*De novo* transcriptomes built from hundreds of eye tissues reveal hundreds of novel gene isoforms

Vinay S Swamy, David McGaughey

# Results

## A rigorous analysis pipeline finds thousands of novel gene isoforms

|  |  |  |  |
| --- | --- | --- | --- |
| subtissue | number of samples | number of studies | transcriptome size |
| Retina Adult Tissue | 105 | 8 | 49026 |
| RPE Fetal Tissue | 49 | 7 | 40098 |
| Cornea Adult Tissue | 43 | 6 | 27028 |
| Retina Fetal Tissue | 89 | 6 | 60990 |
| RPE Adult Tissue | 48 | 4 | 39172 |
| Cornea Fetal Tissue | 6 | 2 | 18697 |

Table 1. Statistics about ocular sample dataset. Transcriptome size is defined as the number of unique transcripts expressed in a given tissue type

       We built transcriptomes using 340 publicly available ocular tissue samples. We include both adult and fetal tissue from the Cornea, Retina, Retinal Pigmented Epithelium(RPE) mined from 32 different studies(Table 1) Our fetal tissues consist of both human fetal tissues and human induced pluripotent stem cell(iPSC) derived tissue. We supplemented our ocular data set with 905 samples across 46 body locations from the GTEx project.



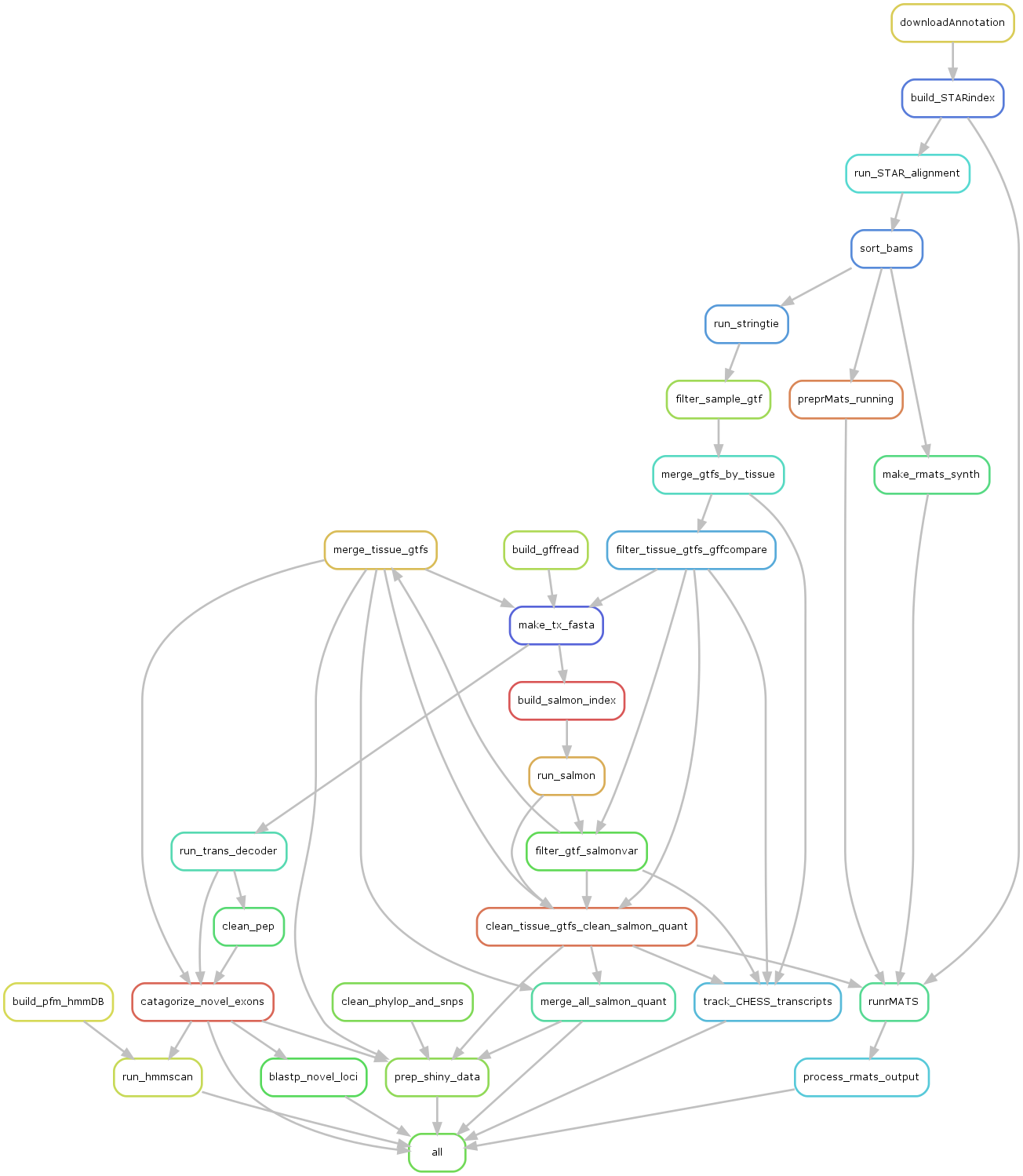


Figure 1. Directed Acyclic Graph of *de novo* transcriptome construction pipeline We designed a Snakemake pipeline to efficiently and reproducibly process and build our *de novo* transcriptomes(fig1), used it to construct a separate tissue specific transcriptome for each tissue type.

       After initial construction of transcripts, we found over two million distinct transcripts (sup table 1). We saw that many of these transcripts were detected in one or two samples(sup fig 1) so we refined our set of transcriptomes using series of filtering steps(methods) to remove low confidence transcripts. Our final transcriptome contains 340,456 distinct transcripts with 102,323 previously annotated and 238,677 novel transcripts, and incudes 300 kilobases of previously unannotated genomic sequence. We define novel as any region of genomic sequence that is not found in Gencode, Ensembl and Refseq annotation databases. Novel transcripts are split into two categories: novel isoforms which are novel variations of known genes, and novel loci, which are previously unreported, entirely novel regions of transcribed sequence.

