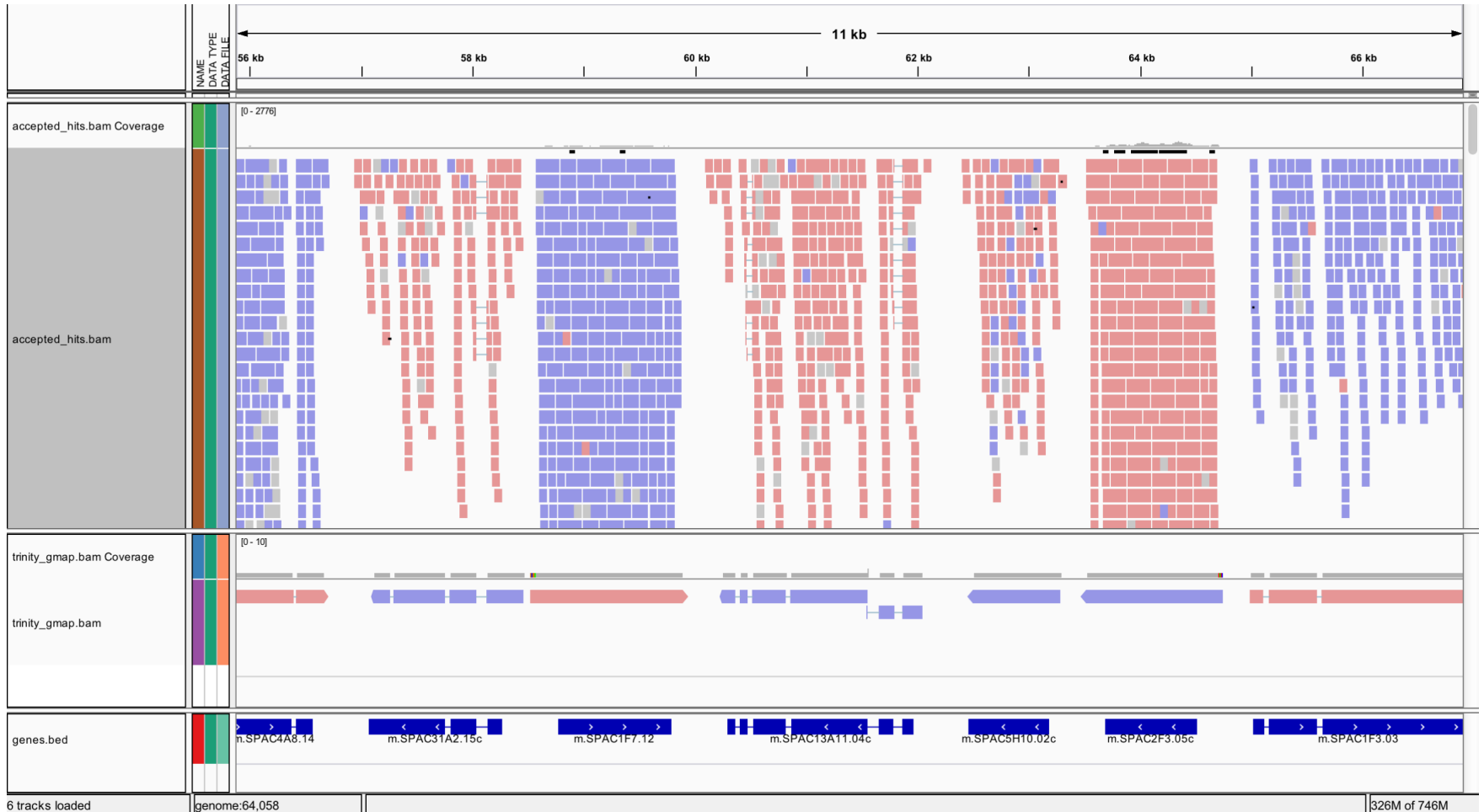
Fraction of reads explained by "1++,1--,2+-,2-+": 0.4941

