In last week’s lecture, I introduced an example study that we will be using as a case study to discuss RNA-seq analysis. The example study, which is described in the paper, “**An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex**” has the following major goals and steps:

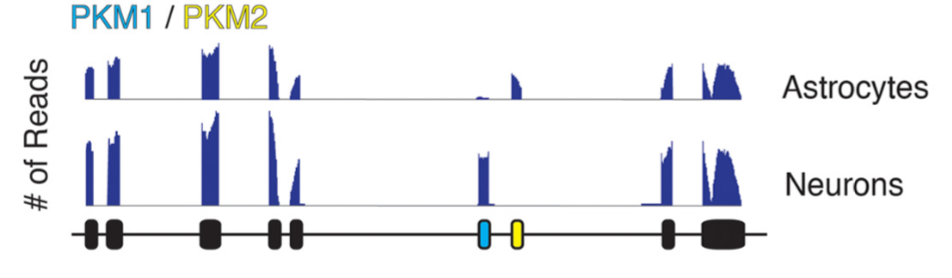
* Goal 1: Figure out the full transcriptomes of individual cell types in the brain.
  + Step 1: Collect samples of individual cell types
  + Step 2: Sequence them
  + Step 3: Put sequencing reads in context of genome and transcripts
  + Step 4: Quantify transcripts
* Goal 2: Compare transcriptomes to each other to see what can be learned.
  + Step 5: Differential Expression Analysis
  + Step 6: Gene Set Enrichment Analysis

After introducing the study, its goals and major steps, we then discussed general approaches and considerations that one might apply for the first three steps depending on the goals of our own particular study. This week, we will continue using “**A survey of best practices for RNA-seq data analysis**” and the case study, as well as my own commentary to explore the general approaches and considerations that apply to the final two steps above.

**STEP 4: QUANTIFY TRANSCRIPTS**

As we discussed in last week's lecture, once sequencing reads are put within the context of the reference by being mapped to either the genome or to the transcriptome, the next step is usually to count the number of reads that aligned. This is simple when there is only one splicing isoform, but becomes much more complex when there are multiple isoforms.

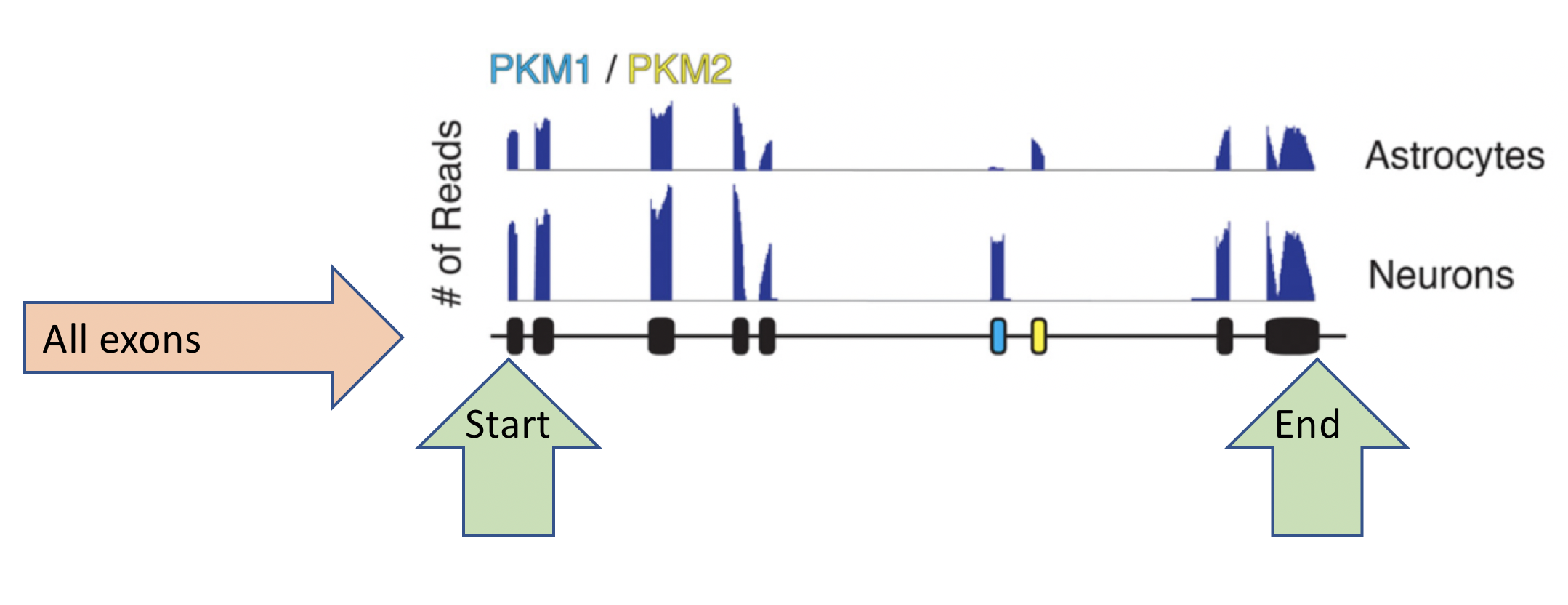
For example, the picture below (taken from “**An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex**”) shows two different isoforms of the PKM gene. In this case, one isoform, PKM1, is primarily expressed in astrocytes while PKM2 is primarily expressed in neurons. PKM1 has an exon that is not included in PKM2 (colored blue below) and PKM2 has an exon that is not included in PKM1 (colored yellow below).



