recount2 BRAIN

Analysis-ready human curated sample metadata for brain RNA-seq studies

recount-brain: a curated repository of human brain RNA-seq datasets metadata

Leonardo Collado-Torres [6,*]

@fellgernon

leo.collado@libd.org

Ashkaun Razmara[1] Shannon E Ellis[2] Dustin J Sokolowski[3] Sean Davis[4] Michael D Wilson[3] Jeffrey Leek[5] Andrew E Jaffe[5,6]

[1] *Frank H. Netter MD School of Medicine at Quinnipiac University, North Haven, CT*

[2] *Department of Cognitive Science Department, University of California San Diego, La Jolla, CA*

[3] *Department of Molecular Genetics, University of Toronto*

[4] *Center for Cancer Research, National Cancer Institute, NIH*

[5] *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore*

[6] *Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore*

1 Introduction

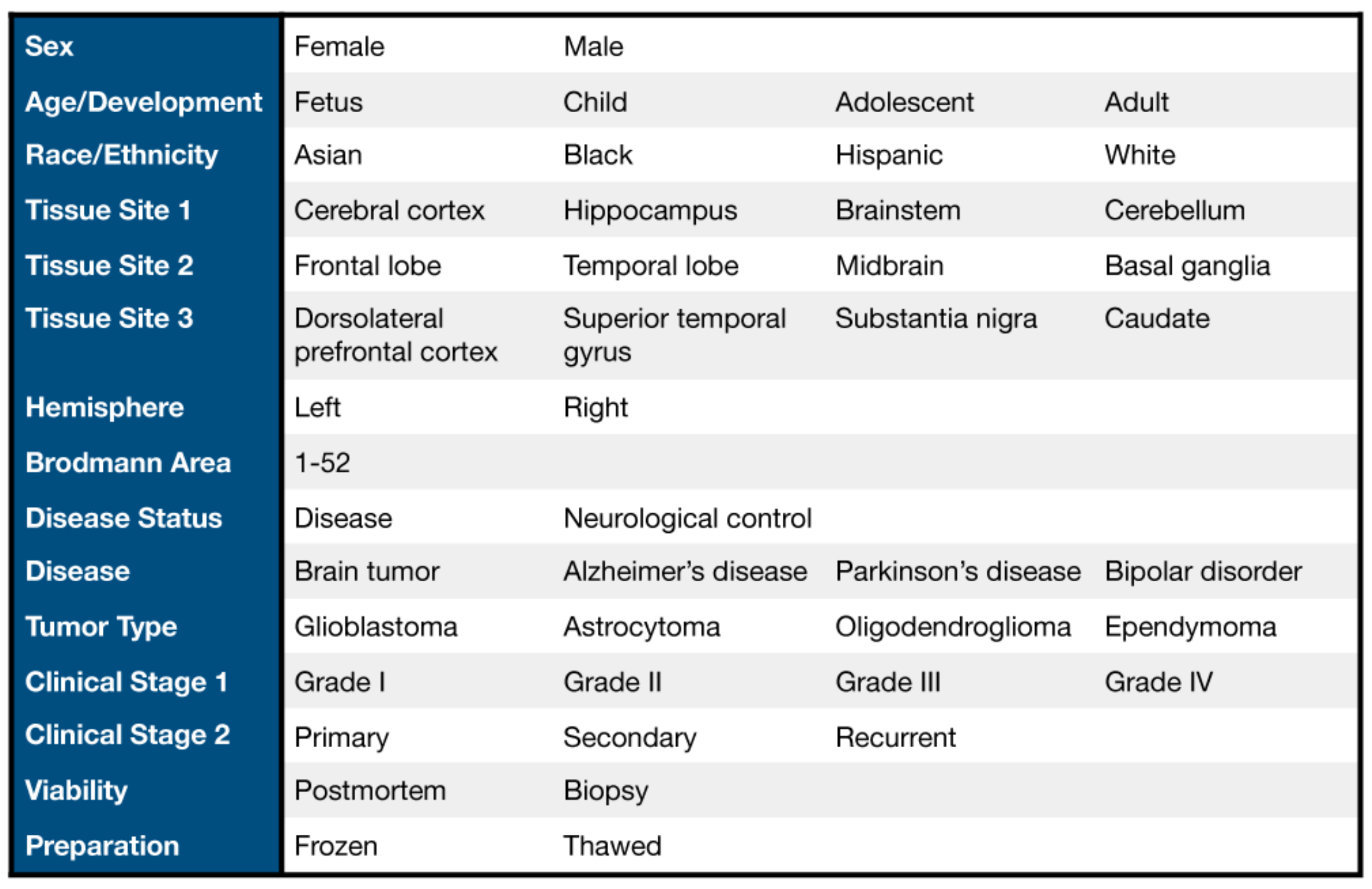
1. Uniformly-processed RNA-seq is available in `recount2` (Collado-Torres et al. 2017) and other projects;
2. Sample metadata from SRA is inconsistent, thus re-using this public data is challenging;
3. Metadata can be predicted from expression (Ellis et al. 2018) and mapped to ontologies (Bernstein, Doan, and Dewey 2017).