Fraction of reads explained by "1+-,1-+,2++,2--": 0.5059

Infer experiment strandness



Infer experiment strandness



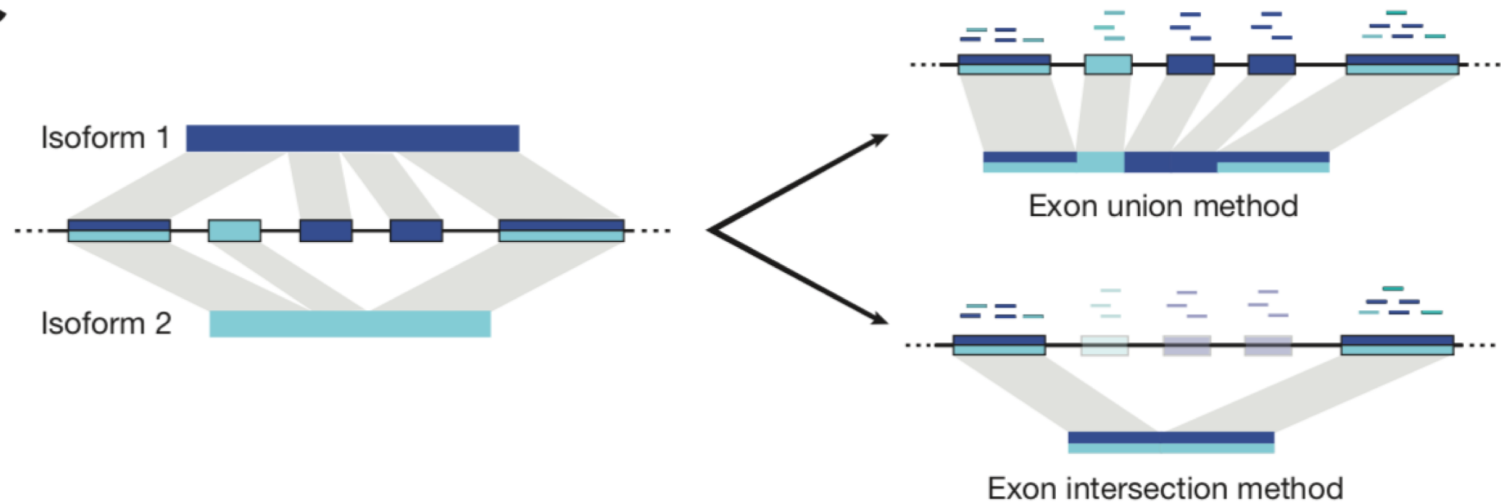
Quantification

- Count uniquely mapped reads only
- What to count?
 - Fragments
 - Reads
- Where to count?
 - Gene body
 - Exons

