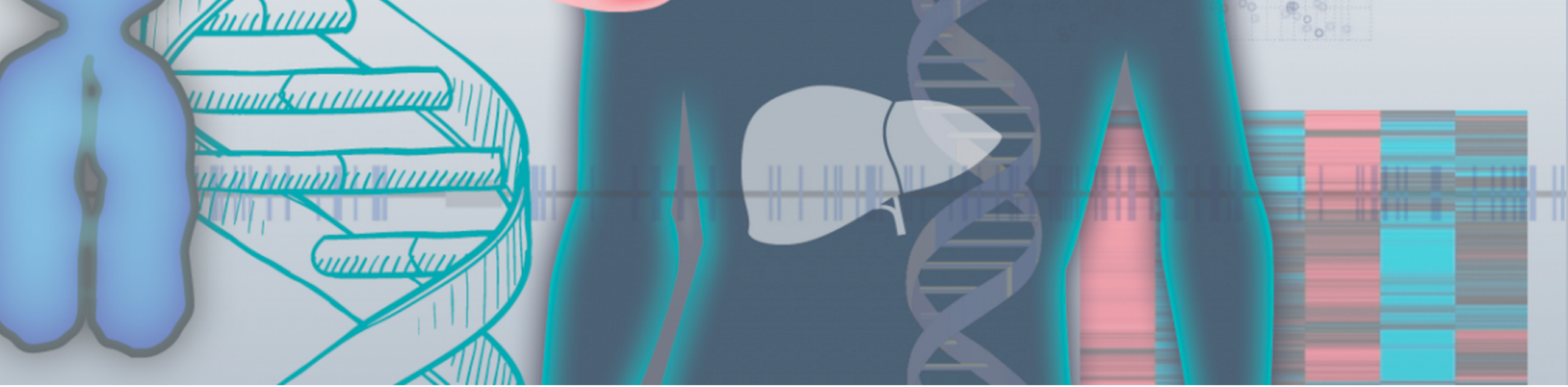


Harmonization and integration of RNA-seq datasets across cohorts: GTEx & Kids First case study