**Quantification using genomic coordinates**

For genomic alignments, the most common practice is for researchers to identify regions of the genome based on annotated transcript isoforms and then to count the number of sequencing reads that align between the coordinates. The simplest way is to count all of the sequencing reads that align between the beginning and end of the full gene. The gene coordinates are taken from the annotations. The starting coordinate for the gene is based on the first annotated exon and the end coordinate for the gene is based on the last annotated exon, even if the exons are from different transcript isoforms.

For example, the start and end of PKM would be like this:



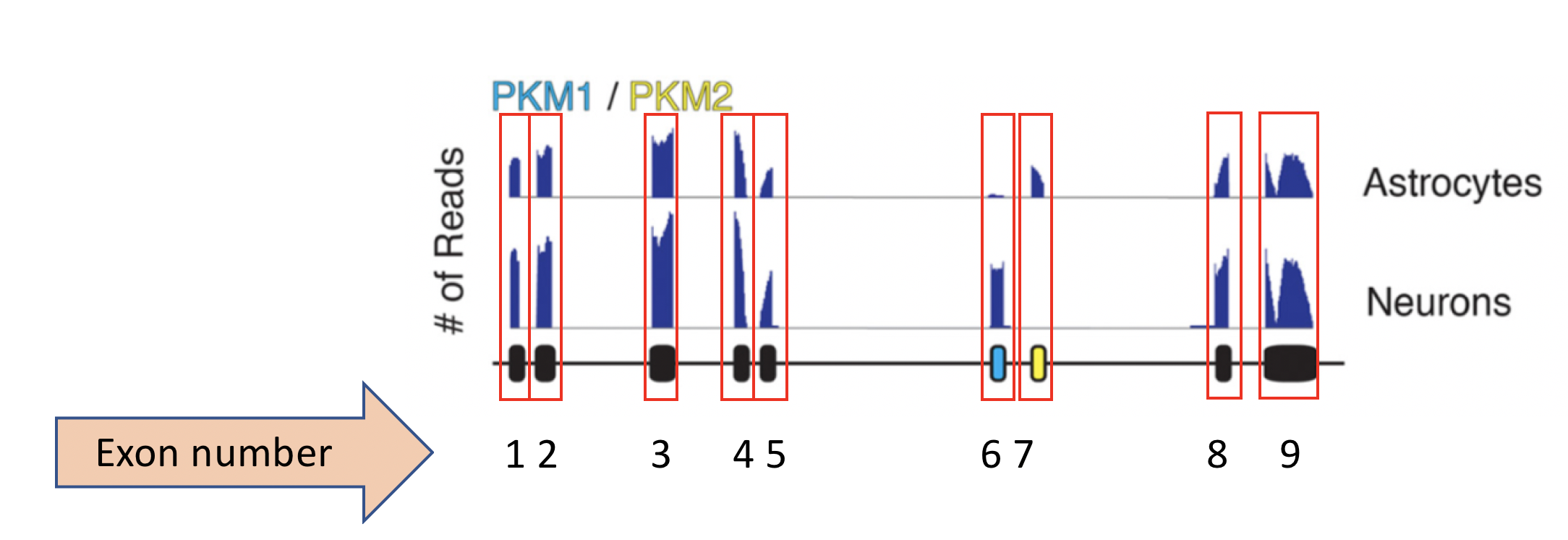
To quantify at the whole-gene level, one would then add up all of the sequencing reads (represented by the vertical blue lines along the genomic coordinates) that aligned anywhere between the start and the end in each sample. The benefit of this method is that is conceptually very simple and can reproducibly tell us whether PKM is more expressed in one sample than another.

**Question 1**: Based on your estimation of the number of sequencing reads (the number and height of the blue bars), is PYM likely to differentially expressed between astrocytes and neurons?

Yes, because there are two distinctly different peaks between PKM1 and PKM2

The drawback of this method is that we can’t tell anything about which isoform is being expressed. Also, if there were an unannotated isoform (which would look like there were sequencing reads aligning to the introns) we would add those reads to the total PKM count and would not identify the fact that there is an unannotated isoform. Even without trying to reconstruct our sequencing reads into transcript isoforms, we can answer these questions in part by simply counting exons and/or introns.

For example, if we count each exon independently as in the picture below, we can get a picture of whether each exon is differentially expressed between the samples. Then, we can infer what is happening at least in general terms by interpreting the results.



**Question 2**: Based on your estimation of the blue bars indicating the number of reads, which of two exons are most differentially expressed between the two samples?

Exons 6 and 7 seemed to be the most differentially expressed, as they are the ones most clearly different between PKM1 and PKM2.

**Question 3**: Based on your estimation of the blue bars indicating the number of reads, which of two exons are least differentially expressed between the two samples?

Exons 2 and 3 appear to be some of the most similar. Though it is worth noting that all of the exons, bar 6 and 7, are relatively similar between PKM1 and PKM2.

**Question 4**: What general conclusion can you make based on the totals for the individual exons about which splicing isoforms are expressed in the two samples? Explain your answer.

The PKM1 isoform is expressed in Neurons, while the PKM2 isoform is expressed in Astrocytes. These isoforms differ based on the presence and absence of exons 6 and 7. PKM1 has exon 6, but lacks exon 7. PKM2 has exon 7 but lacks exon 6.

This method gives us more detailed information about what is happening transcriptionally than the whole-gene counting method and still has the benefit of being conceptually and computationally simple. Because of the simplicity of the method, it is much more reproducible and less likely to have strange biases between the samples than transcript reconstruction methods.

The drawback of this method is that we need to account for the number of tests that are being performed. Instead of performing one test to see if PYM is differentially expressed we are performing nine separate tests. This leads to large multiple testing penalties if you are looking at every exon in the genome independently.