2 Methods

We identified SRA studies present in `recount2` that had at least 4 samples with at least 70% of them were predicted to correspond to the brain using `phenopredict` (v0.0.03) (Ellis et al. 2018). Figure 6 of (Razmara et al. 2019) shows the reproducible curation workflow we followed that briefly involved: creating a list of metadata variables of interest, documenting which part of the paper/supplement the information came from, and any custom modifications. We merged `recount-brain` with GTEx and TCGA brain sample metadata and linked to controlled vocabulary terms for Brodmann region, tissue and disease.

3 Results

In total, there are 6,547 samples with metadata in `recount-brain` with 5,330 (81.4%) present in recount2 from 62 SRA studies, GTEx (n=1,409) and TCGA (n=707). The curated metadata can be interactively explored through jhubiostatistics.shinyapps.io/recount-brain/. Figure 3.1 exemplifies some of the metadata information available for these studies.

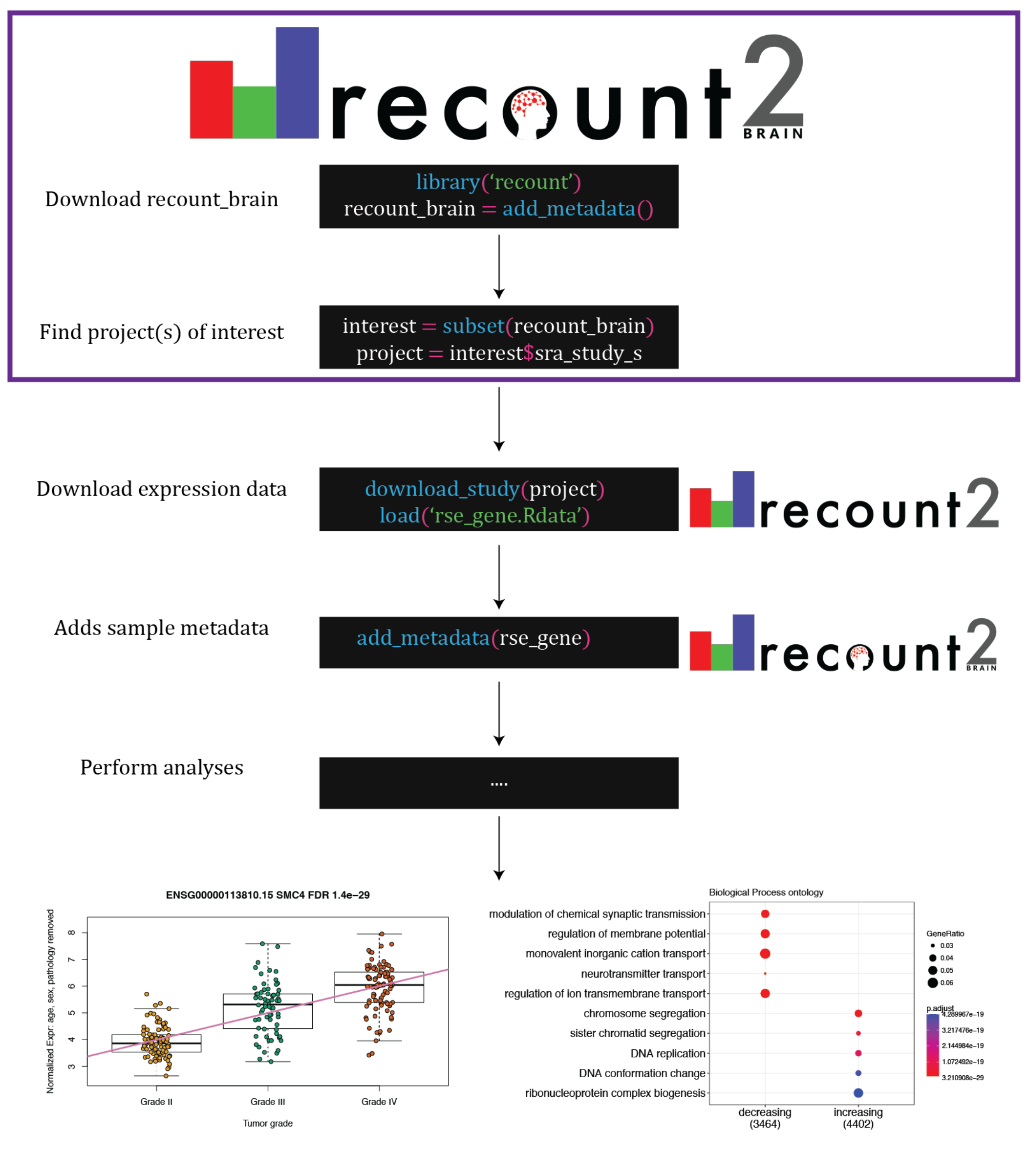


Sex	Female	Male		
Age/Development	Fetus	Child	Adolescent	Adult
Race/Ethnicity	Asian	Black	Hispanic	White
Tissue Site 1	Cerebral cortex	Hippocampus	Brainstem	Cerebellum
Tissue Site 2	Frontal lobe	Temporal lobe	Midbrain	Basal ganglia
Tissue Site 3	Dorsolateral prefrontal cortex	Superior temporal gyrus	Substantia nigra	Caudate
Hemisphere	Left	Right		
Brodmann Area	1-52			
Disease Status	Disease	Neurological control		
Disease	Brain tumor	Alzheimer's disease	Parkinson's disease	Bipolar disorder
Tumor Type	Glioblastoma	Astrocytoma	Oligodendroglioma	Ependymoma
Clinical Stage 1	Grade I	Grade II	Grade III	Grade IV
Clinical Stage 2	Primary	Secondary	Recurrent	