CFDE Steering Committee meeting :: 10/12/2021

François Aguet

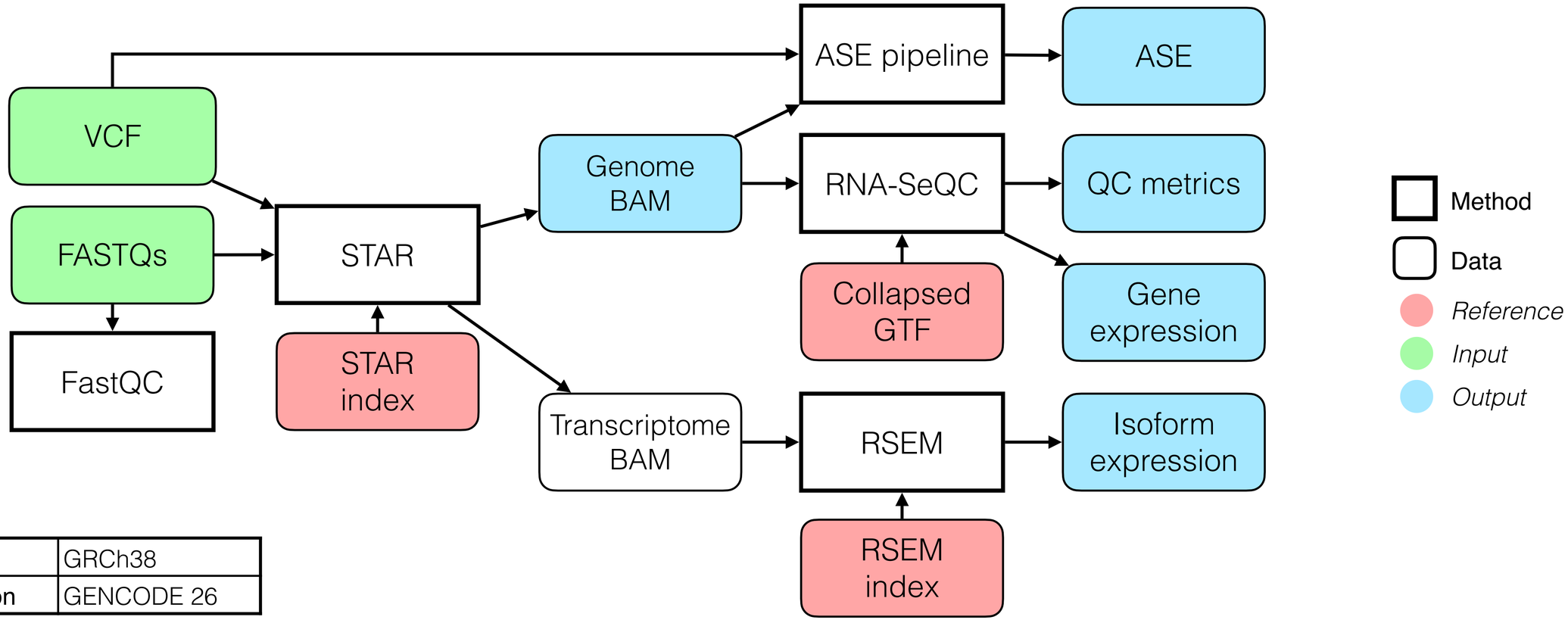
BROAD INSTITUTE

GTEx

Outline

- Response to NIH CFDE request to develop Scientific Use Case(s) with other DCCs
- ***Specific Aim***: Working with the Kids First DCC, we proposed an integrative comparison of 'matched' GTEx and KF RNA-seq datasets of interest to our user communities who want to compare gene expression from pediatric tumors against a GTEx reference normal.
- ***Goal:*** assess considerations for interoperability (i.e., integration and comparison) of RNA-seq datasets from different experiments/DCCs/etc.
  - Technical harmonization vs. biological differences
  - Uniform processing with a harmonized pipeline
  - Impact of gene/transcript annotation: comparison of GENCODE versions

RNA-seq alignment, quality control, and quantification



References

Reference genome	GRCh38
Transcript annotation	GENCODE 26

Pipeline components

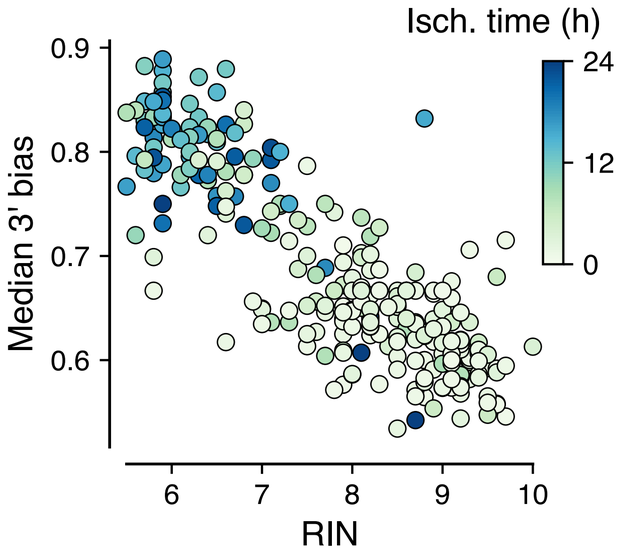
Alignment	STAR v2.7.5c
Gene expression	RNA-SeQC v2.4.2
Transcript expression	RSEM v1.3.3
QC metrics	RNA-SeQC v2.4.2

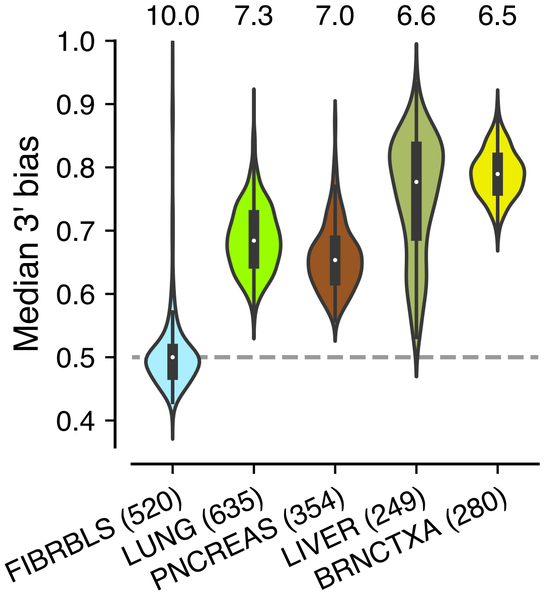
github.com/broadinstitute/gtex-pipeline

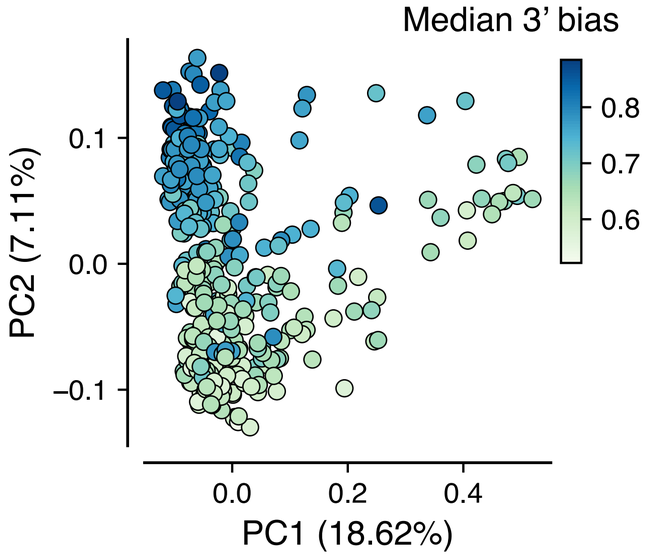
STAR: Dobin et al., *Bioinformatics*, 2013
RSEM: Li et al., *Bioinformatics*, 2010
RNA-SeQC 2: Graubert, Aguet et al., *Bioinformatics*, 2021

