Reviewer Comments for Eye in a Disk: eyeIntegration human pan-eye and body transcriptome database version 1.0

2019-04-22

Note to all reviewers: Reviewer 2 asked why we had not included the Ratnapriya et al. large AMD retina RNA-seq study1. The reason was because the data was made available after our last data pull. While addressing the reviewer criticisms we decided to pull new GTEx samples along with all eye tissue datasets as of May 2019. This will change all sample counts in the document. We **have** corrected the incorrect code that generated the wrong sample numbers.

AE Comments to Author:  
The manuscript "Eye in a Disk: eyeIntegration human pan-eye and body transcriptome database version 1.0" is an interesting compilation of information from many datasets. The reviewers have a number of comments and queries for consideration.   
  
  
Editorial Board Member (Comments for the Author (Required)):  
  
This manuscript describes a database on genes from tissues of the eye into an accessible system for other researchers to access. They use one particular example, fetal eye tissue and organoid retina, as an example of the power of the database as constructed. While there is no hypothesis here, the ability of others to mine this curated data is likely to be of great interest to many researchers. Both reviewers have requested clarification and elaboration.  
  
1. Figure 4 is not readable. If this is going to be useful illustration of potential, it would be better to do each of the 5 subparts as a separate figure. Similarly, Figure 5 would be better separated into 2 figures with increased readability.  
2. Minor typos in the manuscript, but otherwise clearly written.

We have reduced the number of panels in Figure 4 to two and moved the remaining panels to a new supplementary figure (Supplemental Figure 7). We would prefer to keep Figure 5 as one figure, especially as we envisioned this being a fairly short manuscript and having five figures in the main body is … a lot.

Reviewer #1 (Comments for the Author (Required)):  
  
Building on the eyeIntegration web resource released in 2017, the authors re-wrote the bioinformatics pipeline and created one of the most comprehensive "reproducible and versioned transcriptome dataset" - eye in a Disk (EiaD) and updated the eyeIntegration web app, now integrating deep sequencing data from 371 human eye tissue samples and 848 GTEx samples across 44 tissues to allow quick search and visualization of human eye-related transcriptomes and make comparisons and identify differentially expressed genes among 14 various eye tissues and 44 other body tissues. It also included newly-published single cell seq data of developing and adult mouse retina. It is a powerful addition to the bioinformatics tools for the visual science community. In general, the manuscript is straightforward description of the development of the dataset and a guide to the web app. There are some minor clarifications should be done:

1) The authors quantified expression "across 42 gene and 46 transcript types" (supplemental table 2). Please describe the difference or relationship between of the "gene" and "transcript" types. Are they pretty much overlapping to each other?

The original design of Supplemental Table 2 was unnecessarily confusing. There are no differences between the “gene type” whether at the gene or transcript level. We have reworked the table, adding the Gencode description of the biotypes to the table (Supplementary Table 2, line 400) and given the counts as “Gene Counts” or “Transcript Counts” for each biotype.

2) "Figure 2: Raw RNA seq data ....... Visualizations across 63 tissues": are these 63 tissues the same as described in the abstract "44 body and 14 eye tissues", what are the extra 5 [63-(44+14)]?

We originally had 63 tissues - several were removed as they had an insufficient number of samples after quality control. I’ve updated the text for this figure legend to reflect the number after removing samples for QC reasons. The new numbers are 19 eye and 54 body tissues (which again, are post-QC). There are a few tissue types which entirely drop out after QC. For example, we have a small number of eye lid tissues that all failed QC.

3) Line 179: "To create a high quality final dataset across the 1311 initial samples (suppl fig and table 2)". Are these "1311 initial samples" the same as "371 human eye tissue samples and 848 GTEx" described on line 158-159. If so, 371+848 ≠ 1311?

916 eye and 1,375 GTEx samples now go into the pipeline. Table 1 (line 235) gives the counts pre and post-QC for the cornea, ESC, lens, retina, and RPE tissues. Eye lid and retinal endothelium were initially included, but all samples failed QC. For GTEx were go from 1,375 to 1,314 after QC (Supplemental Table, line 397)

4) In Figure 3, the colors of ESC and lens, retina and RPE are too close to each other, and hard to be distinguished. More distinctive, different colors would help. What is the total number of clusters? A table listing all the clusters and the composition of each cluster should be included as a supplemental table.