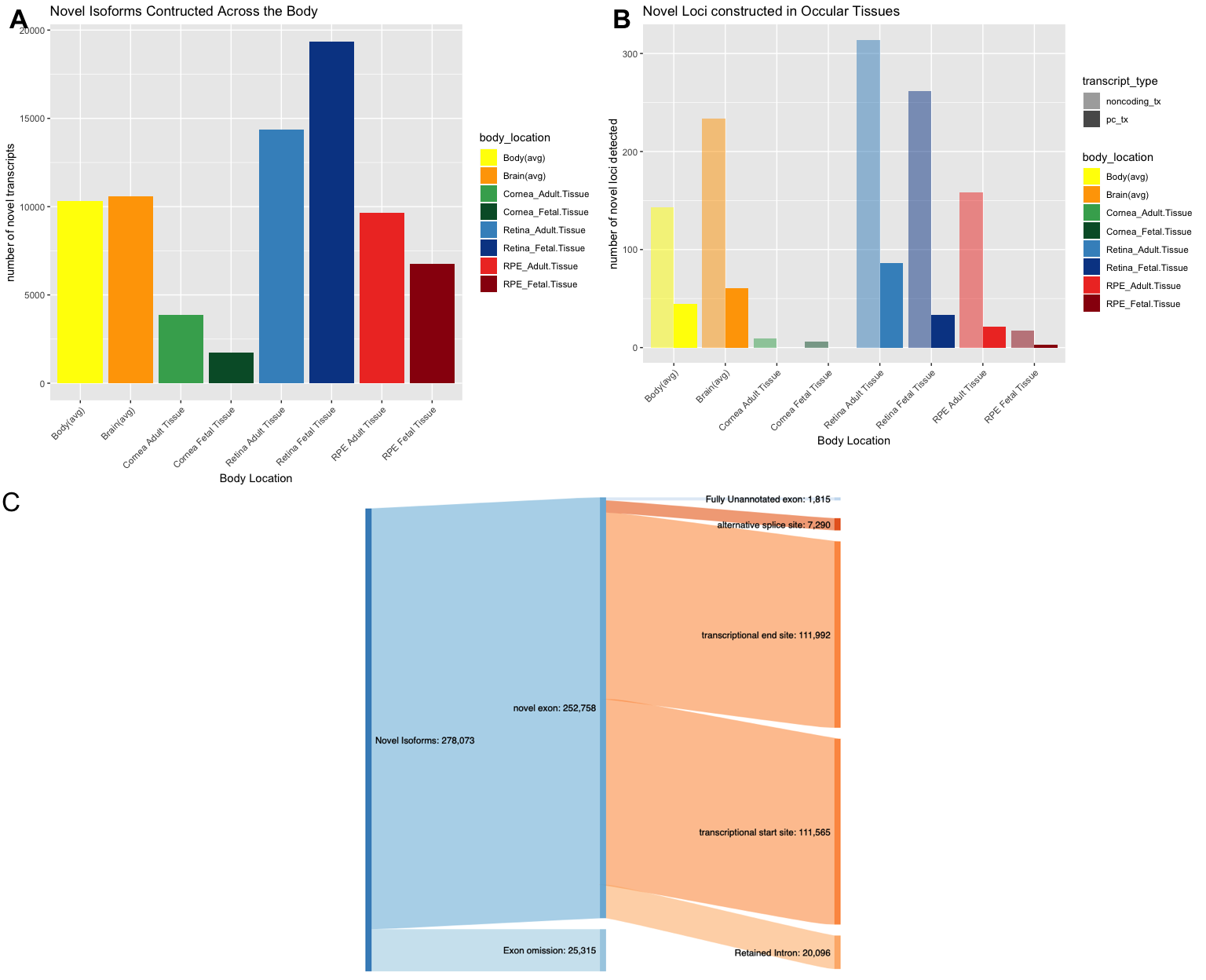


Figure 2 A,B - Number of novel gene isoforms and novel loci constructed across ocular and other body tissues.Brain and body represent an average of 13 and 34 distinct subtissues, respectively C - Classification of novel gene isoforms based on source of novelty

       Novel isoforms occur due to a novel arrangement of known exons or contain a novel exon which in turn is either a modification of a known exon to generate a novel modified exon, or an addition of a previously unannotated exon, or fully novel exon. For both classes of novel exon, we further annotate them with the putative biological process driving their inclusion: alternative splicing, alternative promoter usage, or alternative polyadenylation. We find that the majority of novel exons with our dataset are novel first and last exons.

## *de novo* transcriptomes improve sample mapping rates

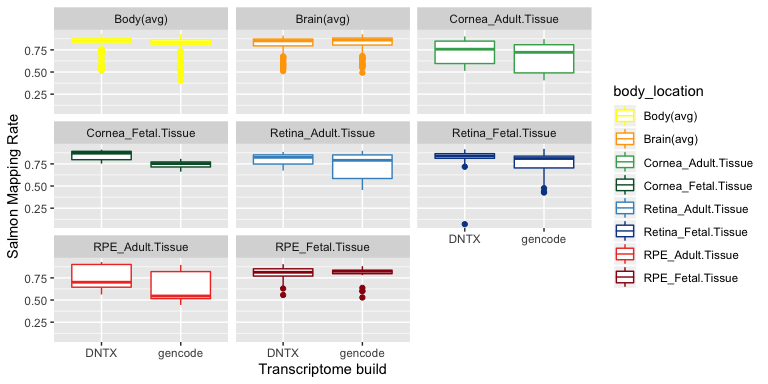


Figure 3 Salmon pseudo-mapping rates using both tissue specific *de novo* transcriptome and gencode transcriptomes as the underlying reference for salmon.