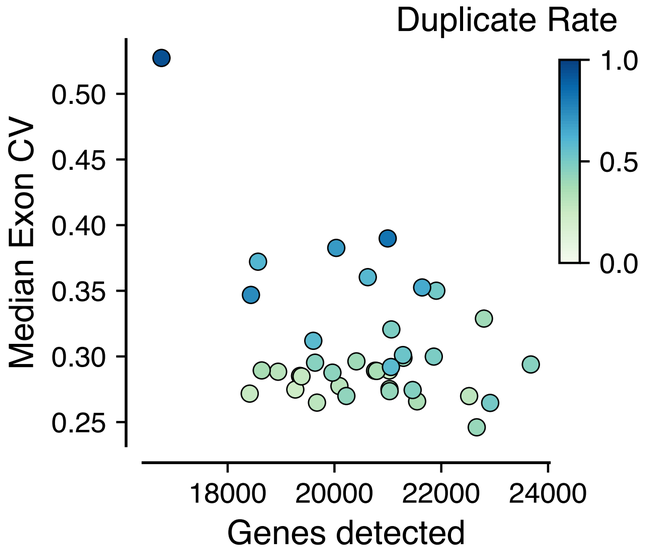
Assessing sample quality with RNA-SeQC metrics

- QC metrics are most informative in combination with sample metadata, e.g.:
  - Measures of RNA quality (RIN/ RQS, ischemic time)
- QC metrics yield insights into technical sources of expression variation
- QC metrics are frequently (anti)correlated, e.g.:
  - Coverage variability, duplication rate, and genes detected