Quantification

- Count fragments over all exons

c

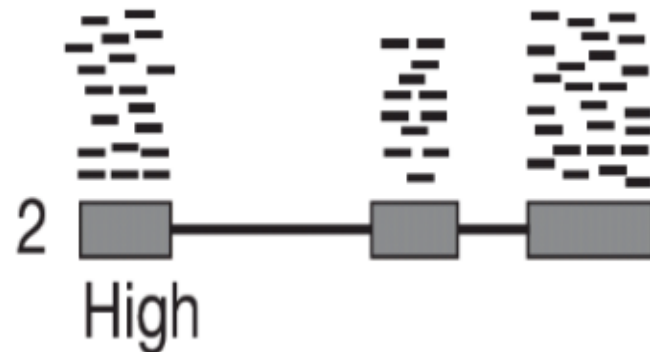


Quantification

- For comparison, normalize counts by
 - Transcript length



- Sequencing depth



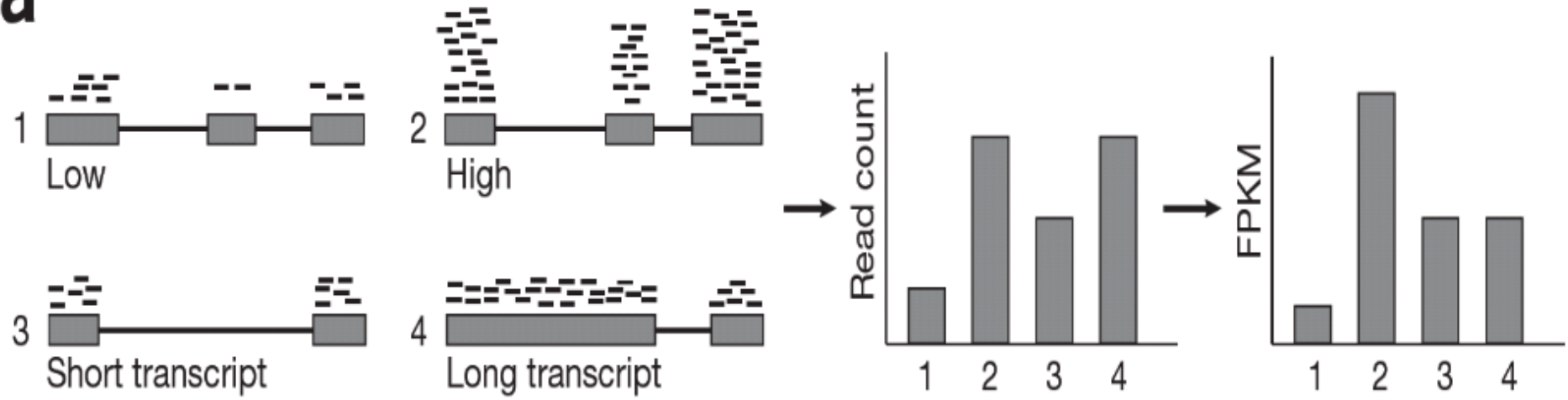
Quantification

- Fragments Per Kilobase Million (FPKM)
 - Total reads of a sample (w), divided by 1Mio
 - Divide the exon read counts (x) by the “per million” scaling factor
 - Divide this values by the length (l) of the gene (sum of exon lengths), devided by 1000 (kilobase scaling)

$$FPKM = 10^9 \times \frac{x}{wl}$$

Quantification

a



Quantification

- Tool box: R Subread - featureCounts
- Requirements:
 - BAM
 - GFF/GTF
 - Sequencing strandness information

Quantification: SE

Tools

featurecounts

Annotation

[goseq](#) tests for overrepresented gene categories

RNA Analysis

[limma](#) Perform differential expression with limma-voom or limma-trend

[edgeR](#) Perform differential expression of count data

[DESeq2](#) Determines differentially expressed features from count tables

[featureCounts](#) Measure gene expression in RNA-Seq experiments from SAM or BAM files.

[Remove Unwanted Variation](#) from RNA-seq data

Workflows

- [All workflows](#)

featureCounts Measure gene expression in RNA-Seq experiments from SAM or BAM files. (Galaxy Version 1.6.3+galaxy2)

Alignment file

157: HISAT2 on data 137, data 136, and data 1: aligned reads (BAM)
156: HISAT2 on data 133, data 132, and data 1: aligned reads (BAM)
155: HISAT2 on data 129, data 128, and data 1: aligned reads (BAM)
154: HISAT2 on data 95 and data 1: aligned reads (BAM)
153: HISAT2 on data 94 and data 1: aligned reads (BAM)

This is a batch mode input field. Separate jobs with a new line.

The input alignment file(s) where the gene expression has to be counted. The files must be in the History, these files must have the extension .bam or .sam

Specify strand information

Unstranded

Indicate if the data is stranded and if strand-specific read counting should be performed. If the data is unstranded, the program will count reads mapped to both strands. If the data is stranded, the program will count reads mapped to the specified strand. If the data is paired-end, the program will count reads mapped to both strands.

Gene annotation file

in your history

Gene annotation file

2: hg38.goi.gtf

The program assumes that the provided annotation file is in GTF format. If the file is in a different format, the program will fail.