You are correct - the originals colors were poorly designed. We have updated the color scheme. We hope the update is more clear.

There are 37 total clusters (up from 33 in the last submission, as we have some more GTEx tissue types and the AMD retina study). We have added a supplementary table (line 404, Supplemental Table 4) summarizing the Sub Tissue composition of each one.

5) Line 208, ".... Then genes (or transcripts), then tissues." The "genes" should be "gene IDs by official gene name". The tables derived from "expression" search (e.g. Fig.4A) should be further annotated, e.g. define what are "decile", "log2(PMT+1)", etc.

Line 208 now reads “The user first selects either the 2017 or 2019 gene or transcript EiaD dataset, then Hugo Gene Nomeclature Committee (HGNC) genes names (or ENSEMBL transcripts), then tissues” (it is now line 247). We define decile, log2(TPM + 1) in the Figure 4 text (lines 255-258) as follows “Figure 4: Screenshots from eyeIntegration web app. A. Pan-tissue gene expression box plots with accompanying data tables. The data tables display the rank (lower is more highly expressed) of each gene in each sub tissue, the decile of the rank (10 is the highest decile of expression), and the mean gene’s log2(TPM + 1) score for each sub tissue. B. Heatmap visualization.”

6) Since this dataset and web app is built in the way that "future sample updates can be streamlined in with less effort", perhaps advice on how future deep sequencing and single cell seq data should be organized for streamline updates should be added to the discussion.

We added a short section to the discussion giving some tips on how to help with future updates to EiaD from lines 348-351: “If you wish to have your data added to EiaD in the future, we suggest you 1. deposit data into GEO/SRA, 2. use clear, descriptive, consistent, and detailed metadata for each sample, and 3. (optional) contact the corresponding author. Contacting the corresponding author is only necessary if you feel your data should be included in EiaD and was deposited into the SRA before May 8th, 2019.”

7) Many typos and some unfished edit should be corrected, e.g. Line 285 "citation" should be added!

Fixed the “CITATION” that was missing. We re-read several times and have fixed many typos.

Reviewer #2 (Comments for the Author (Required)):  
  
In this manuscript the authors describe aggregation of RNA-seq data from 371 ocular tissue samples into an "Eye in a Disk (EiaD)" database. Samples from adult and fetal retina, retinal pigment epithelium (RPE), cornea, lens, induced pluripotent stem cell (iPSC) derived retinal organoids are included. Data from two single cell RNA-seq (scRNA-seq) datasets generated from developing mouse retina are also included. The aggregated data are presented for general use via the eyeIntegration web interface. Data from 844 Genotype-Tissue Expression (GTEx) project samples derived from 44 different tissues are included for reference. To generate the EiaD database, data from RNA-seq studies that were available in the NIH Short Read Archive (SRA) were re-analyzed using the alignment free quantification tool Salmon, using the Gencode transcriptome index as a reference. Transcripts with low levels of expression (<2 TPM) were removed from the resulting dataset, as were samples with low calculated mapping rates. The resulting data were stored in an SQLite database. By using a workflow management system called SnakeMake, the authors ensure that reproducibility is maintained with each iteration of the data, facilitating sample tracking and updates of the database and portal, as new transcriptome data for the eye and other tissues become available. The whole analysis pipeline is available via Github, as is the R package used to generate the eyeIntegration web application.   
  
In addition to generating the EiaD, the authors created an interactive website with various useful features for visualizing and exploring expression data, such as differential gene expression between tissues, gene and tissue clustering based on expression profiles, weighted gene correlation networks around genes specified by the user, and t-SNE plotting, a data visualization tool for comparing expression data between different sample sets. To demonstrate these tools, the authors used t-SNE and PCA analyses to compare the gene expression profiles of the ocular and non-ocular tissues in the EiaD. They found that   
in general data from eye tissues cluster together, and that these eye tissues clusters are mostly separate from non-ocular GTEx tissue clusters. Further, analyses of RNA-seq data from iPSC derived retinal organoids and fetal retinal samples showed that the retinal organoids are similar but not identifical to fetal retina with regard to gene expression profiles. For example, HOXB genes are expressed in the retinal organoids, not fetal retina.   
  