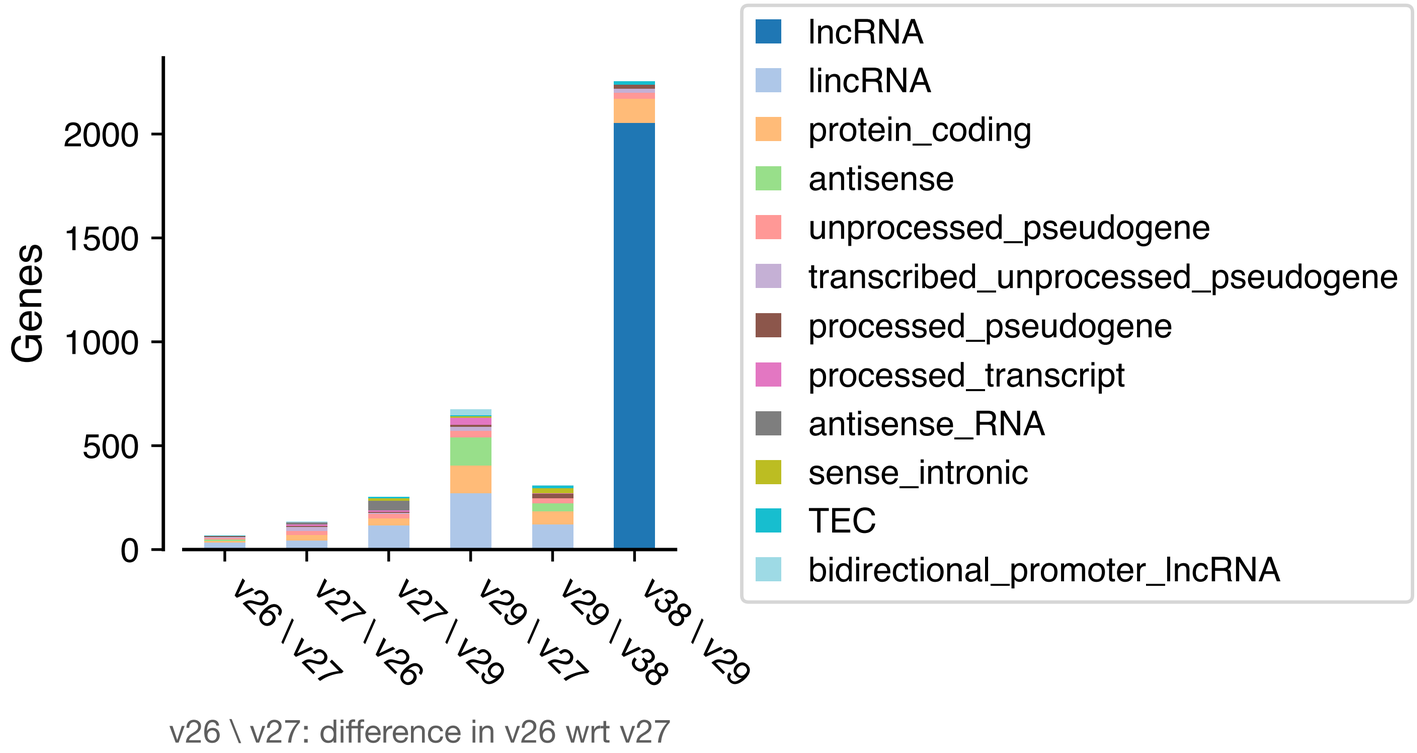


Graubert, Aguet, et al. *Bioinformatics*, 2021

Benchmarking effect of GENCODE version on expression quantification

- 10 GTEx samples (5 Adrenal Gland, 5 Cerebellum)
- 4 GENCODE versions:
  - v26 (GTEx v8)
  - v27 (KF)
  - v29 (ENCODE)
  - v38 (latest release)
- Processed with GTEx pipeline (RNA-SeQC and RSEM quantifications)