Output format

Gene-ID "t" read-count (MultiQC/DESeq2/edgeR/limma-voom compatible)

The output format will be tabular, select the preferred columns here

Count multi-mapping reads/fragments

Disabled; multi-mapping reads are excluded (default)

If specified, multi-mapping reads/fragments will be counted (ie. a multi-mapping read will be counted as many times as it is mapped). The program uses the -M tag to find multi-mapping reads. (-M)

Minimum mapping quality per read

0

The minimum mapping quality score a read must satisfy in order to be counted

Exon-exon junctions

Yes No

If specified, reads supporting each exon-exon junction will be counted (-J)

Long reads

Yes No

If specified, long reads such as Nanopore and PacBio reads will be counted. (-L)

Count reads by read group

Yes No

If specified, reads are counted for each read group separately. The 'RG' tag must be present in the BAM file. (-R)

Largest overlap

Yes No

If specified, overlapping reads (fragments) will be assigned to the target that has the largest overlap. (-O)

Minimum bases of overlap

10

Specify the minimum required number of overlapping bases between a read and a feature. (--minOverlap)

Quantification: PE

Tools

featurecounts

Annotation

- [goseq](#) tests for overrepresented gene categories

RNA Analysis

- [limma](#) Perform differential expression with limma-voom or limma-trend
- [edgeR](#) Perform differential expression of count data
- [DESeq2](#) Determines differentially expressed features from count tables
- [featureCounts](#) Measure gene expression in RNA-Seq experiments from SAM or BAM files.
- [Remove Unwanted Variation](#) from RNA-seq data

Workflows

- [All workflows](#)

featureCounts Measure gene expression in RNA-Seq experiments from SAM or BAM files. (Galaxy Version 1.6.3+galaxy2)

Alignment file

160: HISAT2 on data 149, data 148, and data 1: aligned reads (BAM)
159: HISAT2 on data 145, data 144, and data 1: aligned reads (BAM)
158: HISAT2 on data 141, data 140, and data 1: aligned reads (BAM)
157: HISAT2 on data 137, data 136, and data 1: aligned reads (BAM)
156: HISAT2 on data 133, data 132, and data 1: aligned reads (BAM)

This is a batch mode input field. Separate jobs will be triggered for each file. If you are using a Gene annotation file in the History, these files must have the database file in the same directory.

Specify strand information

Unstranded

Indicate if the data is stranded and if specific read counting should be performed. If specified, reads supporting each strand will be counted separately. (-s)

Gene annotation file

in your history

Gene annotation file

2: hg38.goi.gtf

The program assumes that the provided annotation file is in GTF format. Make sure the file is used for the alignment

Output format

Gene-ID "t" read-count (MultiQC/DESeq2/edgeR/limma-voom compatible)

The output format will be tabular, select the preferred columns here

Create gene-length file

Yes No

Creates a tabular file that contains the effective (nucleotides used for counting reads)

Options for paired-end reads

Count fragments instead of reads

Enabled; fragments (or templates) will be counted instead of reads

If specified, fragments (or templates) will be counted instead of reads. (-p)

Count multi-mapping reads/fragments

Disabled; multi-mapping reads are excluded (default)

If specified, multi-mapping reads/fragments will be counted (ie. a multi-mapping program uses the 'RG' tag to find multi-mapping reads. (-M))

Minimum mapping quality per read

0

The minimum mapping quality score a read must satisfy in order to be counted

Exon-exon junctions

Yes No

If specified, reads supporting each exon-exon junction will be counted (-J)

Long reads

Yes No

If specified, long reads such as Nanopore and PacBio reads will be counted. (-L)

Count reads by read group

Yes No

If specified, reads are counted for each read group separately. The 'RG' tag is required. (-L)

Largest overlap

Yes No

If specified, overlapping (fragments) will be assigned to the target that has the largest overlap. (-l)