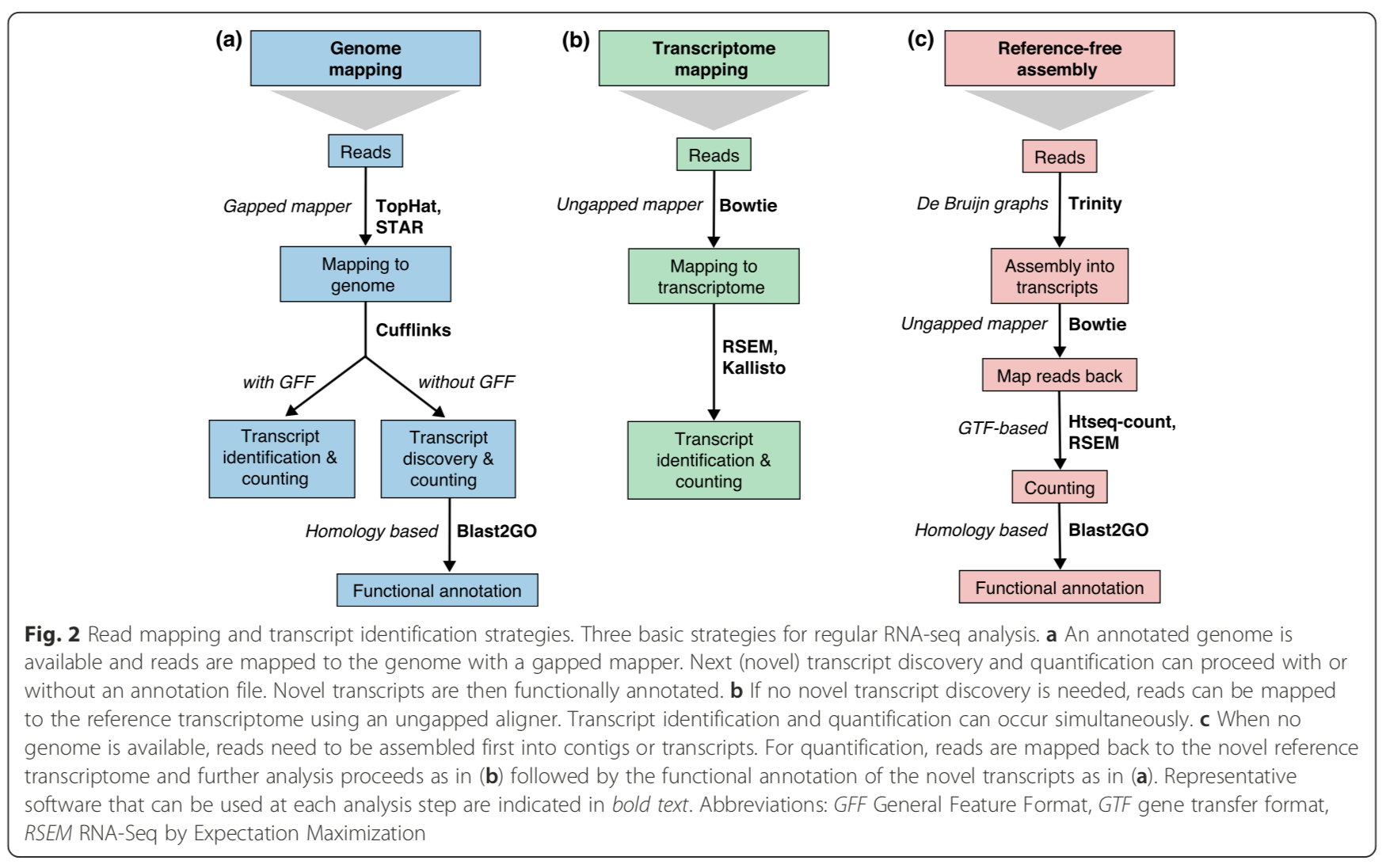
**Question 5**: Based on what you know about multiple testing correction, what does the increased number of tests mean in terms of our results? In other words, as we perform and correct for more tests, can we still detect the same differences, or do the differences in expression for each exon need to be more significant in order to pass the multiple testing correction?

More tests means a higher chance of error occurring just by chance. This can be adjusted for, but also typically results in more stringent standards for significance. As a result, the more tests that are done, the stronger detectable differences will need to be in order to remain significant.

**Quantification using transcript abundance estimation**

In contrast to methods that quantify transcription based purely on genomic coordinates, there are many different methods that estimate the abundance of full-length transcript isoforms.

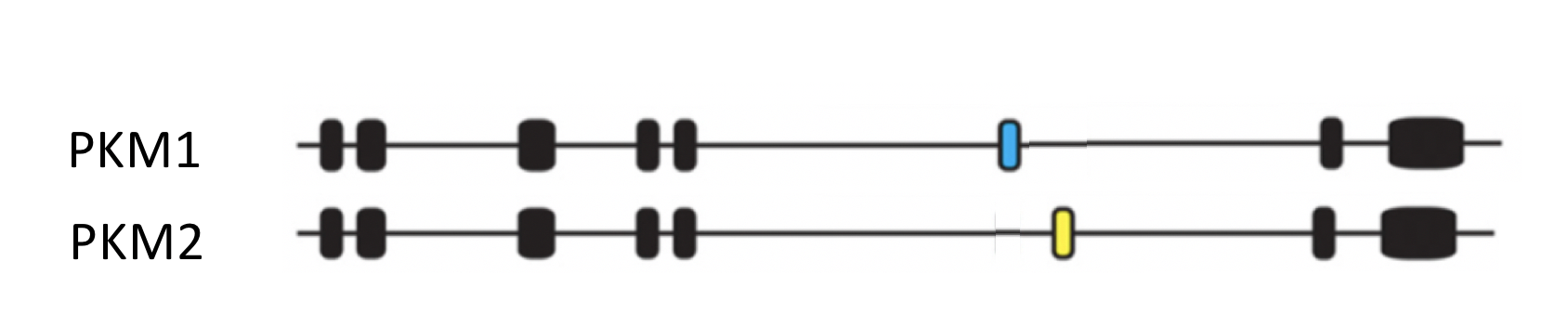
Transcript abundance estimation methods can start with a genome alignment followed by estimation of isoforms are supplied via an annotation file (Gene Feature Format, GFF) as in figure (**2a**) below. Alternatively, the transcript abundance might be estimated directly based on mapping of the short reads to the transcript isoforms supplied in the GFF instead of to a full genome as in figure (**2b**) below:



