

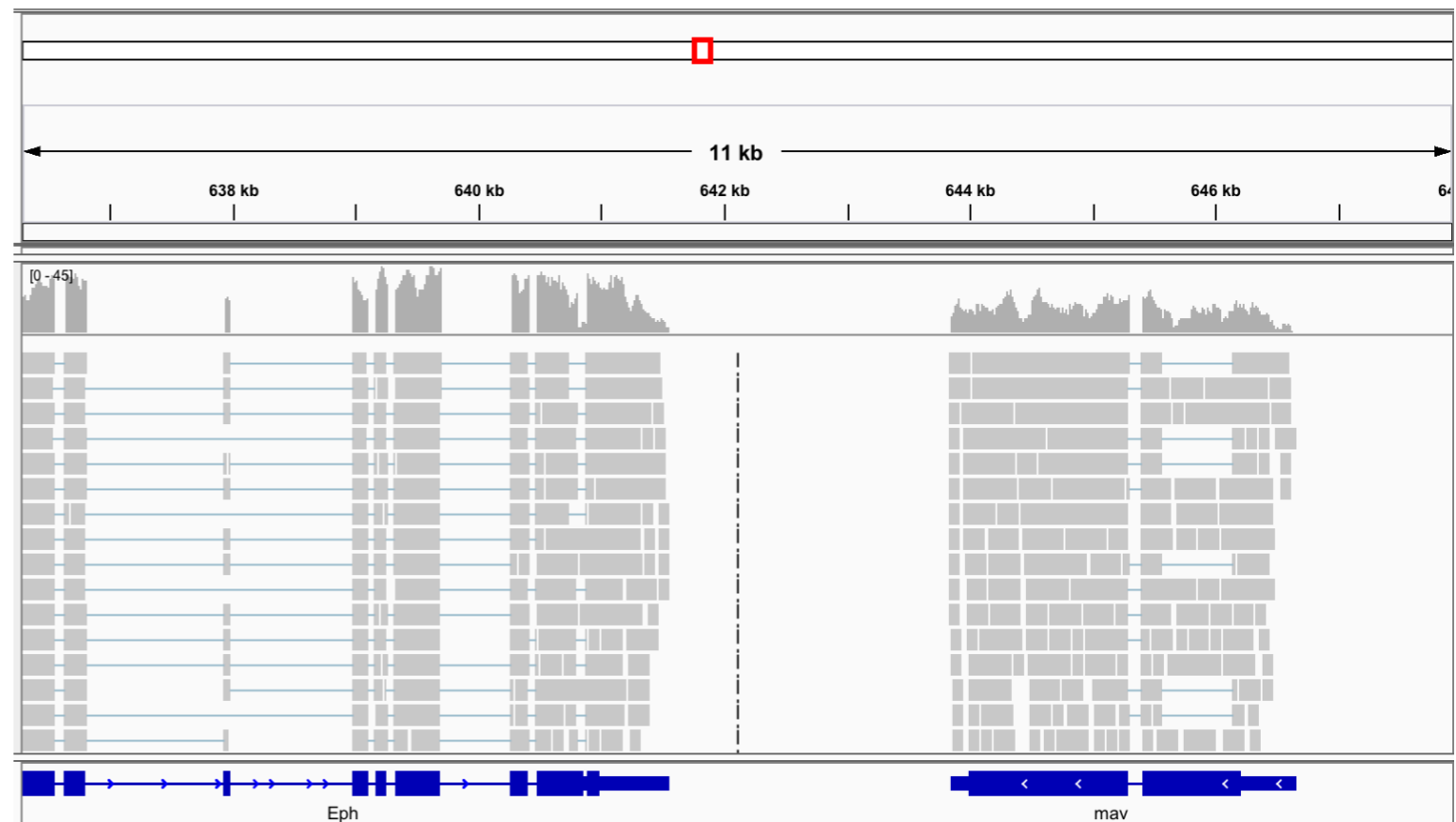
Expression Quantification and Differential Expression Analysis

How do we analyze RNA-Seq data?