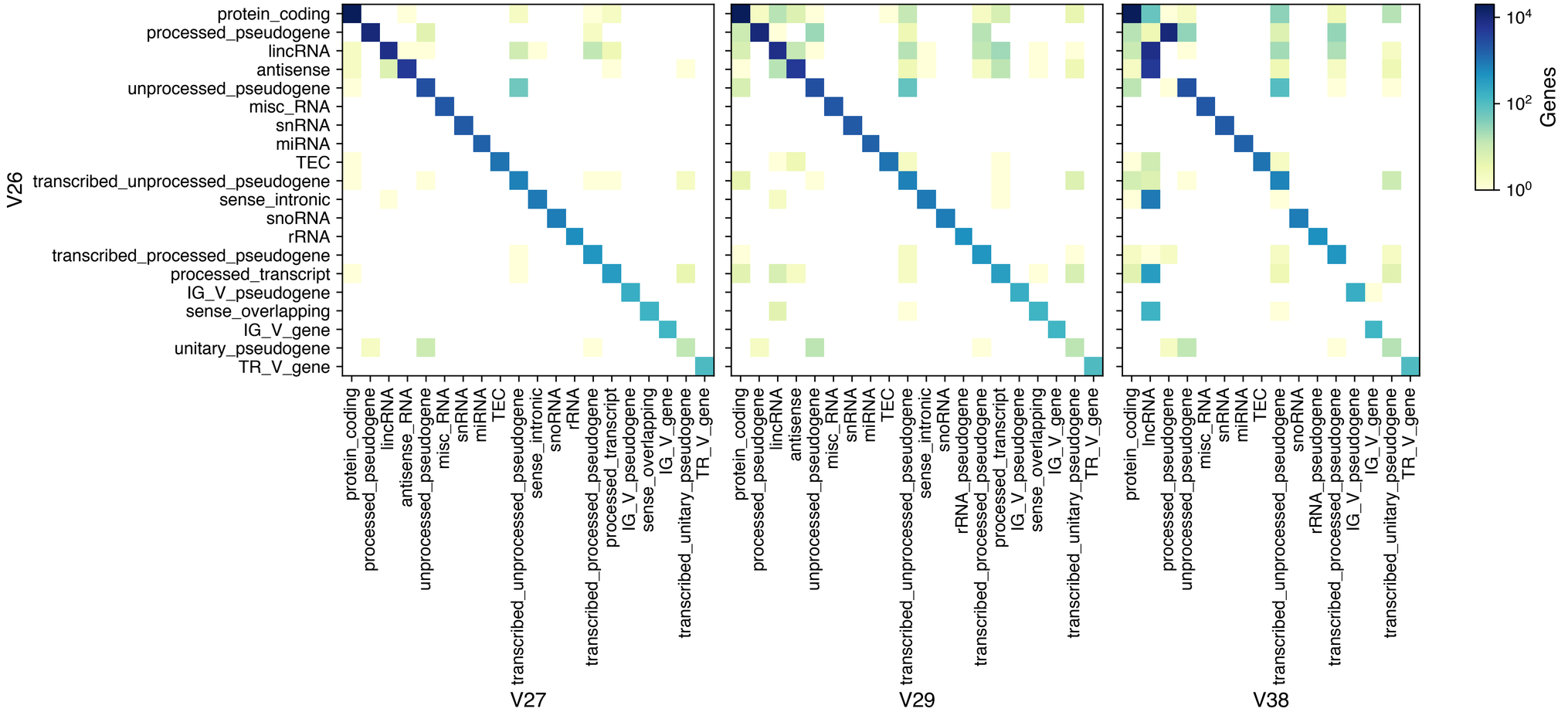
Differences between GENCODE versions



v26 \ v27: difference in v26 wrt v27

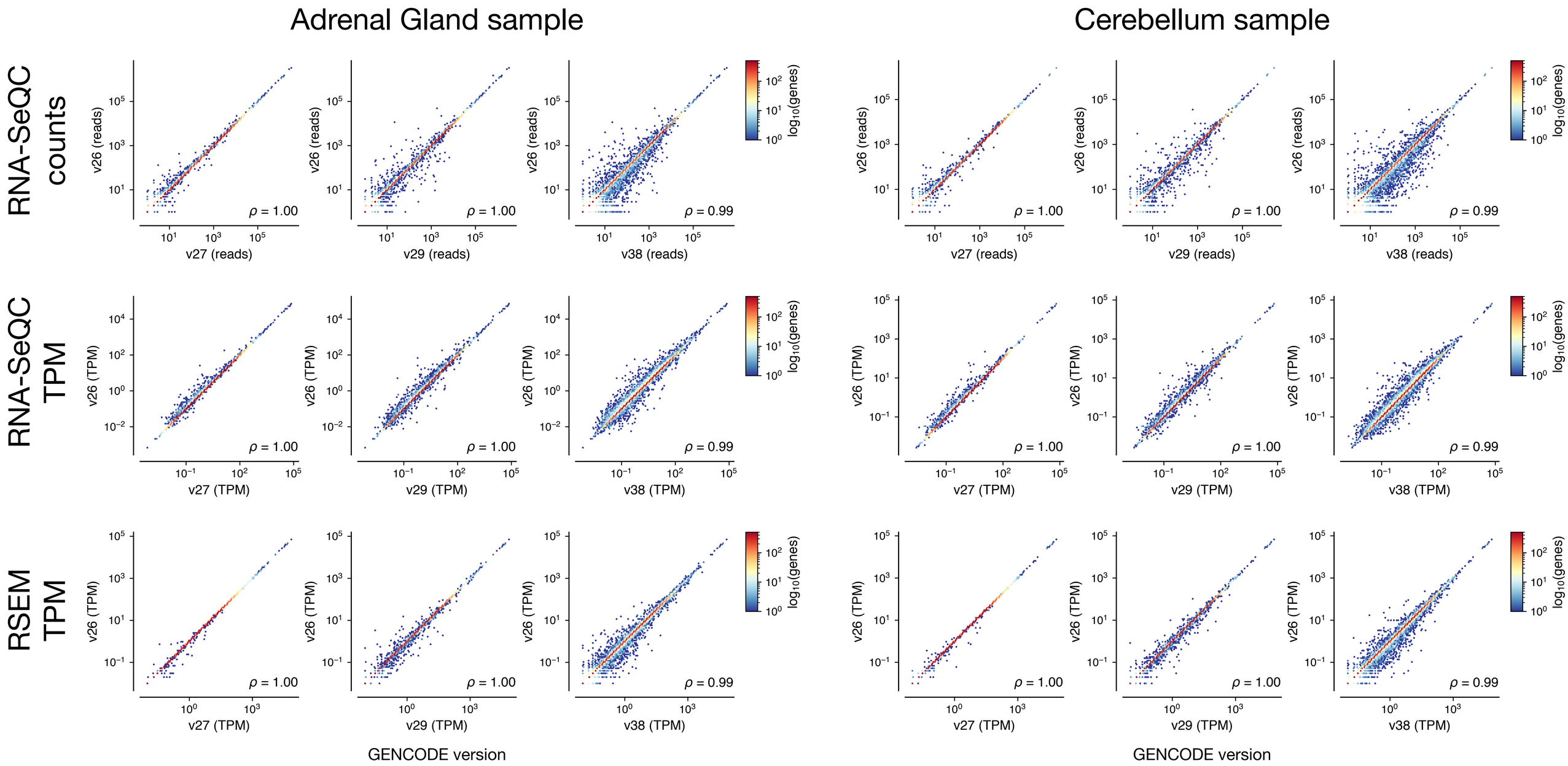
- Figure shows Ensembl gene IDs that are added/removed between versions
- The vast majority of changes affect long non-coding genes