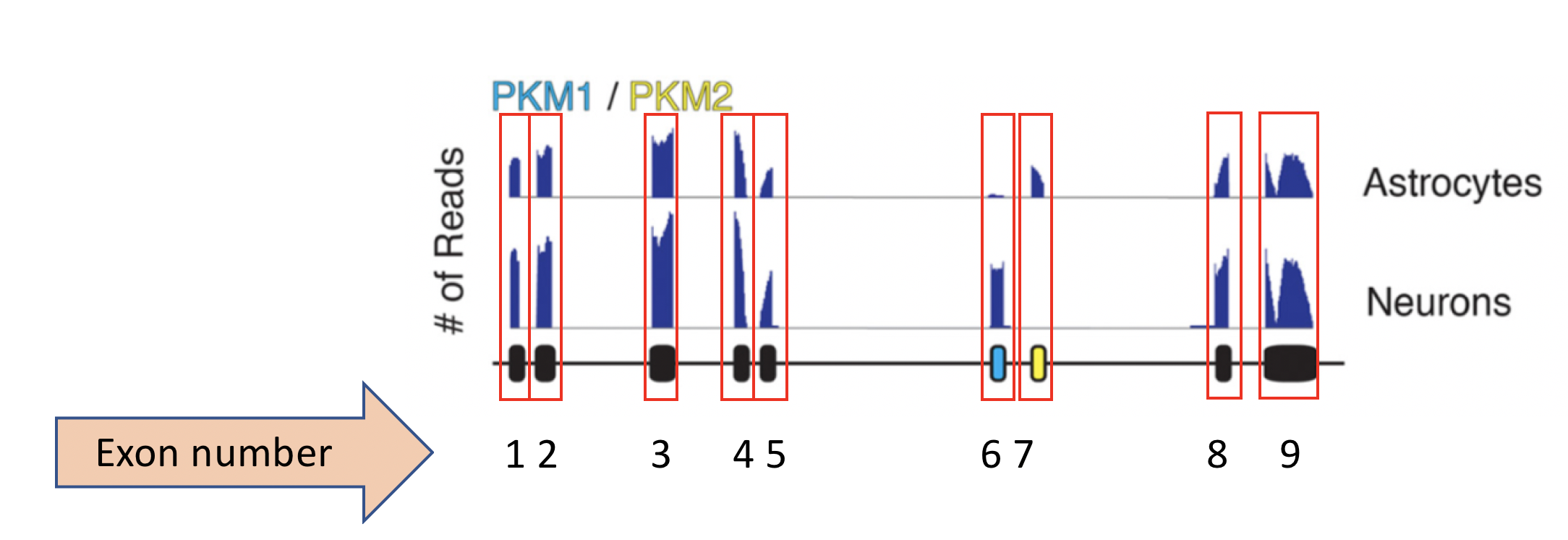
Credit: Conesa et al. Genome Biology (2016) 17:13

Transcript abundance estimation methods suffer from the same multiple testing problem as the exon-counting example above – each transcript isoform counts as a separate test. However, the other (bigger) problem with this method is that it can be very difficult to figure out which transcript isoform a read might have come from.

Let’s consider the PYM gene from the examples above. If we were to try to estimate the abundance of PYM1 relative to PYM2, it might look something like this:



Using the same data from the exon counting example, we can start to see why it is so difficult to differentiate between the isoforms.