- **STEP 1: EVALUATE AND MANIPULATE RAW DATA**
- **STEP 2: MAP TO REFERENCE, ASSESS RESULTS**
- **STEP 3: ASSEMBLE TRANSCRIPTS** Optional
- **STEP 4: QUANTIFY EXPRESSION**
- **STEP 5: TEST FOR DIFFERENTIAL EXPRESSION**
- **STEP 6: VISUALIZE AND PERFORM OTHER DOWNSTREAM ANALYSIS**

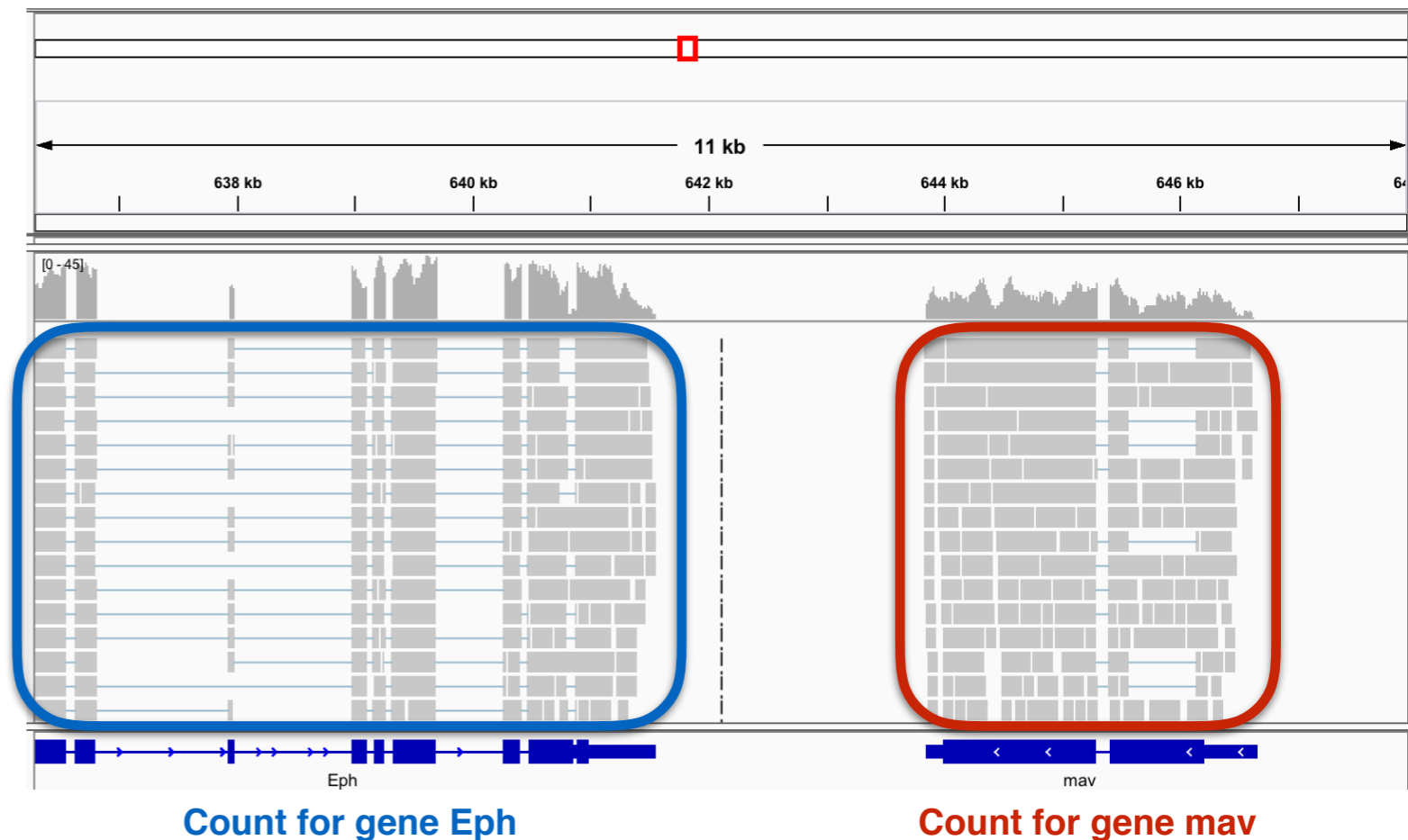
STEP 4- Quantify Expression

- Mapping tells us where every read came from.
- How do we go from that to gene expression?
 - What genes are expressed?
 - What is the expression level for each gene/gene isoform?



