In today’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.

For reference, the details of most of what we discussed are explained in “**A survey of best practices for RNA-seq data analysis**”.

First, read through “**A survey of best practices for RNA-seq data analysis**”. Then, **identify what was done by the authors of** “**An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex**” **to accomplish Goal 1**. Starting from the point where the cells have already been collected and sorted into samples of individual types of cells, what did the authors do next? Specifically:

1. How did the authors ensure that the RNA was high quality?

They used a Bioanalyzer that would calculate a RIN for each of their samples to help ensure RNA quality. Only those with a RIN of >8 were used for library construction.

1. Which types of RNA (microRNA, snRNA, mRNA, etc.) were included in the sequencing library?

They used only poly(A) RNA for their library. This would mean their library is composed primarily of mRNA, with some lncRNA and any other types that are polyadenylated.

1. Were the sequencing libraries paired-end or single-end?

They used a paired-end sequencing library.

1. Are the sequencing libraries strand specific?

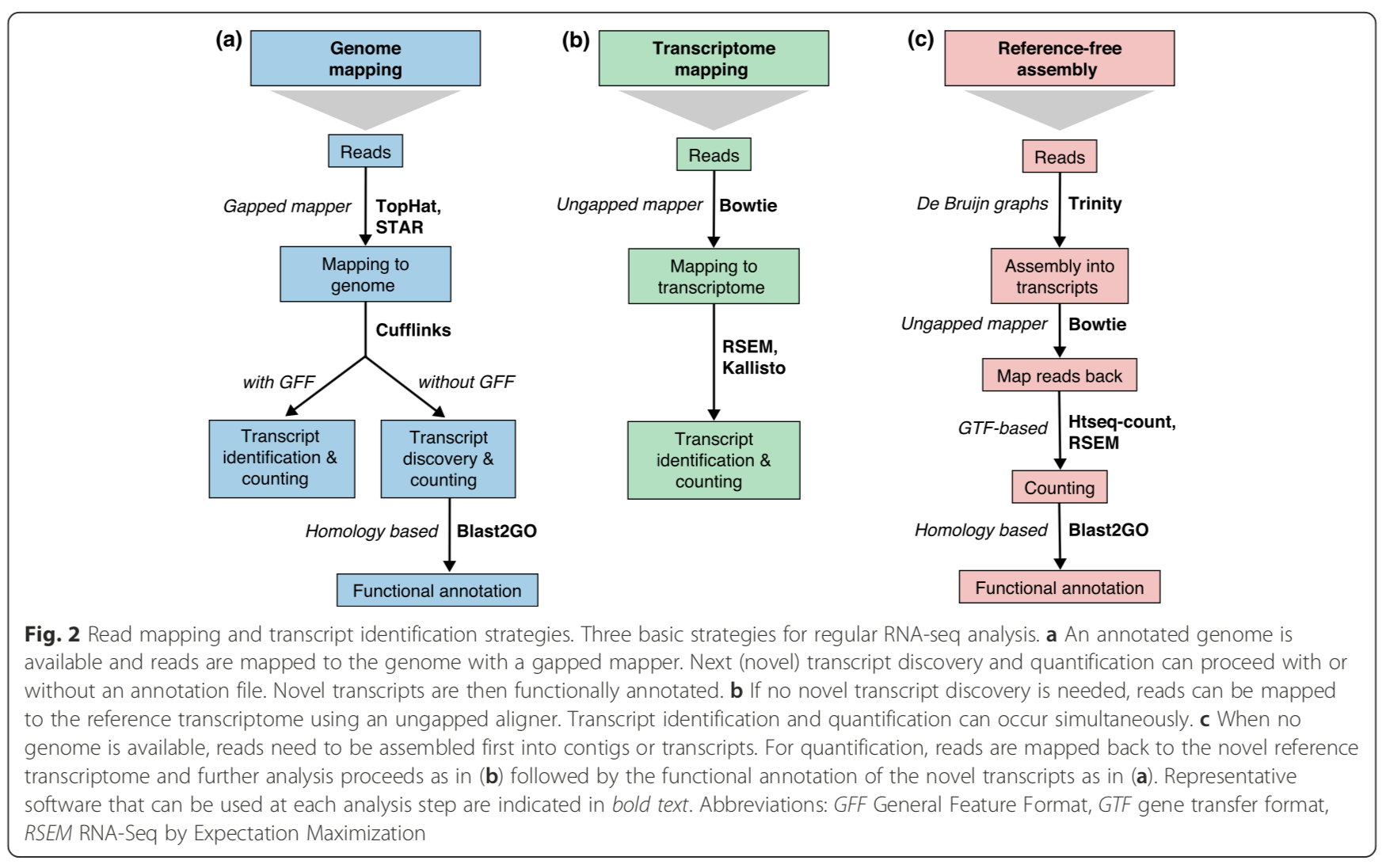
They used the TruSeq RNA test kit and library creation procedure. This would indicate that their libraries are strand specific.