Viability	Postmortem	Biopsy		
Preparation	Frozen	Thawed		

Figure 3.1: Overview of some recount-brain sample metadata variables

3.1 Example usage

Select studies or add the sample metadata to the expression data with `recount::add_metadata()` (Figure 3.2).



Download recount_brain

```
library('recount')
recount_brain = add_metadata()
```

Find project(s) of interest

```
interest = subset(recount_brain)
project = interest$sra_study_s
```

Download expression data

```
download_study(project)
load('rse_gene.Rdata')
```

Adds sample metadata

```
add_metadata(rse_gene)
```

Perform analyses

```
....
```

recount2 BRAIN
recount2
ENSG00000113810.15 SMC4 FDR 1.4e−29
Normalized Expr: age, sex, pathology removed
Grade II
Grade III
Grade IV
Tumor grade
Biological Process ontology
modulation of chemical synaptic transmission
regulation of membrane potential
monovalent inorganic cation transport
neurotransmitter transport
regulation of ion transmembrane transport
chromosome segregation
sister chromatid segregation
DNA replication
DNA conformation change
ribonucleoprotein complex biogenesis
decreasing
increasing
GeneRatio
p.adjust

Figure 3.2: Access recount-brain using the recount Bioconductor package

As an example of how you can use `recount-brain`, we used studies with post mortem interval (PMI) information to assess whether expression of *RNASE2* is associated with PMI. In studies present in `recount-brain` we did find an overall association as shown in Figure 3.3 in contrast to (Ferreira et al. 2018)'s findings. A sensitivity analysis releaved study variability which is why Ferreira et al likely did not observe this association.