STEP 4- Quantify Expression

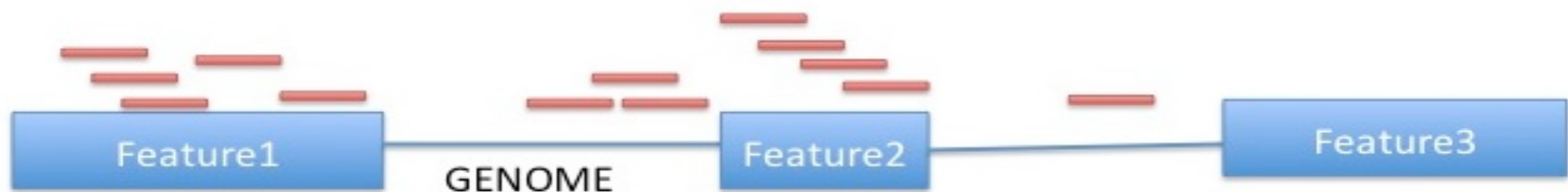
- What is gene expression?
 - A gene is expressed when it's corresponding DNA sequence is transcribed into mRNA (for translation into protein).
- What is gene expression level?
 - The amount of mRNA detected in a sample.



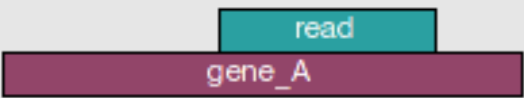
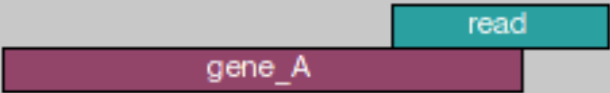



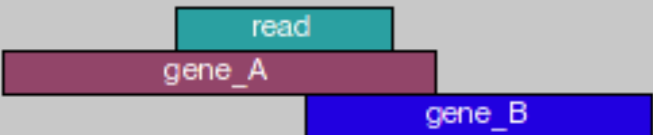
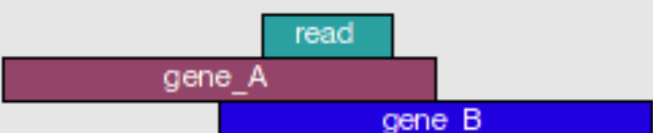
- **Read depth= mRNA amount= expression level of gene**

STEP 4: Quantify Expression

- Bedtools
 - **Bedtools multicov** : Takes a feature file (GFF/GTF) and counts how many reads in the mapped output file (BAM) overlap the features.
 - Remember that the chromosome names in your gff file should match the chromosome names in the reference fasta file used in the mapping step.