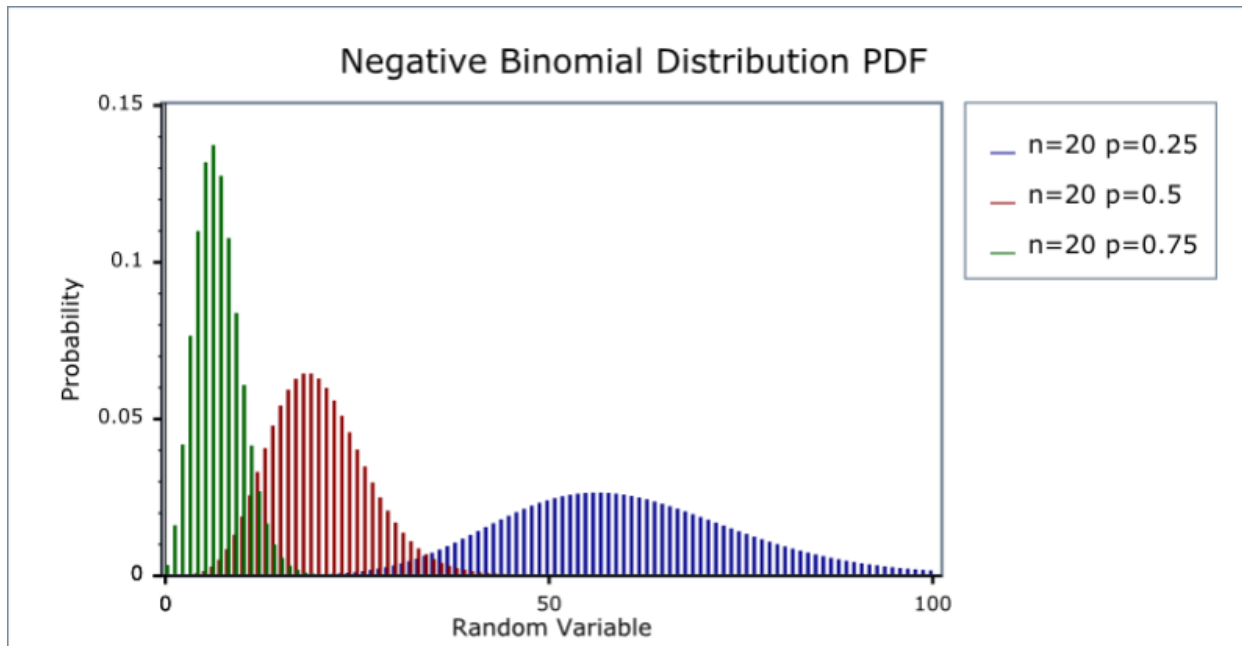
Minimum bases of overlap

10

Specify the minimum required number of overlapping bases between a read and a reference. (--minOverlap)

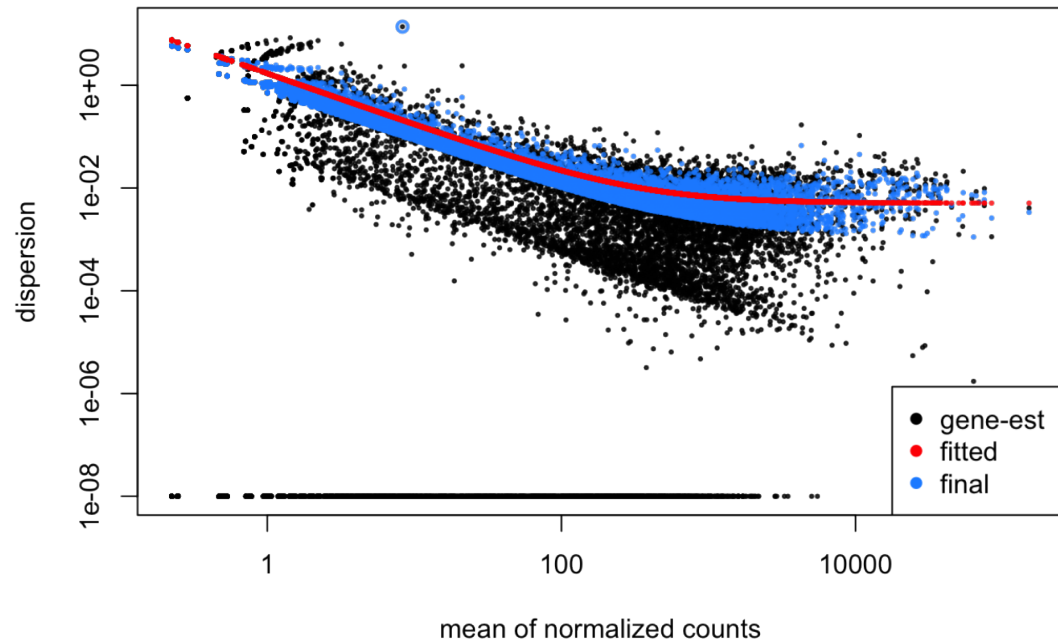
Differential Gene Expression

- Sophisticated read count normalization
 - per gene estimation of dispersion for Negative Binomial read count distribution (NB can model over-dispersion due to biological noise)