       As an initial benchmark of the accuracy and utility of our transcriptome builds, we quantified transcript expression of our samples using the alignment free quantification tool salmon. Each sample was quantified twice, once using the gencode V27 human transcript annotation, and once using its associated tissue specific transcriptome. We found that globally, salmon mapping rates increased when using the *de novo* transcriptomes to as the reference transcriptome for quantification, despite average of 3 fold reduction in annotation size. Using our transcriptomes, we quantified a total of 13420034 reads across all of our samples. Additionally, for most tissue types using *de novo* transcriptomes led to a more precise estimation of gene expression.

## Novel Isoforms in Ocular tissues

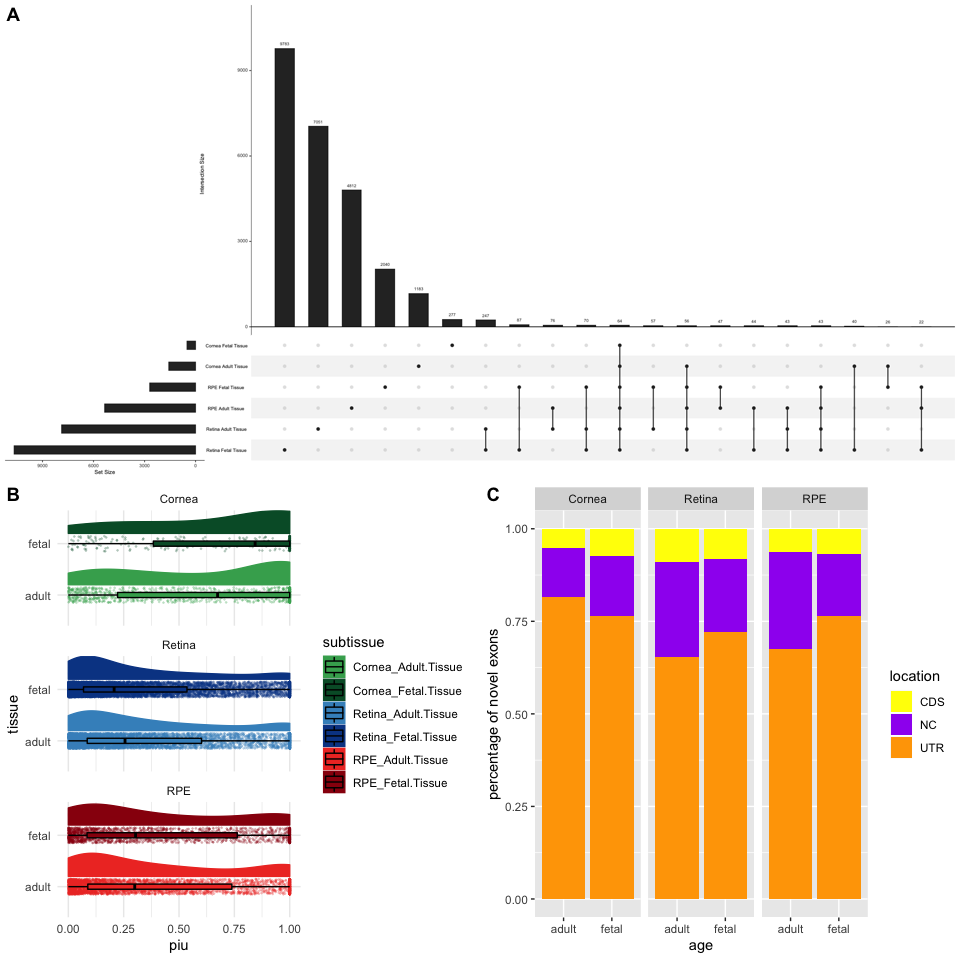


Figure 4. Analysis of novel isoform chaacteristics A. Set intersection of novel isoforms in ocular transcriptomes B. Raincloid plots of novel isoform piu. Boxplots are overlaid over piu data with estimated distribution of data set above each boxplot C. Location of Novel exons with in the body of associated parent transcript. CDS refers to coding region of transcript, UTR to untranslated region and NC to noncoding transcript