STEP 4 : Quantify Expression

	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 –

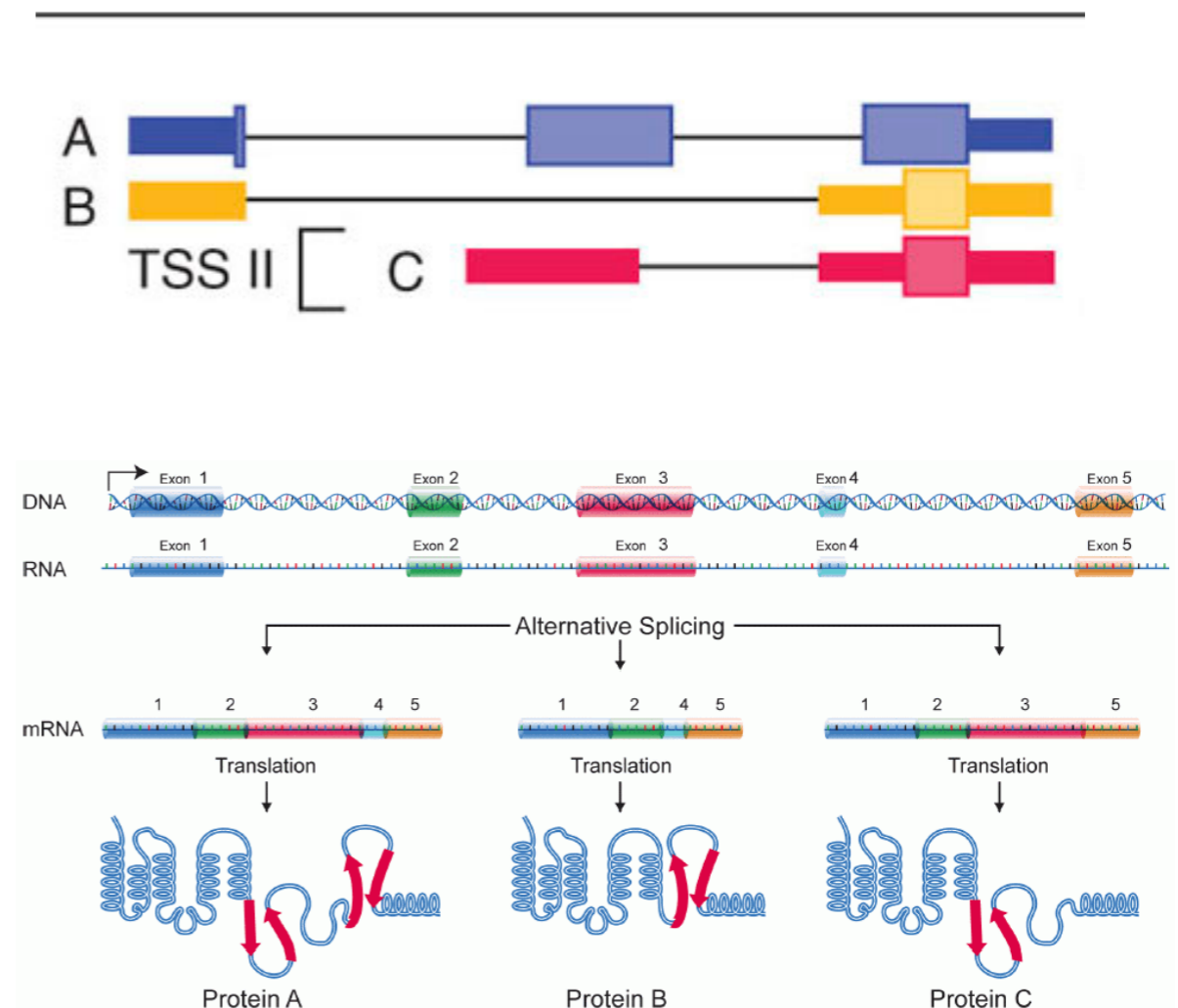
– Gives you fine grained control over how to count genes, especially when a read overlaps more than one gene/feature.

STEP 4- Quantify Expression

- Quantifying a gene is simpler than quantifying its different isoforms/transcripts.
- Tools: kallisto, stringtie, and cufflinks