Biotype annotation changes across GENCODE versions

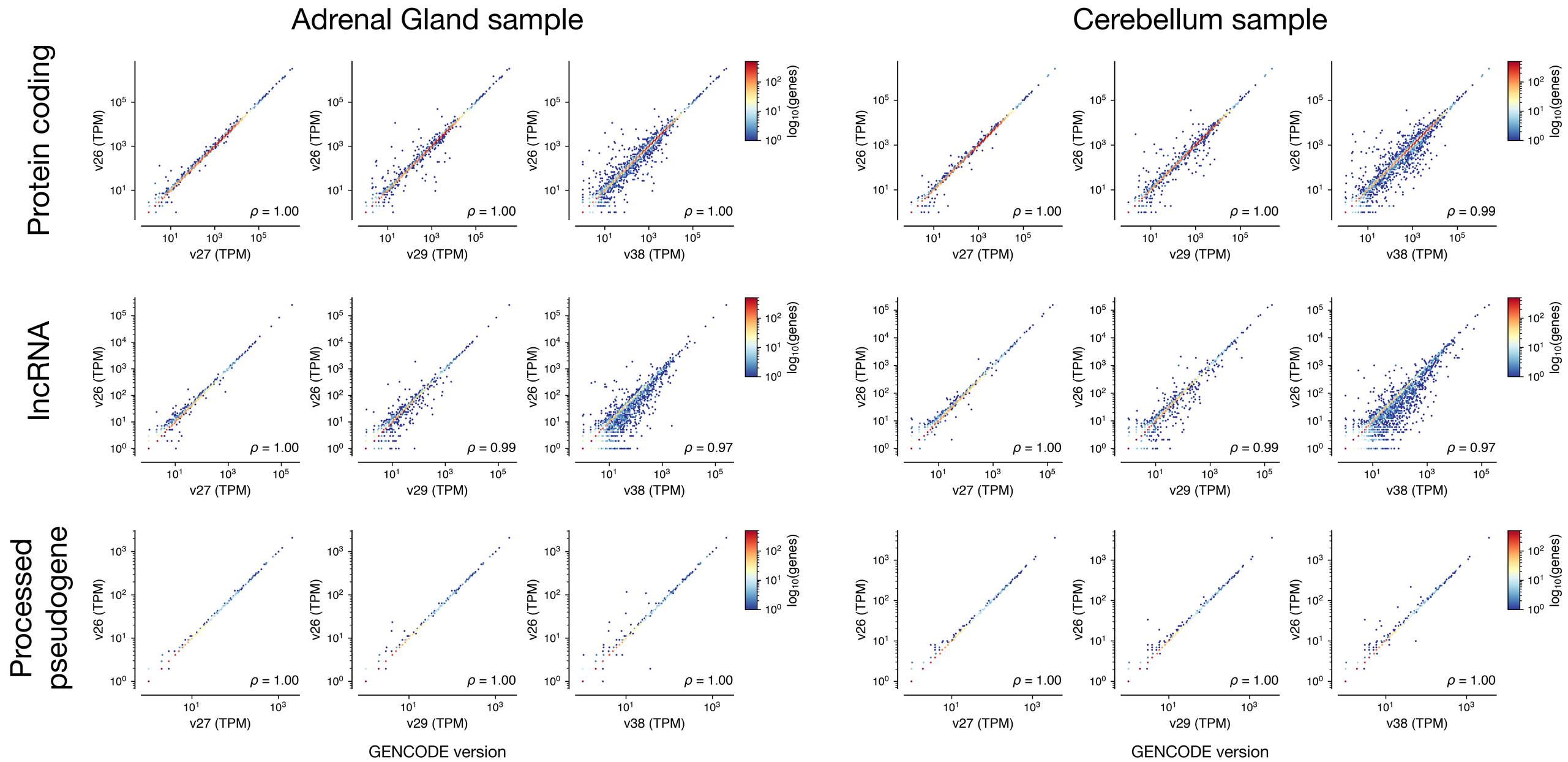


- In addition to changes in gene/transcript structure, the functional annotations of genes also varies across annotation versions
- The most significant recent change was a reannotation of several classes of noncoding RNAs as lncRNAs

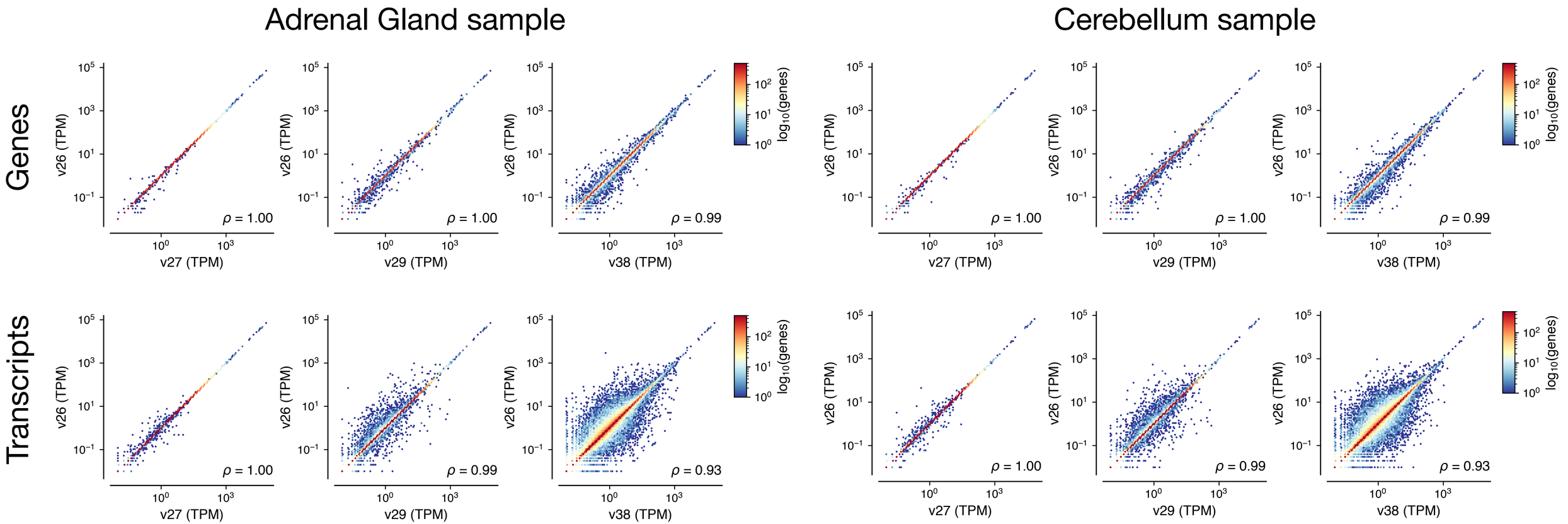
Effect of annotation version on expression quantification