The EiaD and eyeIntegration web application are a unique transcriptome database and data visualization portal centered on the eye, which enables quick exploration of gene and transcript expression between various healthy eye tissue types and a range of non-ocular human tissues. While the database and web tools are a valuable resource for the vision research community, there are a number of issues that need to be addressed/clarified.  
  
Comments/Issues:  
  
1. A common practice following identification of differential gene expression between two tissue types is experimental validation of a subset of the identified differences. Given the "dry lab" nature of this manuscript, this does not seem appropriate to request. But, it would be helpful to know if the differences detected in gene expression between iPSC derived retinal organoids and fetal retinal samples have been reported by other investigators. Similarly, it would be helpful to compare the output of the EiaD pipeline with similar data from the GTEx pipeline. For example, since the differences in gene expression between tissue types has been pre-calculated, it should be relatively easy to compare differential expression results for comparisons between GTEx tissues produced from GTEx portal and from the EiaD pipeline.

To address this concern about the accuracy of our process we directly compared the GTEx provided count table against the counts we calculated for the identical GTEx samples. The code below can be run in R to re-produce our finding: which is that we have a pearson (R^2) correlation of 0.89

# R code  
# need about 64GB of memory!  
library(tidyverse)  
gtex <- read\_tsv('https://storage.googleapis.com/gtex\_analysis\_v7/rna\_seq\_data/GTEx\_Analysis\_2016-01-15\_v7\_RNASeQCv1.1.8\_gene\_tpm.gct.gz', skip = 2)  
tpm <- read\_tsv('/data/mcgaugheyd/datashare/eyeIntegration/2019\_gene\_TPM\_04.tsv.gz')  
metadata <- read\_tsv('/data/mcgaugheyd/datashare/eyeIntegration/2019\_metadata\_04.tsv.gz')  
metadata <- metadata %>% mutate(si = str\_extract(sample\_attribute, "GTEX\\S+"))  
samples <- metadata$si[metadata$si %in% colnames(gtex)]  
genes <- tpm$ID[tpm$ID %in% gtex$Description]  
# make data long to join GTEx and EiaD results  
long\_tpm <- tpm %>% gather(Sample, TPM, -ID)  
long\_gtex <- gtex %>% select(-Name) %>% gather(Sample, GTEX\_TPM, -Description)  
# filter EiaD tpm down to just GTEx samples  
long\_tpm2 <- left\_join(long\_tpm, metadata %>% select(sample\_accession, si) %>% unique(), by = c("Sample" = "sample\_accession")) %>% filter(si %in% samples)  
# merge EiaD and GTEx  
both <- left\_join(long\_tpm2 %>% select(-Sample) %>% rename(Sample = si), long\_gtex %>% filter(Sample %in% samples) %>% rename(ID=Description), by = c('ID','Sample'))  
# remove NA  
both2 <- both %>% filter(!is.na(TPM), !is.na(GTEX\_TPM))  
# calculate correlations with log2(TPM + 0.01 pseudocount) as per Zhang et al.   
cor(log2(both2 %>% pull(TPM) + 0.01), log2(both2 %>% pull(GTEX\_TPM) + 0.01), method = 'spearman')  
# spearman 0.91  
cor(log2(both2 %>% pull(TPM) + 0.01), log2(both2 %>% pull(GTEX\_TPM) + 0.01), method = 'pearson')  
# pearson 0.89

To ensure there are no substantial differences in quantification of gene TPM values, we calculated the between GTEX and EiaD generated TPM values for our shared GTEx samples (see methods); we computed an of 0.89. Zhang et al. report that RNA-seq quantifications done between alignment-free methods (used in EiaD) and alignment-based methods (used by GTEx) get a ranging from 0.89 to 0.932. As Zhang et al. compared quantification methods with identical gene references (we use Gencode GRCh38 gene models and GTEx uses hg19) and did not scale TPM score differently, our result falls in line with expectations.

We have added a section to our methods and results section discussing this briefly (lines 126-129, 227-232).

2. It would be helpful to provide some information about the accuracy of the Salmon transcript quantification tool for the reported analyses. For example, a search for Rhodopsin expression in the single cell RNA seq data shows relatively high expression in cone cells, in which rhodopsin is not usually expressed. Is this an artefact of the scRNA-seq method, or a possible read assignment error from the Salmon tool?

The single cell expression section of eyeIntegration is still in active development. Right now we are using the counts quantified by the original research groups. As more scRNA-seq datasets come available we will be re-processing the data ourself. The high rhodopsin expression is most likely because (some) rods are being incorrectly labelled as cones by the original authors.