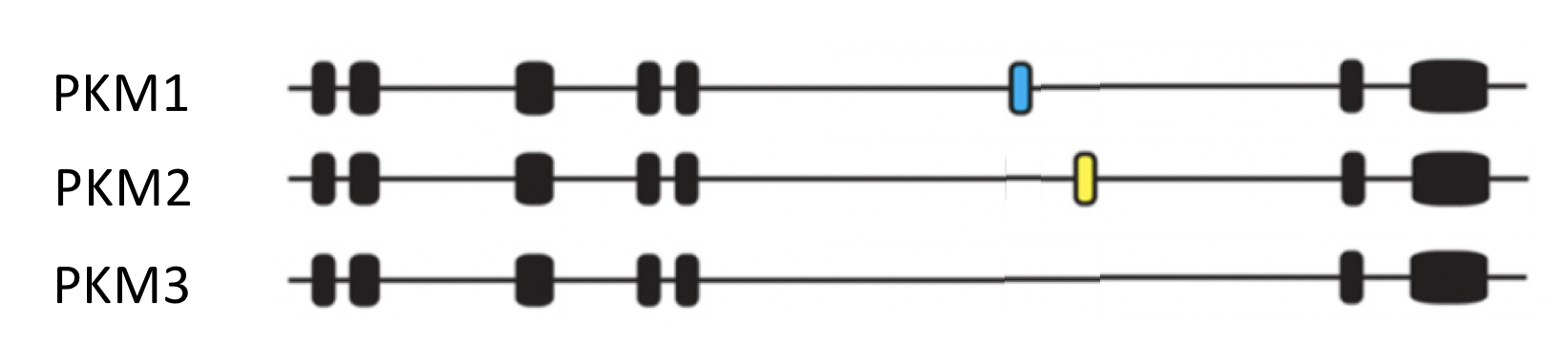


Since exons 1-5 and 8-9 are shared between the two isoforms, it is impossible to use reads that align to any of those exons to differentiate between the two isoforms. In other words, all of the reads that align to exon 1 will be equally assigned between PYM1 exon 1 and PYM2 exon 1. All of the reads that align to exon 2 will be equally assigned between PYM1 exon 2 and PYM2 exon 2, and so forth. A common practice is to give a partial count to each isoform that could be the source of the read, so in this case a half count would be assigned to PYM1 and a half count would be assigned to PYM2 for each read that aligns to exon 1. After all of the reads have been counted, the program then uses the information from reads that map unambivalently only to a single isoform in order to figure out the relative abundance of the isoforms compared to each other.

**Question 6:** Which reads from the picture above would be used to determine the relative abundance of PYM1 relative to PYM2?

Reads of exons 6 and 7 would be distinguish between PKM1 and PKM2.

While this makes sense in principle, it can start to get both conceptually and computationally difficult quite quickly. For example, let’s imagine that there was a third PYM isoform, PYM3, that only has exons 1-5 and 8-9 as in the picture below:

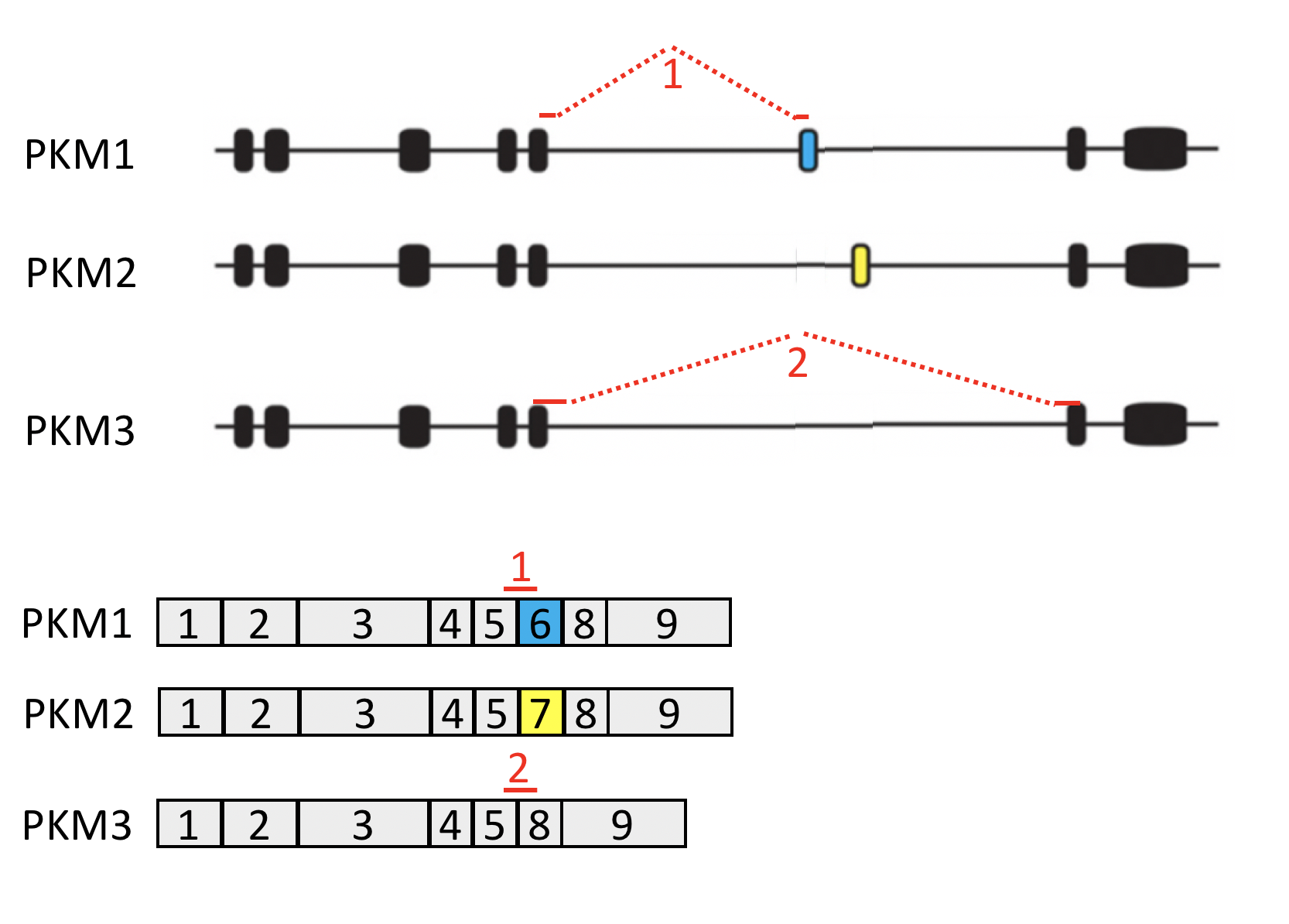


**Question 7:** Which reads from the exon counting picture could be used to differentiate between PKM3 and either PKM1 or PKM2?