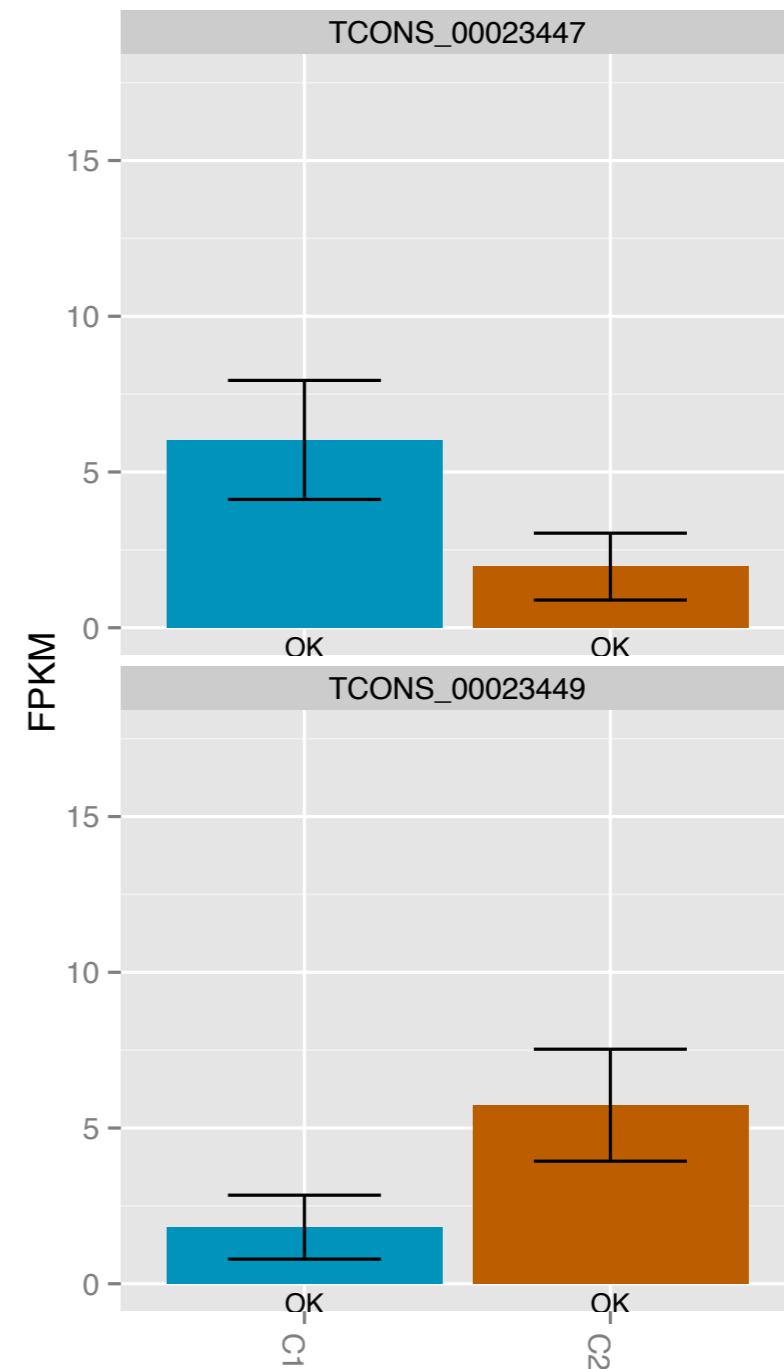