The methods section for scRNA-seq procesesing reads “We use the processed gene count data directly from each group, as well as their cluster assignments which specify what cell type each individual cell is. The count data is mean averaged to the cell type, age, and gene level for the single cell expression section of eyeIntegration.”

This question is also being asked because Salmon is an alignment-free method that performs light weight mapping of reads to transcript models and estimates expression levels of transcripts correcting for potential biases such as GC content. Given the limitations of transcript reconstruction from short RNA-seq reads, a commonly used approach to unify data from different studies, including GTEx, ENCODE, TopMed, and TCGA has been uniform realignment, gene expression quantification, and batch effect removal (e.g. Sci Data 2018 - DOI: 10.1038/sdata.2018.61). While Salmon has been reported to be among the top performing methods for transcript quantification, it also has limitations. For example, Salmon does not seem to perform as well for lowly expressed transcripts, and its efficacy can be affected by incompleteness of the gene/transcript annotation model or total number of isoforms per gene (BMC Genomics 18:583, 2017; Briefings in Bioinformatics 18: 260, 2017). It would be helpful to include information about the choice and use of this method in the manuscript.

We have added a “limitations” section to the results section of the manuscript:

“Salmon quantification, while highly performant and accurate, has a higher variance when with lower read depth and with shorter transcripts2. Extra care should be taken with comparisons with lower counts of samples (cornea, RGC) as smaller sample numbers decrease the confidence in differential expression. We recommend that if you wish to directly compare your RNA-seq sample quantifications with EiaD, you should not directly compare our TPM values with yours as there are many important variables that will differ. Instead run our Snakemake pipeline (<https://www.github.com/davemcg/EiaD_build>) and add your samples. Finally, we would like to remind any users that RNA-seq methods measure mRNA levels, but the functional unit is the protein; westerns are still the gold standard with which to evaluate expression and localization.”

3. It would be helfpul to learn more about how the authors handled non-biological batch effects between studies in their differential expression analyses. For example, in the pan tissue t-SNE plot shown in Figure 1 the majority of eye tissues cluster next to each other, but there are two separate clusters of retina and RPE samples that are distant from the other retina and RPE sample sets. Are these differences real, or due to differences in the RNA-seq library preparations and sequencing protocols between the different samples downloaded from SRA? One potential approach to this issue is to remove non-specific variation from the expression values in the differential expression model, as reported by Wang et al in their unification of GTEx and TCGA RNA-seq sample data in order to compare expression levels between normal and tumor samples from two different studies (Sci Data 2018, DOI: 10.1038/sdata.2018.61). One method that can be used to infer hidden covariates that capture much of the non-specific variation in expression datasets is ComBat in the R package SVAseq that enables indicating known variables that should be adjusted for, such as different studies (Nucleic Acids Res. 42:e161, 2014; Biostatistics 8:118-127, 2007). It is recommended that this be applied to each tissue type separately, combining samples from the same tissue type, such as retina, across different studies, and adding the study as an indicator in ComBat.

The major way we handle non-biological batch effects is through the use of <https://github.com/stephaniehicks/qsmooth>, which reshapes the distribution of gene expressions values to be consistent among each tissue type (e.g. Retina/RPE/cornea).

While ComBat is an extremely popular and powerful tool to remove batch effects, I am hesitant to use as there is some evidence that these types of tools may overinflate confidence in group differences (see <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4679072/>)3.

We have made a significant change to how our TPM values are computed, using limma’s batchEffects() tool to correct the TPM values for mapping rate and sub tissue. These two changes clean up the t-SNE plotting for the eye tissues. The major difference we see within a single sub tissue type is adult retina - for some reason we don’t understand the Ratnapriya / Swaroop retina AMD samples (both non-AMD and AMD) cluster apart from all of the other adult retina tissue samples we have collected. This may be related to the lower mapping rate (and RIN scores) from this cohort.

We do not use the tranformed TPM values for the limma / voom differential testing. These transformed values are only used the the t-SNE, boxplots, and the heatmap visualizations.

4. Further information about the quality control and filtering applied to the RNA-seq data would be helpful to include in the manuscript. For example, how were the parameters for filtered by a mapping rate of 40% per sample and median read count >2 across all genes per sample selected? Also, basic statistics related to read mapping such as min, max, average, 25 quartile, 75 quartile of mapping should be included.