Again, one would use the exon 6 and 7 reads to classify PKM1 and PKM2. But this time, any sequences that lacked both exons 6 and 7 would be classified as PKM3.

If a genome alignment was performed in addition to the transcript abundance estimation (as in 2a above), then we can use split reads that cross the splice junctions in order to help to differentiate between the isoforms. Alternatively, if we are mapping strait to the transcriptome, the same reads will align only to one isoform or another without being split.

Example:

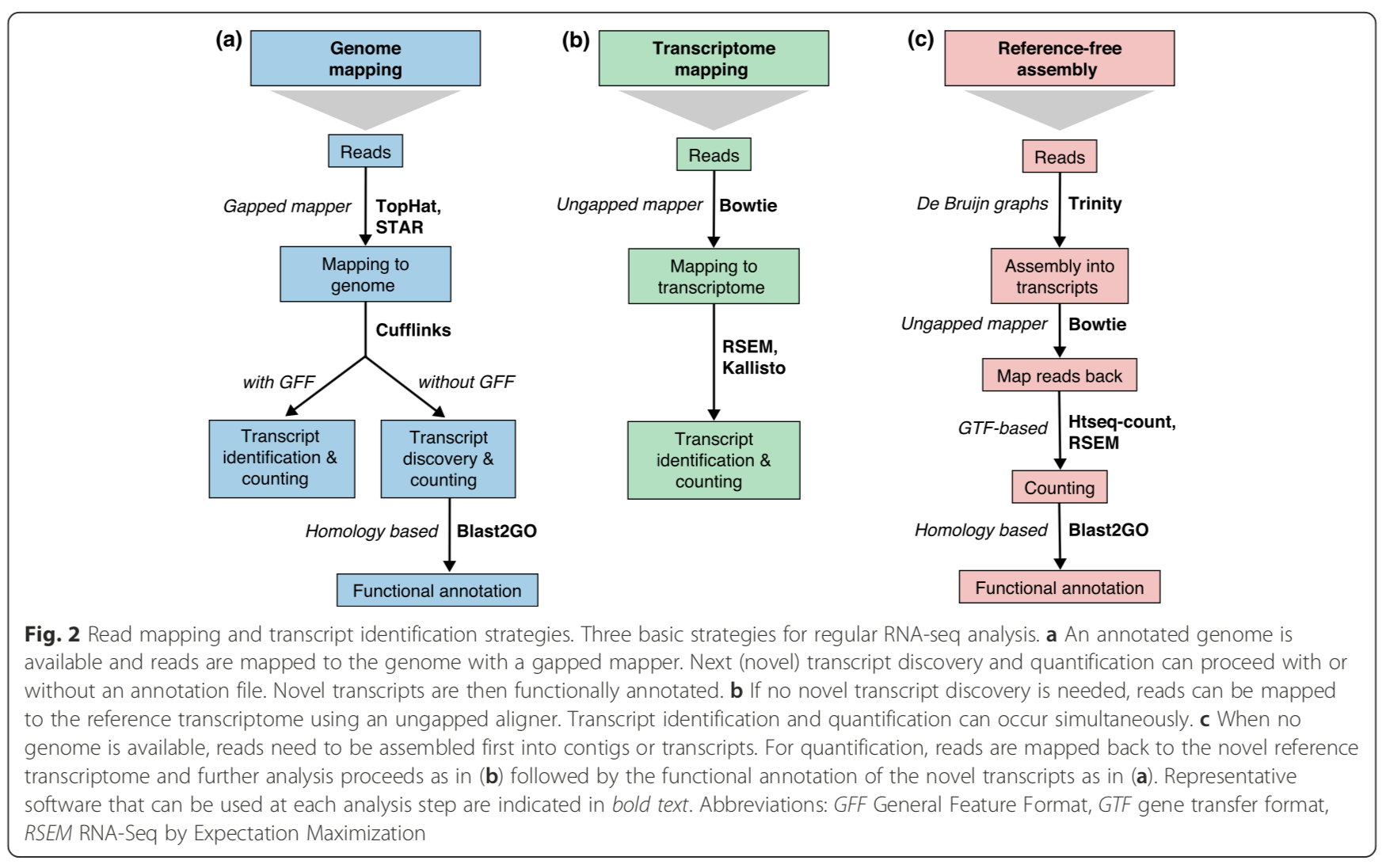


Read 1 will only align to PKM1, while read 2 will only align to PKM3

The difficulty is that only a very small number of the reads actually cross exon-exon junctions, so the majority of reads are uninformative when it comes to differentiating between the alternative isoforms. In addition, if a transcript isoform is rare, then there may not be detectable sequencing reads that map to some of the exon-exon junctions. There are many different methods for addressing the problems of assigning reads to different isoforms and calculating their abundance. Because different algorithms are used, the different methods tend to produce different results. This is not necessarily a bad thing as long as you are aware of the biases that are inherent to the method that you use and also as long as you expect the biases to affect all of the samples that you are comparing equally. If you are interested in learning more, please check out the optional reference papers included in this learning module.

**De novo transcript assembly followed by transcript-level abundance estimation**

The final gene quantification technique attempts to assemble the short reads into transcript isoforms rather than mapping them to annotated transcript sequences. De novo assembly of transcript isoforms followed by transcript abundance estimation methods can start with a genome alignment followed by transcript discovery, as in figure (**2a**) below. In this case, split reads that cross exon-exon junctions can drive detection of the transcript isoforms. Alternatively, if no reference genome exists, the reads might be assembled into transcript isoforms based on reads that map to a particular exon-exon junction as in figure (**2c**) below. In either case, the transcript assembly step is followed by another step where the relative abundance of each transcript isoform is calculated using the same abundance estimation methods that we described above.