1. What steps were taken to minimize batch effects and other sources of technical bias during library construction and sequencing?

Samples were collected and sequenced in the largest feasible group size and performed by the same individuals.

The authors of **“An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex”** use the sequencing reads to analyze transcription in two different ways. First, they perform a gene-level analysis to identify which genes are expressed in each cell type. Then, they perform a transcript-level analysis to compare the abundance of individual transcript isoforms generated by alternative splicing.

1. How were the sequencing reads put into context for the **gene-level** analysis?
   1. Which of the strategies from Figure 2 of “A survey of best practices for RNA-seq data analysis” (below) best matches what the authors did in “An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex”?



The authors used TopHat which utilizes the genome mapping method.

* 1. Did the authors align the sequencing reads to the genome?

Yes. They mapped the reads to a mouse reference genome.

* 1. Did the authors attempt to reconstruct full transcripts from the sequencing reads?

Yes. They used Cufflinks which is used in transcriptome reconstruction.

* 1. Did the method that the authors used allow them to detect and measure novel/unannotated transcripts?

The TopHat genome mapping approach would have allowed them to potentially detect novel or unannotated transcripts.

1. Once the reads were put into context, what metric did the authors use to determine the level of expression of each gene?

They used FPKM to measure levels of gene expression.

1. How did the authors use their expression level estimates to decide whether a gene was expressed or not in each type of cell?

They conducted an analysis based on the 95% confidence intervals of FPKM values as calculated by Cufflinks. An FPKM value of 0.04 was deemed to be the minimum threshold for gene expression.

1. How did the gene-level expression results compare to the microarray results from the same samples?
   1. Which technique proved more sensitive?

They confirmed that RNA-Seq was more sensitive.

* 1. How do the authors know?

They compared the distributions of cell type-specific gene enrichments. Fold enrichment was calculated for each gene in both the RNA-Seq and Microarray datasets. Plotting the data against each other on a scatter plot revealed that RNA-Seq was more sensitive.

1. How were the sequencing reads put into context for the **transcript-level** alternative splicing analysis?
   1. Did the authors align the sequencing reads to the genome?

Yes. They briefly *de novo* mapped some sequences to the mm9 reference genome. The algorithm OLego was used to assist them.

* 1. Did the authors attempt to reconstruct full transcripts from the sequencing reads?

No. They used a “gapless” aligner to infer transcript structure.

* 1. Did the method that the authors used allow them to detect and measure novel/unannotated transcripts?

No. The “gapless” aligner would indicate as much. However, they did use OLego to identify novel exon junctions.

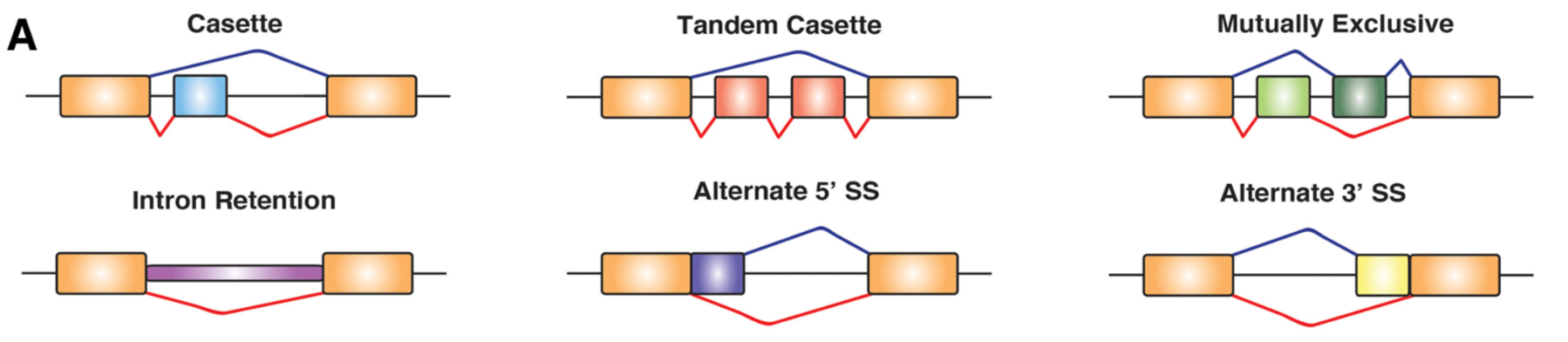
1. What were the major findings of the alternative splicing analysis?
   1. How many genes are alternatively spliced (had at least two splicing isoforms expressed) in at least one cell type?

Their analysis identified 6588 total genes that are alternatively spliced in at least one cortical cell type.

* 1. Roughly how many genes are alternatively spliced per cell type?

About the same amount of genes were alternatively spliced across all cell types. The paper said this is approximately 5549 events.

* 1. Which of the types of alternative splicing events pictured below was most commonly found?



The cassette exon comprised the majority of alternative splicing events.

* 1. Which was second most common?

Alternative 3’ SS

* 1. Which was third most common?

Intron Retention