Differential Gene Expression

- Tool box: DESeq2
 - Dispersion shrinkage to regression fit
 - Hypothesis test
 - FDR correction: Benjamini Hochberg rank sum test



Differential Gene Expression

The image shows the Galaxy DESeq2 tool interface. On the left is a sidebar with tool categories and a search bar. On the right is the main configuration panel for the DESeq2 tool. Red triangles with numbers 1 through 9 point to specific elements in the interface.

Tools

deseq2

Annotation

- [Annotate DE\(X\)Seq results](#)

RNA Analysis

- [edgeR](#) Perform differential expression of count data
- [DESeq2](#) Determines differentially expressed features from count tables
- [Annotate DESeq2/DEXSeq output tables](#) Append annotation from GTF to differential expression tool outputs
- [featureCounts](#) Measure gene expression in RNA-Seq experiments from SAM or BAM files.
- [Remove Unwanted Variation](#) from RNA-seq data
- [StringTie](#) transcript assembly and quantification

Peak Calling

- [PEAKachu](#) Calls Peaks in CLIP data
- [DiffBind](#) differential binding analysis of ChIP-Seq peak data

Workflows

- [All workflows](#)

DESeq2 Determines differentially expressed features from count tables (Galaxy Version 2.11.40.6)

how

Select datasets per level

Factor

1: Factor

Specify a factor name, [selects_drug_x](#) or [cancer_markers](#)

placenta

Only letters, numbers and underscores will be retained in this field

Factor level

1: Factor level

Specify a factor level, [values could be 'tumor', 'normal', 'treated' or 'control'](#)

3t

Only letters, numbers and underscores will be retained in this field

Counts file(s)

- 188: featureCounts on data 2 and data 157: Counts
- 187: featureCounts on data 2 and data 156: Summary
- 186: featureCounts on data 2 and data 156: Counts
- 185: featureCounts on data 2 and data 155: Summary
- 184: featureCounts on data 2 and data 155: Counts

2: Factor level

Specify a factor level, [values could be 'tumor', 'normal', 'treated' or 'control'](#)

1t

Only letters, numbers and underscores will be retained in this field

Counts file(s)

- 182: featureCounts on data 2 and data 154: Counts
- 181: featureCounts on data 2 and data 153: Summary
- 180: featureCounts on data 2 and data 153: Counts
- 177: Convert GTF to BED12 on data 2
- 3: hg38.promoter.bed

(Optional) provide a tabular file with additional batch factors to include in the model.

Annotations:

- 1: Search bar
- 2: DESeq2 tool description
- 3: Factor name input
- 4: Factor level input
- 5: Factor level input
- 6: Counts file(s) input
- 7: Factor level input
- 8: Factor level input
- 9: Counts file(s) input

Differential Gene Expression

The image shows a screenshot of the Galaxy web interface. On the left is a 'Tools' panel, and on the right is the configuration page for the 'Cut columns from a table (cut)' tool. Red arrows with numbers 1 through 6 point to specific elements in the interface.

Tools Panel (Left):

- 1** points to the 'Tools' header.
- 2** points to the 'Cut columns from a table (cut)' tool link under the 'Text Manipulation' section.

Tool Configuration Page (Right):

- 3** points to the tool title 'Cut columns from a table (cut) (Galaxy Version 1.1.0)'.
- 4** points to the 'File to cut' input field, which contains the text '247: Compute on data 246'.
- 5** points to the 'Operation' dropdown menu, which is set to 'Keep'.
- 6** points to the 'Cut by' input field, which contains the text 'fields'.
- The 'List of Fields' section shows a 'Select/Unselect all' checkbox and two selected fields: 'Column: 1' and 'Column: 3'.
- At the bottom is an 'Execute' button.