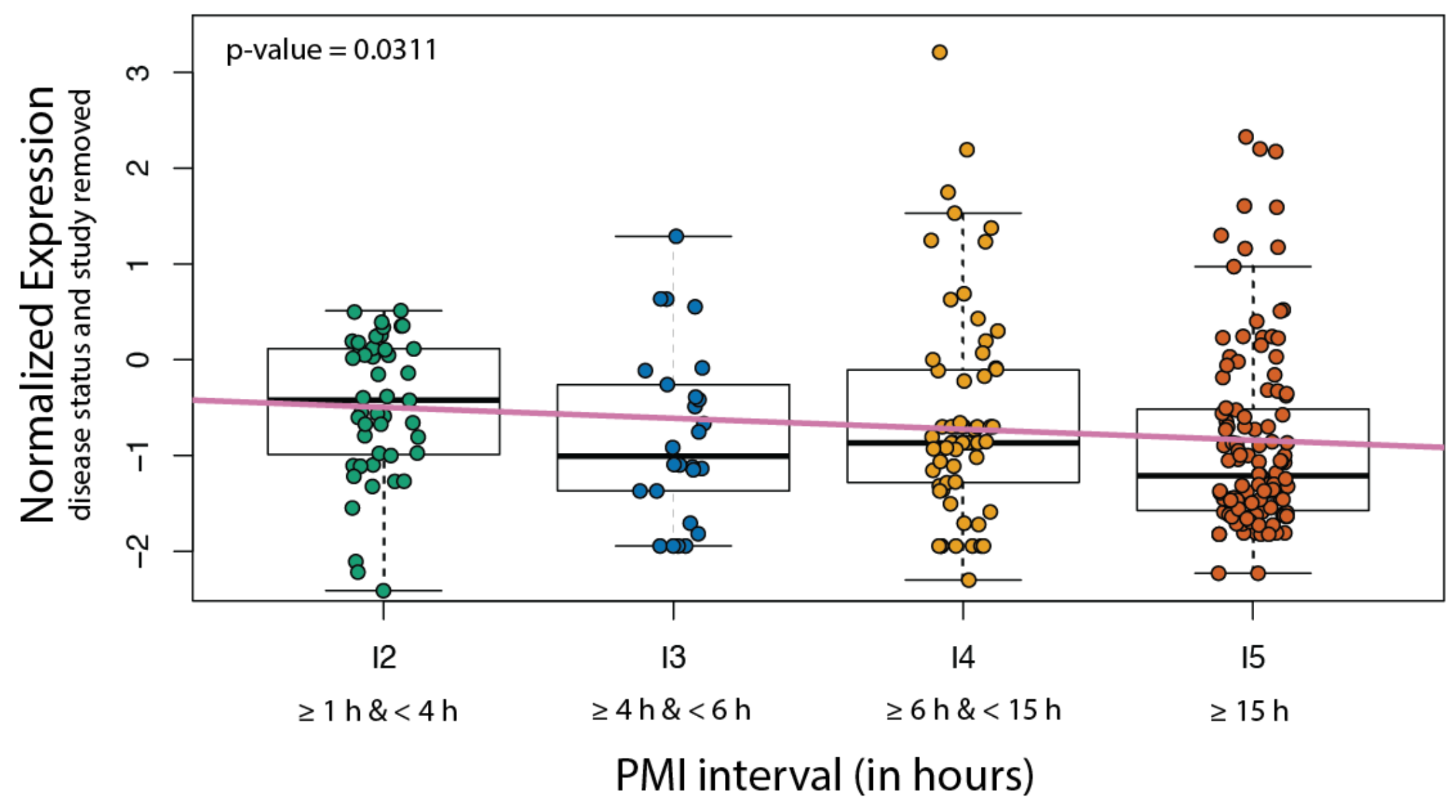


Figure 3.3: Replicate findings from other studies using recount-brain

We used `recount-brain` to determine the consistency of gene variability across glioblastoma studies SRP027383 and SRP044668 as well as TCGA (Figure 3.4).