Differences by biotype

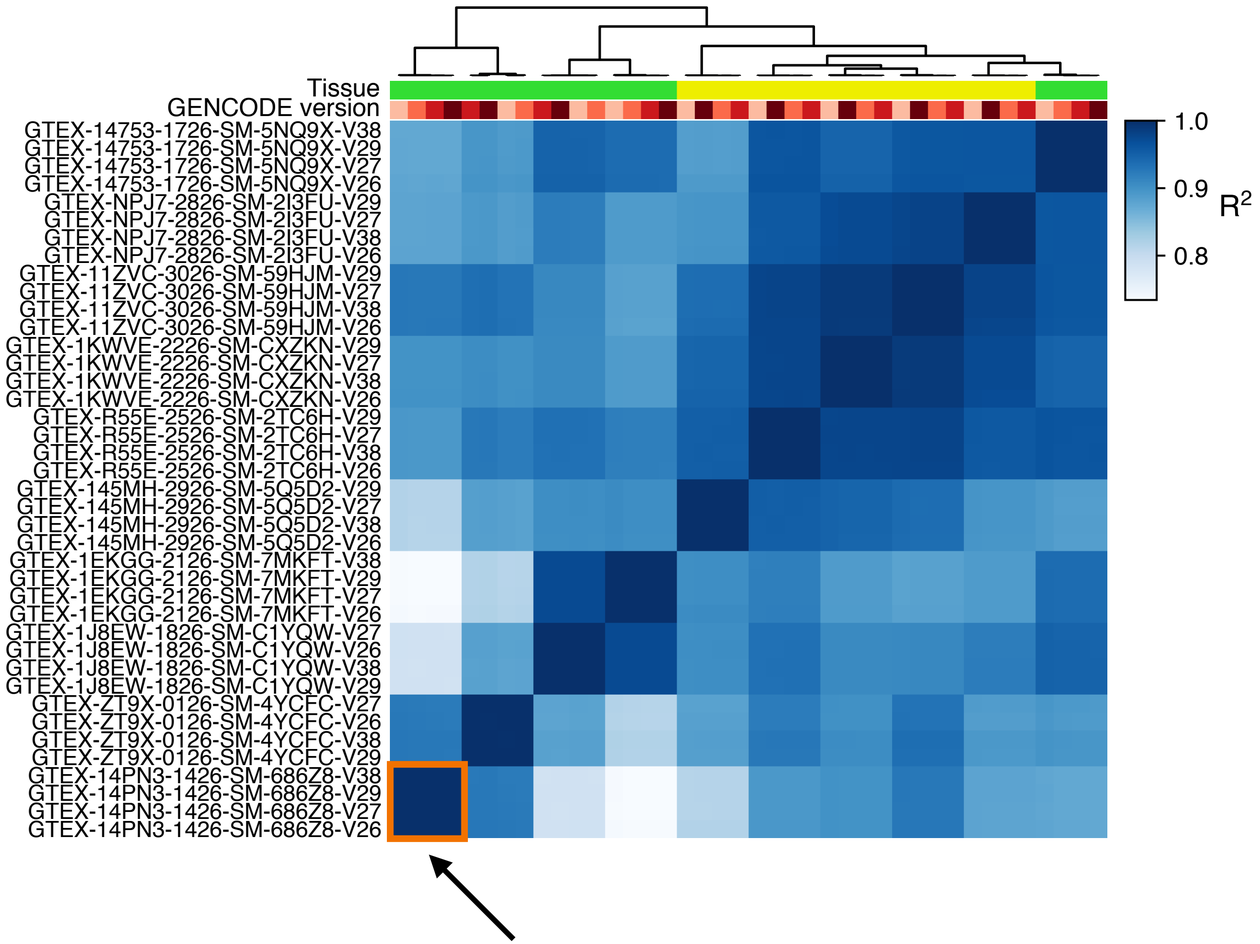


Effect on gene- vs transcript-level quantifications



- Most changes between annotation versions occur at the transcript level, e.g., addition/removal of isoforms, changes in 3’ UTRs, etc.
- Deconvolution of isoform-level expression from short read RNA-seq is highly sensitive to the accuracy of the annotation. Consequently, isoform expression estimates vary more strongly between versions.