       Next, we analyzed the novel isoforms within our ocular transcriptomes. We compared the overlap in constructed novel isoforms across ocular tissues and found that 95 % of novel isoforms are specific to a singular ocular subtissue. (fig 3a) We then calculated for each novel isoform percent isoform usage (PIU), or the fraction of total gene expression a transcript contributed to its parent gene. We found that on average for Retina and RPE, novel isoforms contribute to 30 percent of their parent gene’s expression, whereas in the Cornea we saw that novel isoforms contributed on average to 75% their parent gene’s expression.  
       We next found the longest open reading frame for each novel isoform with a novel exon in order to see if the novel exon in the novel isoform caused a change in the protein coding region of the transcript. Novel isoforms with no detectable open reading frame had associated novel exons marked as noncoding. We found that novel exons lie largely in the untranslated regions of the transcript body. Less than 10% of detected novel isoforms potentially cause a change in its translated protein.

## A companion visualization tool enables easy use of *de novo* transcriptomes

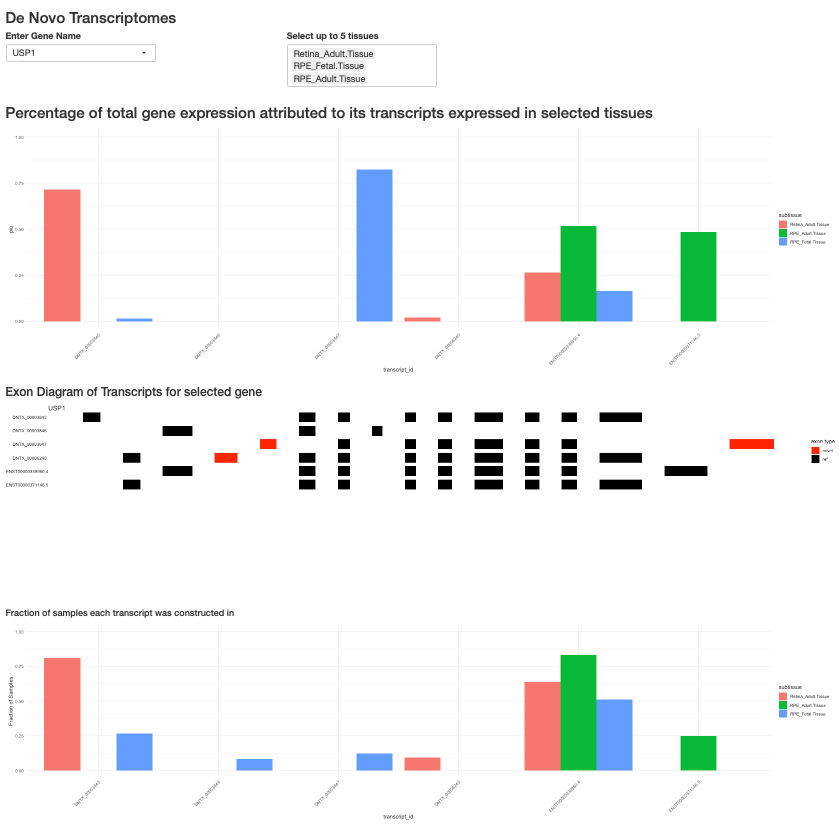


Figure 5. Screenshots from *de novo* transcriptome visualization tool. From Top to bottom: dynamic PIU bar plot for selected gene and tissue. Exon level diagram of transcript body. graph of fractions of Samples transcripts were detecgted in.

       To make our results easily accessible, we designed a webapp for visualizing and accessing our *de novo* transcriptomes. Users start by selecting gene or search for a gene by genomic location, and selecting up to 5 tissues to visualize transcript expression in. For each tissue we show the PIU for each transcript associated with a gene. We show the exon-intron structure of each transcript and mousing over exons show genomic location overlapping SNPs, and phylogentic conservation score. We additionally show a barplot of the fraction of samples in each tissue each transcript was constructed in. Users can also download the *de novo* transcriptomes for selected tissues in GTF format.