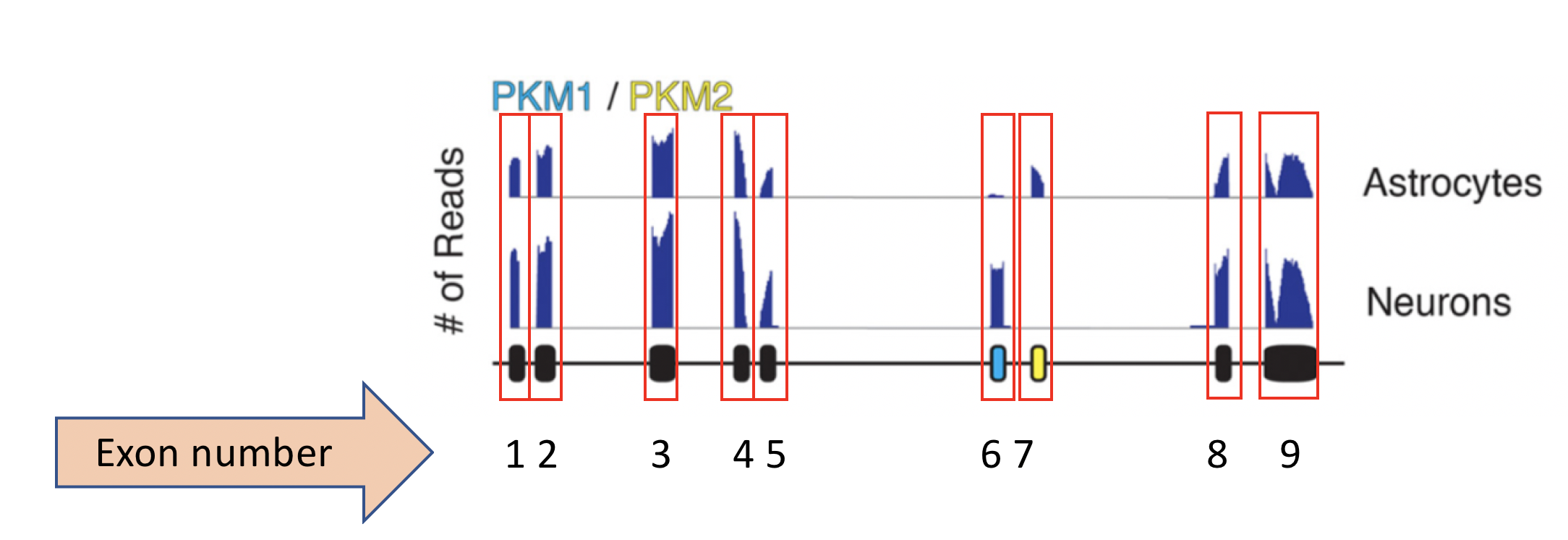


Credit: Conesa et al. Genome Biology (2016) 17:13

The transcript assembly step is another conceptually and computationally complex problem that has been addressed many different ways and the techniques frequently produce different results from each other. What makes these techniques especially problematic is that there may be sample-to-sample differences in the transcript isoforms that are predicted even when comparing samples of the same type. This is because slight sequence differences or differences in the number of sequencing reads (library size) can impact the way that the short reads are assembled into transcript isoforms. Especially if a transcript isoform is rare, there may not be detectable sequencing reads that define some of the exon-exon junctions, so some samples will appear to have more splicing isoforms than others. The program then needs to reconcile the entire list of isoforms that are present across all of the samples before it can reliably quantify the relative abundance of the all of the isoforms in each sample. (Please see the optional review papers for more information.)

**Question 8:** As a simple example of how difficult the assembly problem can be, consider the case where we have the same exon-specific reads from above (pictured again below) and a small number of split reads or reads that cross exon-exon junctions to resolve the various isoforms (not pictured, but you can imagine where they might be). How many different transcript models might we build if we did not know which PYM isoforms existed and simply tried to assemble the sequencing reads together? Just give a ballpark number i.e. more than one, more than 10, more than 100 etc. (Hint: there is no way to tell how many exons are in each transcript without reads that cross the exon-exon junctions, and even when you do have reads that cross the junctions you don’t know whether they came from the same transcript isoform or different transcript isoforms.)

Probably at least 3, maybe 4. We still have the two differing splice sites, which would indicate different isoforms. We should also be able to tell which of the exons are shared based on their higher relative frequency than two mutually exclusive isoform exons. This would in turn help us identify which exons are not shared by all of our sequences.



Now that you understand these concepts a little better, please answer the following questions based on **“A survey of best practices for RNA-seq data analysis”**

In the section of the paper that discusses read alignment as a quality control metric, we learn that the percentage of mapped reads is a global indicator of the overall sequencing accuracy and of the presence of contaminating DNA.

**Question 9:** How does the method that is used to map the reads by putting them into the context of the genome or by mapping them to preassembled transcript isoforms affect the number of mapped reads that one should expect? Why?

If we mapped to a human genome, we would expect around 70 to 90% of reads to map. But if we mapped to a transcriptome, we may expect lower percentages because reads from unannotated transcripts will be lost, and a lot more multi-mapping reads because of reads falling onto exons that are shared by different isoforms of the same read.