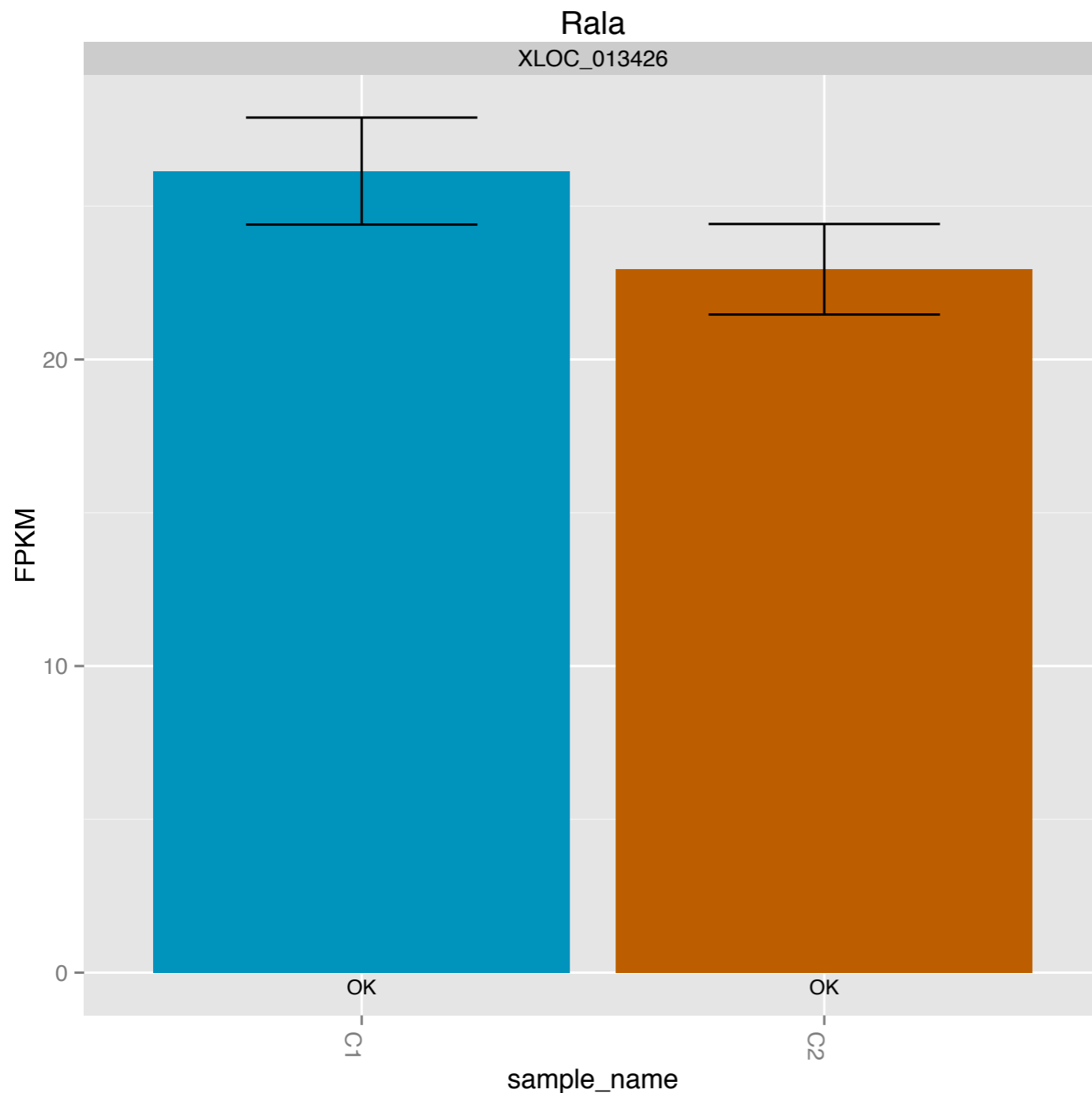
What is a gene? What is a transcript?

A gene can have multiple transcripts!



STEP 4- Quantify Expression

Why quantifying all transcripts of the gene may be important?



STEP 5- Test for Differential Expression

- After mapping and quantifying the genes for each sample:
 - compare gene counts across samples/conditions.
- But first, **normalize!**
 - Normalization evens out the technical variations so that any variation you see between samples is “hopefully” due to real biological reasons.
 - Normalize for **read depth** differences
 - Normalize for **gene/transcript length** differences
 - RPKM = Reads Per **Kilobase of transcript** per **Million mapped reads**
 - Other normalization methods: upper quartile, median read count and more complicated scaling factors (DESeq2 R package)

Gene	Read Count
ABAR1	1200
ATXN1	1345
ATXN2	2
BRAT2	0
GABA	24
GABRA2	456
GABRA4	45345

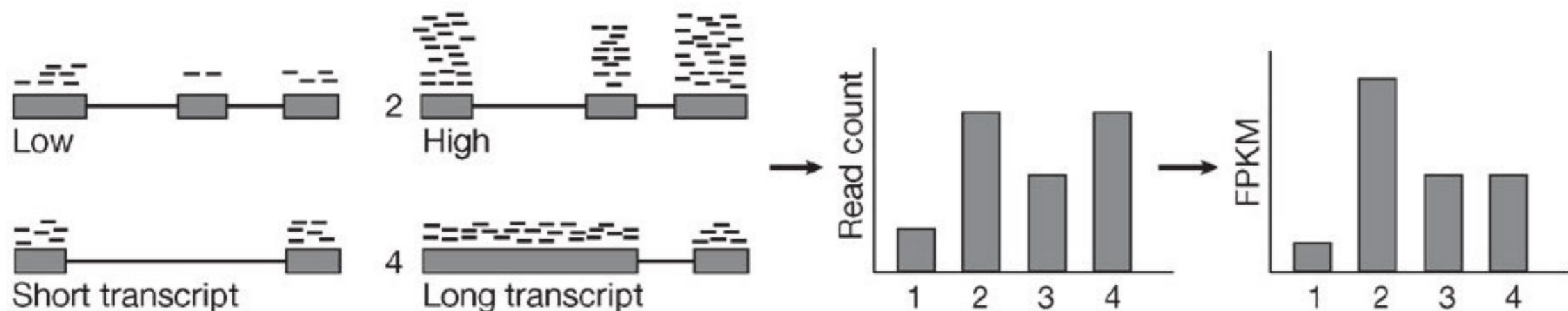


Figure: doi:10.1038/nmeth.1613

STEP 5- Test for Differential Expression

- Input: Gene Expression Matrix

Gene		Sample											
Ensembl	Gene.Name	T1	T2	T3	T4	T5	WT1	WT2	WT3	WT4	WT5	WT6	
ENSMUSG00000000134	Tfe3	312	295	333	258	392	257	344	223	423	277	389	
ENSMUSG00000000142	Axin2	165	171	138	166	203	170	172	119	203	147	178	
ENSMUSG00000000148	Brat1	213	196	207	224	350	204	268	143	300	177	288	
ENSMUSG00000000149	Gna12	684	684	613	545	900	496	672	426	1023	583	797	
ENSMUSG00000000154	Slc22a18	3	2	3	2	2	3	3	2	1	1	3	
ENSMUSG00000000157	Itgb2l	0	0	0	0	0	0	0	0	0	0	0	
ENSMUSG00000000159	Igsf5	0	0	0	0	0	0	0	0	0	0	0	
ENSMUSG00000000167	Pih1d2	15	19	6	10	9	5	5	5	7	6	6	
ENSMUSG00000000168	Dlat	899	777	967	756	1116	777	1047	614	1155	894	1126	
ENSMUSG00000000171	Sdhd	1055	1003	1047	914	1430	939	1192	766	1390	916	1412	
ENSMUSG00000000182	Fgf23	1	0	3	1	0	2	0	2	2	0	0	
ENSMUSG00000000183	Fgf6	0	0	0	0	0	0	0	1	0	0	0	
ENSMUSG00000000184	Ccnd2	1961	1978	1804	1779	2090	1655	2148	1585	2504	1895	2274	
ENSMUSG00000000194	Gpr107	784	733	667	615	889	654	818	483	1034	627	1015	
ENSMUSG00000000197	Nalcn	1120	1009	1047	917	1356	1129	1202	758	1625	1127	1044	

Image from babelomics

- Outputs like:

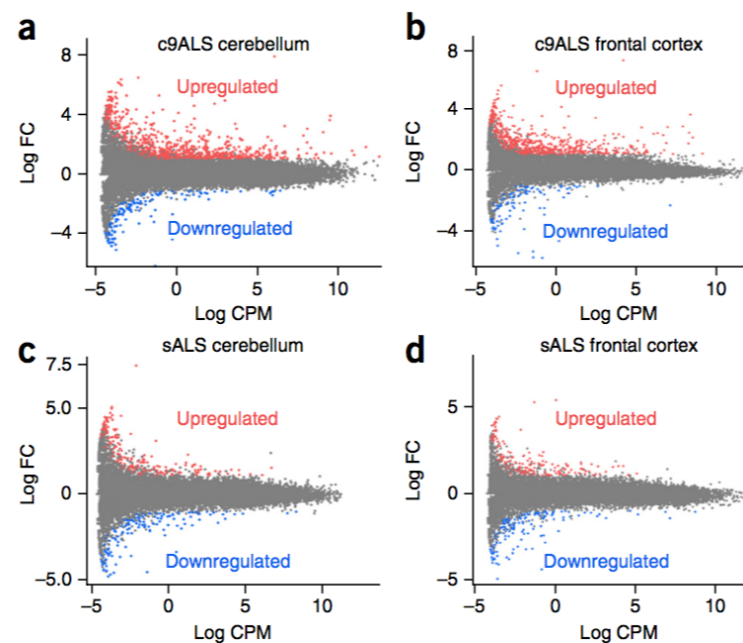
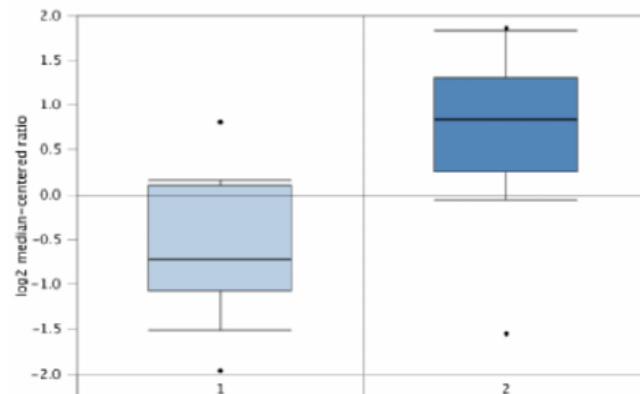


Figure: doi:10.1038/nn.4065