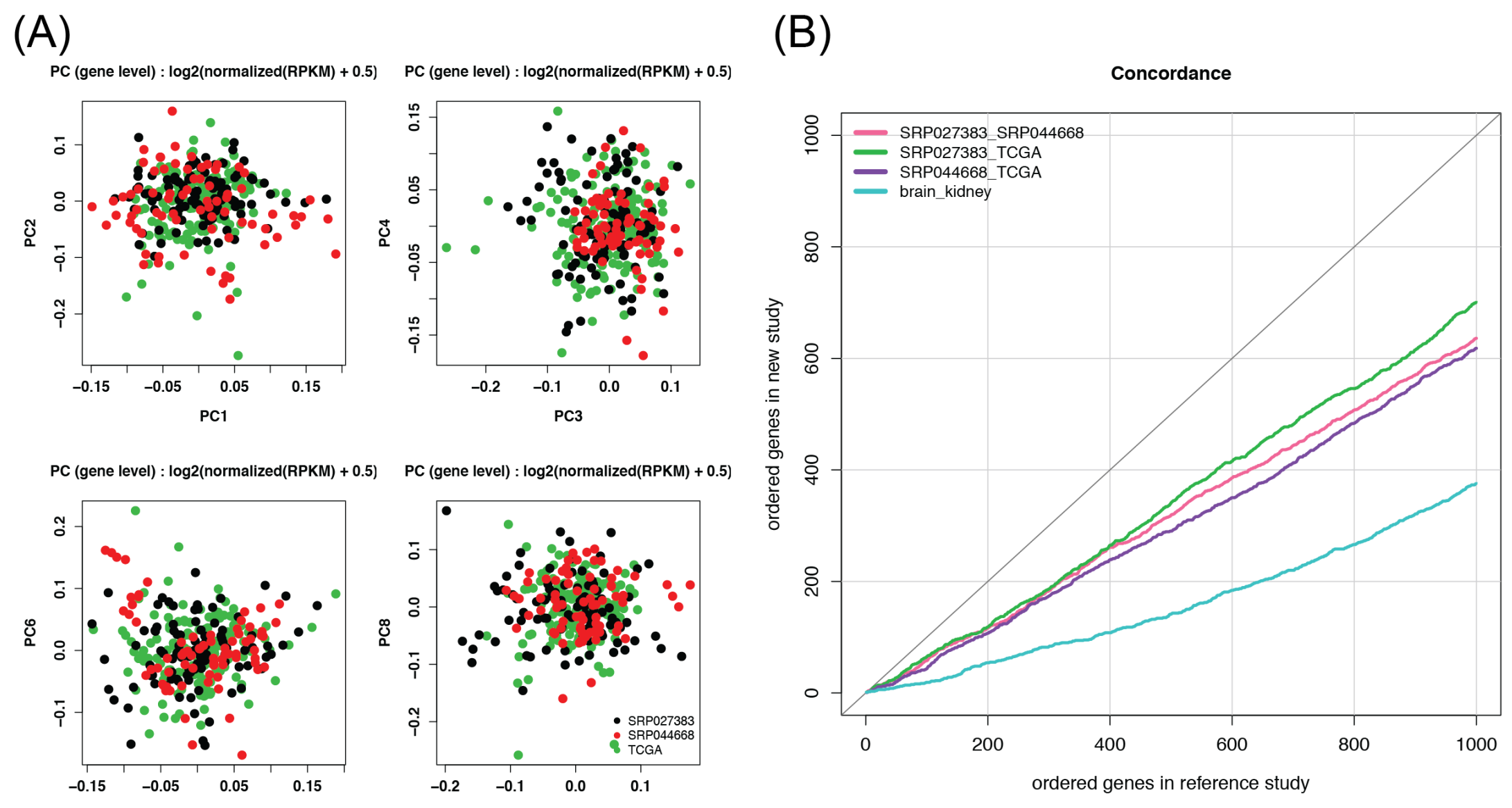


Figure 3.4: Assess consistency of gene variability across glioblastoma studies

4 Conclusions

1. `recount-brain` (Razmara et al. 2019) facilitates human brain RNA-seq analyses.
2. `recount-brain` can be used for reproducing analyses, replicating findings and assessing cross-study variability.
3. Curation efforts are complementary to prediction efforts (Ellis et al. 2018) and automatic ontology mapping (Bernstein, Doan, and Dewey 2017).
4. Our reproducible curation workflow can be adapted to curate more samples and other studies.

References

Bernstein, Matthew N., AnHai Doan, and Colin N. Dewey. 2017. "MetaSRA: Normalized Human Sample-Specific Metadata for the Sequence Read Archive." *Bioinformatics* 33 (18): 2914–23. https://doi.org/10.1093/bioinformatics/btx334.

Collado-Torres, Leonardo, Abhinav Nellore, Kai Kammers, Shannon E. Ellis, Margaret A. Taub, Kasper D. Hansen, Andrew E. Jaffe, Ben Langmead, and Jeffrey T. Leek. 2017. "Reproducible RNA-Seq Analysis Using Recount2." *Nature Biotechnology* 35 (4): 319–21. https://doi.org/10.1038/nbt.3838.

Ellis, Shannon E., Leonardo Collado-Torres, Andrew Jaffe, and Jeffrey T. Leek. 2018. "Improving the Value of Public RNA-Seq Expression Data by Phenotype Prediction." *Nucleic Acids Research* 46 (9): e54–e54. https://doi.org/10.1093/nar/gky102.

Ferreira, Pedro G., Manuel Muñoz-Aguirre, Ferran Reverter, Caio P. Sá Godinho, Abel Sousa, Alicia Amadoz, Reza Sodaei, et al. 2018. "The Effects of Death and Post-Mortem Cold Ischemia on Human Tissue Transcriptomes." *Nature Communications* 9 (1): 490. https://doi.org/10.1038/s41467-017-02772-x.

Razmara, Ashkaun, Shannon E. Ellis, Dustin J. Sokolowski, Sean Davis, Michael D. Wilson, Jeffrey T. Leek, Andrew E. Jaffe, and Leonardo Collado-Torres. 2019. "Recount-Brain: A Curated Repository of Human Brain RNA-Seq Datasets Metadata." *bioRxiv*, April, 618025. https://doi.org/10.1101/618025.

LIBD
LIEBER INSTITUTE *for*
BRAIN DEVELOPMENT
MALTZ RESEARCH LABORATORIES

10.1101/618025
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY