Reference-based RNA-Seq data analysis

Introduction & outline

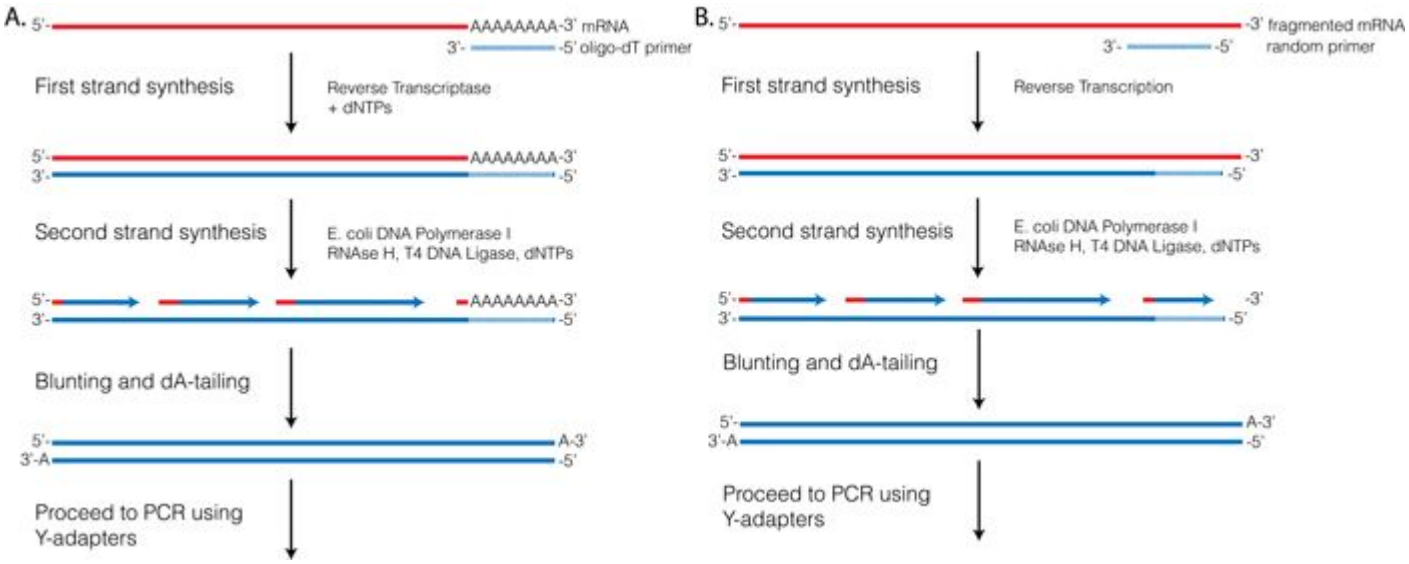
https://galaxyproject.github.io/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html

RNAseq librairies

1. cDNA synthesis

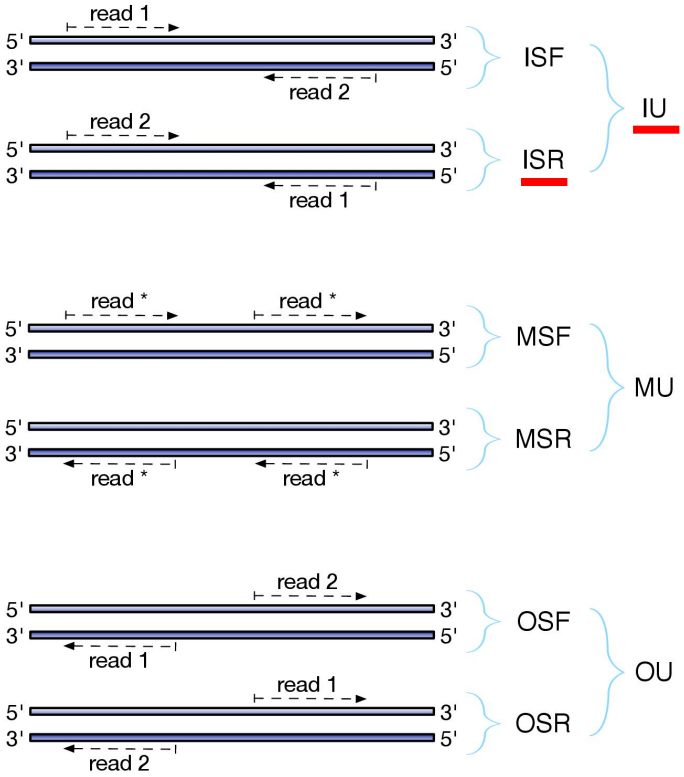
Oligo-dT

random priming



RNAseq librairies

2. Inserts and sequencing strategies



Paired end sequencing “Inward”
(stranded or unstranded)

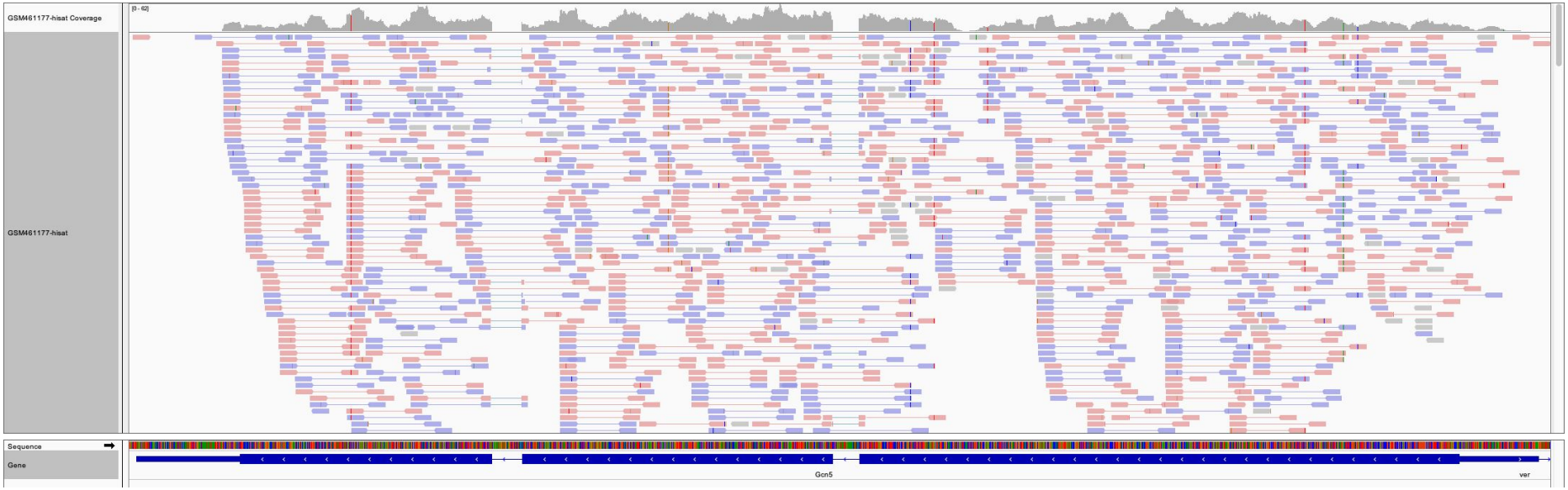
Single read sequencing
(stranded or unstranded)

Paired end sequencing “Outward”
(stranded or unstranded)

in practice, with Illumina **paired-end** RNAseq protocols you will either deal with:

- Unstranded RNAseq data
  (IU type from above. Also called fr-unstranded in TopHat/Cufflinks jargon)
- Stranded RNAseq data produced with Illumina TrueSeq RNAseq kits
  (ISR type from above or fr-firststrand in TopHat/Cufflinks nomenclature).

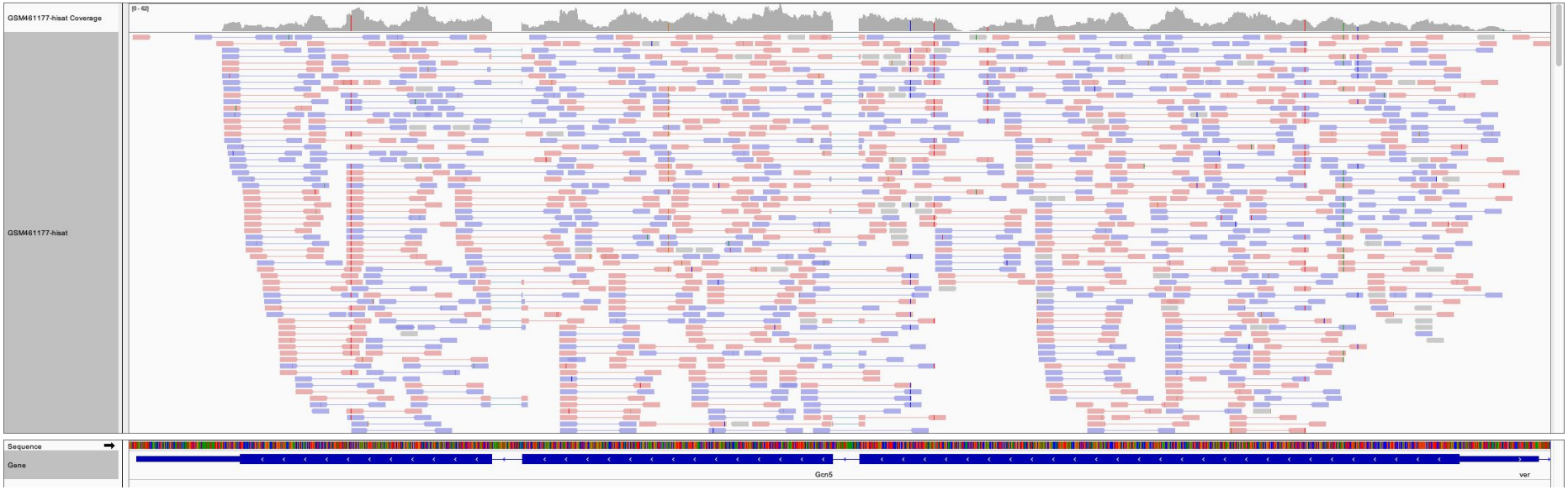
Reference-base Expression analysis: the key idea



- Map reads to a reference genome with aligners TopHat, TopHat2, HiSat, HiSat2 (bowtie based) or STAR
  → **These aligners are "split aware"**
- Use a read counting software and annotation information (GTF, GFF3, BED, …) to count the read spanning a gene / transcript

Read counts are proxies to RNA steady state levels

Focus on quality control & "filtering"

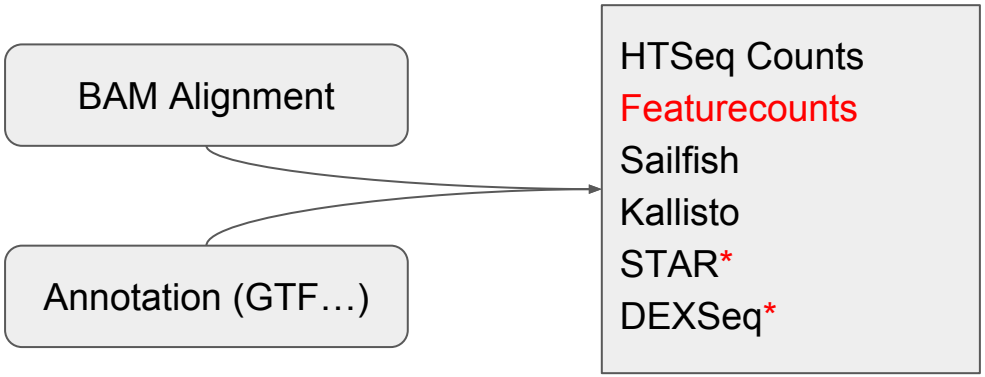


- It is tempting to filter the data to get "good counts" (low quality alignments and PCR duplicates)
- But reflect on the specific purpose of expression analysis with this in mind...

Read counts are proxies to RNA steady state levels

... Then revisit the question what is a "good" read in RNAseq-based expression analysis ?

Transcript Quantification



Counting software

Geneid	s1	s2
FBgn0010247	671	584
FBgn0086378	842	1257
FBgn0263977	1049	1526
FBgn0069923	23	85
FBgn0039955	866	745
FBgn0259821	162	259
FBgn0027341	9	0
…….	.	

read counts by gene

Note that we use absolute read counts because we are going to compare counts across samples.

Other metrics for comparison of genes within the same sample are:

- CPM (Counts Per Million) Each gene count is divided by the corresponding library size (in millions).
- RPKM (reads per kilobase of exons per million mapped reads)
- TPM (Transcript per Million)

1. For each gene, count the number of reads mapping to it and divide by its length in base pairs (=counts per base).
2. Multiply that value by 1 divided by the sum of all counts per base of every gene.
3. Multiply that number by 106

.

Statistical Analysis of Differential expression

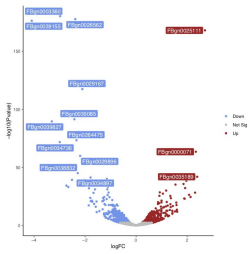
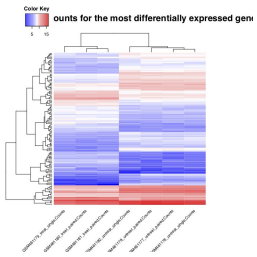
Geneid	s1	s2	s3	s4
FBgn0010247	671	584	842	1257
FBgn0086378	842	1257	23	85
FBgn0263977	1049	1526	1049	1526
FBgn0069923	23	85	866	745
FBgn0039955	866	745	162	259
FBgn0259821	162	259	321	150
FBgn0027341	9	0	1	15
.....	.			

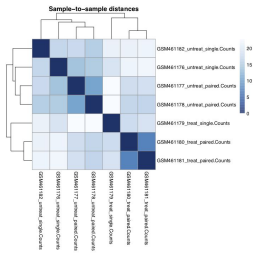
biological condition → other factor ↓	control	treatment
male	s1	s3
female	s2	s4

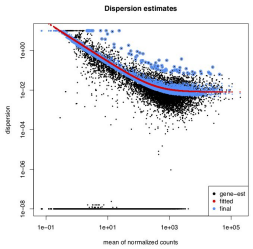
analysis plan = factor table

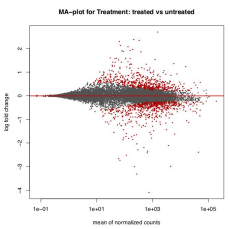
DESeq
edgeR
DEXseq
...