When alignment is discussed in the “Transcript Identification” section of the paper, we learn that one of the metrics that is used by short-read aligners is the percentage of multi-mapped reads -- these are sequencing reads that align to more than one place in the genome or across the annotated transcripts that are being quantified. In general, the more places that a sequencing read can align to the reference, the lower the mapping quality score for the read (MQ score). It is relatively easy to subset out aligned reads based on their MQ scores or based on the number of times that the read aligned to the reference, so it is not unusual for researchers to simply eliminate multiply mapped reads or reads with low MQ scores prior to downstream analysis.

**Question 10a:** What causes multi-mapped reads in a genomic alignment?

Repetitive sequences or shared domains of paralogous genes.

**Question 10b:** In general, should they be discarded? Why or why not?

No, they normally account for a significant fraction of mapping output.

**Question 11a:** What causes multi-mapped reads when the reference is the transcriptome?

Reads that would have been uniquely mapped on the genome would map equally well

to all gene isoforms in the transcriptome that share the exon.

**Question 11b:** In general, should they be discarded? Why or why not

No, they are a valuable tool for helping identify alternatively expressed genes.

**Question 12:** According to the section on transcript discovery, why is the identification of novel transcripts using short-read RNA-seq so very challenging?

Shorts reads don’t usually cross multiple splice sites which makes it difficult to infer their full-length transcripts. It is also harder to identify transcript start and end sites.

**Question 13:** How reliable are pipelines that rely on their ability to reconstruct transcript isoforms from short reads prior to quantification? In other words, do methods that reconstruct transcript isoforms from short reads typically agree with each other?

Accurate transcript reconstruction from short reads is very challenging, and so the different methods used by varying pipelines often show substantial disagreement.

**Question 14:** Why does the “Transcript quantification” section of the paper say that “raw read counts alone are not sufficient to compare expression levels across samples”? In other words, what factors other than differences in expression might affect the number reads that are counted for each transcript?

Other factors like transcript length, total number of reads, and sequencing biases can affect the number of reads counted for each transcript.

**Question 15:** When and why is it important to use measures that correct for transcript length such as RPKM, FPKM, and TPM?

It is necessary to correctly rank gene expression levels within the same sample. This is to account for the fact that longer genes accumulate more reads.

**Question 16:** When and why is it not important to correct for transcript length?

It is not necessary to use when comparing changes in gene expression within the same gene across multiple samples. This is because you are only looking at one gene whose length can be expected to stay relatively consistent. There is no need to correct because you are not comparing genes of different lengths.

Optional references on the accuracy of transcript assembly and abundance estimation methods:

Alignment and mapping methodology influence transcript abundance estimation

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02151-8

Evaluation of Seven Different RNA-Seq Alignment Tools Based on Experimental Data from the Model Plant Arabidopsis thaliana

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7084517/

**STEP 5: DIFFERENTIAL EXPRESSION ANALYSIS**

The following questions based on **“A survey of best practices for RNA-seq data analysis”** pertain to differential expression analysis:

**Question 17:** Why are normalizing methods such as that are based on total or effective counts, such as RPKM, FPKM, and TPM not sufficient for differential expression analysis?

They normalize away sequencing depth either directly or by accounting for the number of transcripts, which can differ significantly between samples. This is critical because sequencing depth is the most important factor for comparing samples.

For microarray analysis, one can perform simple comparison tests, such as t-tests for each gene to measure expression differences across samples once the data has been normalized. However, even once RNA-seq data has been normalized to account for sample-to-sample differences in library size, etc. an additional correction must be applied to normalize for gene-to-gene (or transcript-to-transcript) differences in the overall variance across samples.

**Question 18:** What is it about RNA-seq data that makes this additional correction important?

The negative binominal distribution used to compute differential expression allows for additional variance beyond the variance expected from randomly sampling a pool of molecules that are characteristic of RNA-seq data.

After the authors of “**An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex**” completed their gene-level differential expression analysis, they chose to perform qRT-PCR to validate a subset of 40 differentially-expressed genes.

**Question 19:** How did the authors choose which genes to validate? In other words, why did they pick the ones that they did in order to perform qRT-PCR?

They selected genes that had not been identified previously in literature as cell type-specific. The 40 they chose had low expression values because their validation provided confidence in the overall accuracy of the RNA-seq dataset.

**Question 20:** Which is more sensitive and accurate, qRT-PCR or RNA-seq? (Take your best guess.)

RNA-Seq because it is sensitive enough to detect novel transcripts.

**STEP 6: Gene Set Enrichment Analysis**

Another analysis that the authors of “**An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex**” completed was a gene set enrichment analysis. Unfortunately, they do not go into much detail about the analysis because they use Ingenuity Pathway Analysis, which is a proprietary method using commercial software. The makers of Ingenuity Pathway do not share the details of how their enrichment algorithm works or what genes are in the gene sets that they use. This software is accepted in the field, but please consider any Ingenuity Pathway results that you see with a huge grain of salt. Look very carefully, especially at any results that do not make biological sense.

**Question 21:** In this case, the authors saw enrichment of pathways that make biological sense. What did they see?

Enrichment of neurotransmitter receptor signaling pathways in neurons, immune cell signaling pathways in microglia, epithelial adherence junction signaling pathways in endothelial cells. They also saw enrichment of NF-kB, Wnt/B-catenin, and sonic hedgehog signaling pathways in astrocytes.

The authors also performed a WGCNA analysis, which is a method to determine which genes are co-expressed. Genes that show similar levels of upregulation or downregulation are grouped together and then these groups are compared to a list of gene sets in order to see whether specific pathways may be upregulated or downregulated in each cell type. Unfortunately, the WGCNA analysis and the RNA-seq data are no longer accessible via the website that is listed in the publication. Instead, here is an optional link where you can learn anything that you want to know about WGCNA: https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/Simulated-00-Background.pdf