Parameters were selected by checking distributions and picking the parameter that removes the left tail. We have added mapping rate as a supplementary figure and edited the text in the methods accordingly. Line 111-112: “We first removed samples with a Salmon calculated mapping rate less than 40%. This value was selected as being the far left tail of the distribution of mapping rates across samples (Supplemental Figure 2).”

5. To identify outlier samples the authors measured the distances of each sample from the centroid coordinates per cluster using a t-SNE plot. t-SNE is good for visualization but not for quantification of relative distance between samples and sample sets. The t-SNE retains probabilities not distances and hence the Euclidean distances between points on the 2-dimensional t-SNE plot are not informative. t-SNE documentation suggests that Kullback-Leibler divergences reported by t-SNE can be used for this purpose. This reviewer recommends use of coordinates from other dimensionality reduction methods such as Principal Coordinates Analysis (PCoA) or PCA, or a correlation-based statistic that compares the expression profile of each sample to the profiles of all samples in a given sample set (D statistic) to identify outlier samples, such as described in Nature Genetics 46: 430, 2014.

I first wrote a long explanation of how what we are doing is reasonable, but then thought some more and now have come around to your position. While this t-SNE / euclidean distance cutoff approach works on a empirical level, it is a poor idea.

We have implemented the outlier-by-correlation-to-group approach you suggested (Nature Genetics 46: 430, 2014) (lines 119-125):

“To identify outliers we followed an approach similar to a method in Wright et al4. Briefly, we first selected the 3000 genes with the highest variance across all samples and then for each sub-tissue type and each sample in , we first calculated , the average correlation between and all other samples in . Next, we calculated , where and is the grand mean of all for in . We removed samples with -17.5; we determined this threshold by generating a tSNE plot of our samples, and visually identifying outliers in adult retina tissue. The () amongst these outliers was -17.58 and from this we chose -17.5 as our outlier threshold.”

6. It would be helpful to learn why the authors chose 892 samples from 44 tissues in GTEx, when there are over 11,000 RNA-seq sample available for 50 tissues in the current GTEx release v7 that is available on dbGaP. Also, a recent publication from other investigators at NEI described a large RNA-seq dataset from human retinal samples (Nat Genet. 51:606-610, 2019). Will these data be incorporated into the eyeIntegration database? If so, it would be helpful if this were mentioned in the text. If not, this additional study and the resources derived from it could be referred to.

The value of the full GTEx dataset comes in the power to identify expression quantitative trait loci (eQTL). For differential gene expression by tissue, far fewer samples are needed. A power calculation we ran gives us 80+% power to detect >= 1 log2(FC) gene expression differences with a 20 sample x 20 sample comparision. We would reach a power of 90% with 30 samples, so we have added 10 more samples to each GTEx tissue group. As we have many more eye samples, this is likely an underestimate of our power when doing eye vs (body) comparisons.

We have added the power calculation curve as supplemental figure.

While it would be nice in principle to just use the entire GTEx dataset, there are very practical limitations on why this is inconvenient - the size of the full GTEx datasets, raw, is around 1 petabyte of data, which is very slow to move around and do any compute on. Also this amount of storage space brushes against the limits of our allocation in the compute cluster. As for the Ratnapriya/Swaroop data, we have now added it to EiaD along with a few other smaller studies which were made available after our last EiaD update in January 2019.

1. Ratnapriya R, Sosina OA, Starostik MR, et al. Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat Genet*. February 2019. doi:[10.1038/s41588-019-0351-9](https://doi.org/10.1038/s41588-019-0351-9)

2. Zhang C, Zhang B, Lin L-L, Zhao S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*. 2017;18(1):583. doi:[10.1186/s12864-017-4002-1](https://doi.org/10.1186/s12864-017-4002-1)

3. Nygaard V, Rødland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*. 2016;17(1):29-39. doi:[10.1093/biostatistics/kxv027](https://doi.org/10.1093/biostatistics/kxv027)

4. Wright FA, Sullivan PF, Brooks AI, et al. Heritability and genomics of gene expression in peripheral blood. *Nature Genetics*. 2014;46(5):430-437. doi:[10.1038/ng.2951](https://doi.org/10.1038/ng.2951)