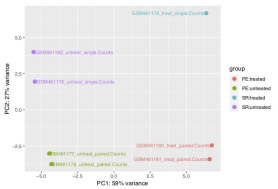
1.GeneID	2.Base mean	3.log2(FC)	4.StdErr	5.Wald-Stats	6.P-value	7.P-adj
FBgn0003360	4524.12972051454	-2.97845052578473	0.103851387878428	-28.6799299136129	6.79040717620437e-181	5.85944235234675e-177
FBgn0026562	45571.1817907476	-2.38486741453747	0.0837445919075783	-28.4778677669054	2.20234439128092e-178	9.50201487618153e-175
FBgn0039155	757.110812249869	-4.08954773698851	0.144182147601647	-28.3637593489543	5.66324182621554e-177	1.62893712394713e-173
FBgn0025111	1561.55924278167	2.70145590945079	0.0975860738085875	27.6828014901965	1.1248425651011e-168	2.42656662356435e-165
FBgn0029167	3770.5285746192	-2.10660983390535	0.0911841682723223	-23.1028025349087	4.33919274593451e-118	7.48857884093378e-115

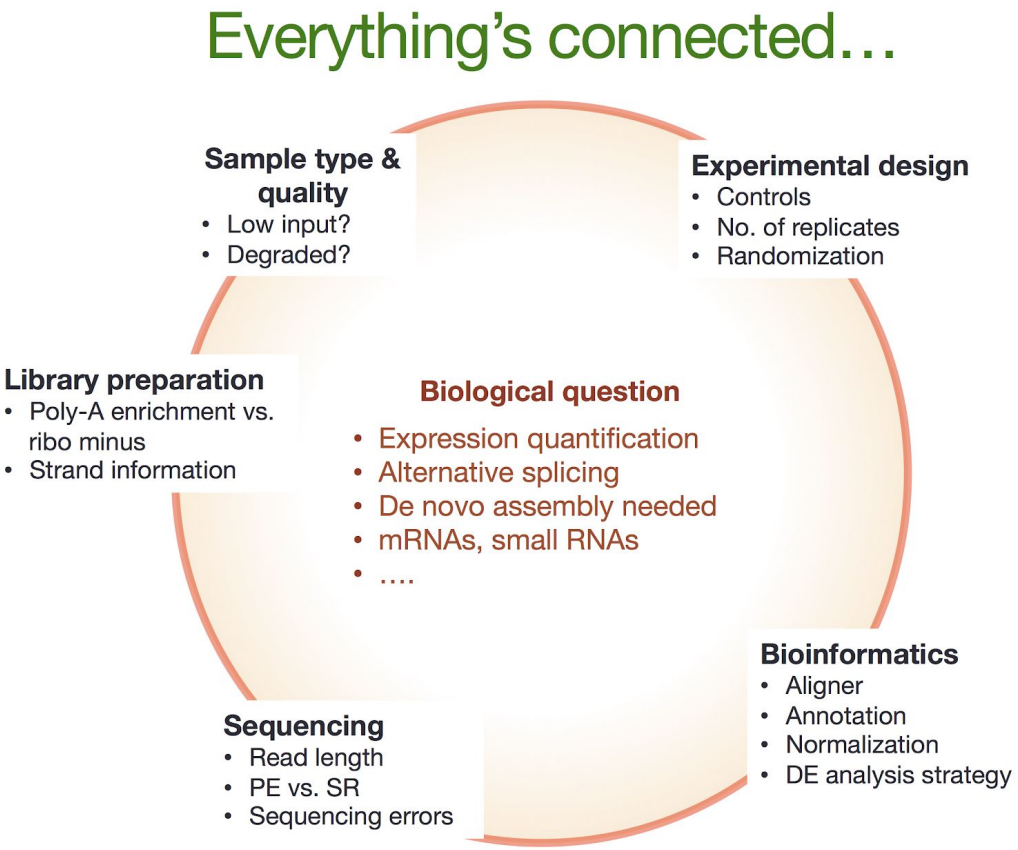








Experimental procedures affect downstream analyses



Additional Material:

How to run your own Galaxy server ? see the GitHub repository Run-Galaxy and the attached tutorial

You will learn how to deploy Galaxy on your own machine (either you labtop or virtual machine in a cloud).

GalaxyKickStart is a software that deploy your Galaxy server using ansible.
We also provide a GalaxyKickStart Docker Image at Dockerhub