Individual variation dominates annotation effects



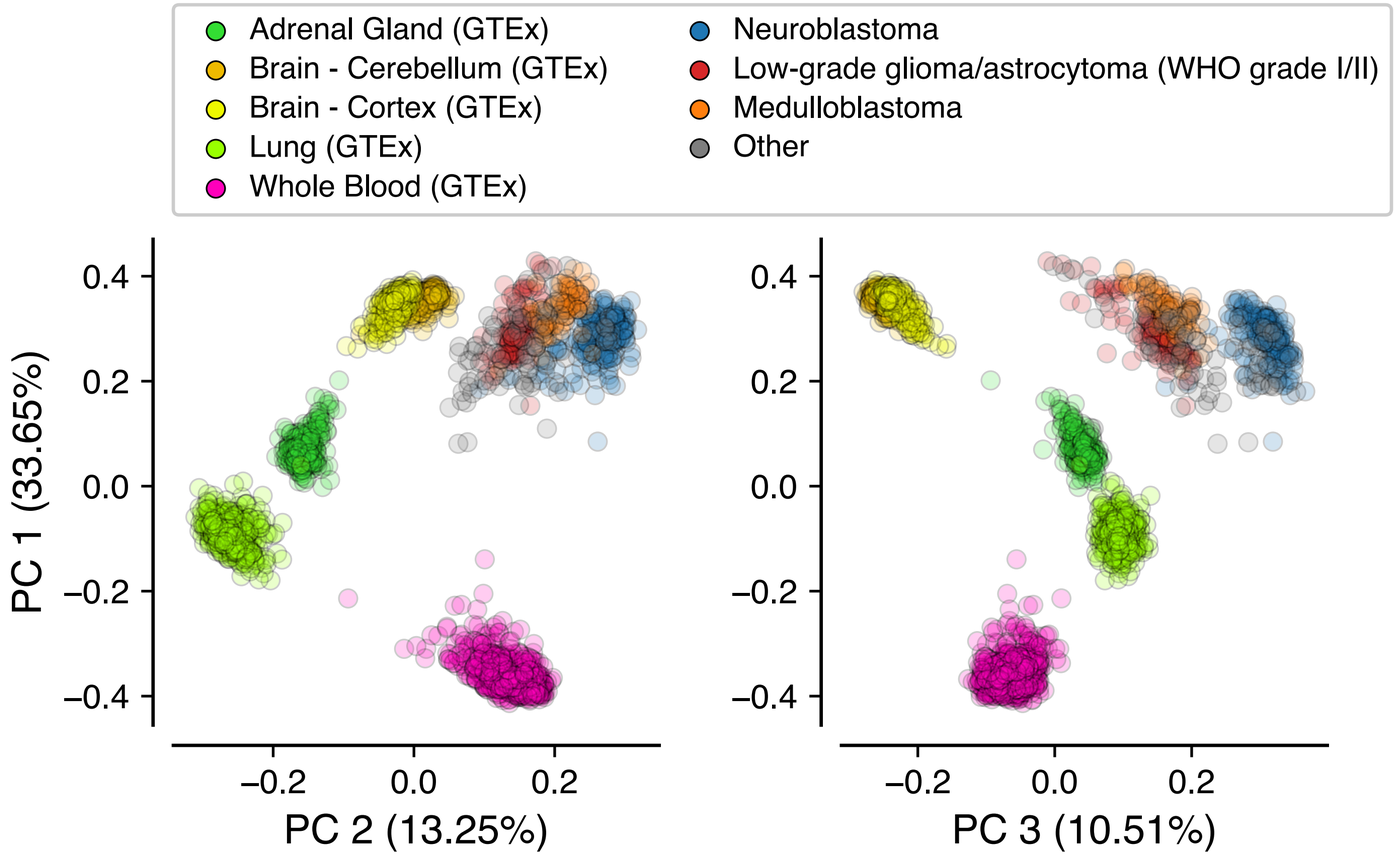
Expression profiles from the same sample processed with different GENCODE versions are highly correlated

- Sample-level* differences in gene expression dominate variation introduced by annotation differences
- Heterogeneity in cell type composition is typically the strongest contributor to expression variation
  - The dominant source of this heterogeneity is typically technical, due to (small) differences in tissue sampling

Comparison of GTEx and Kids First pipelines

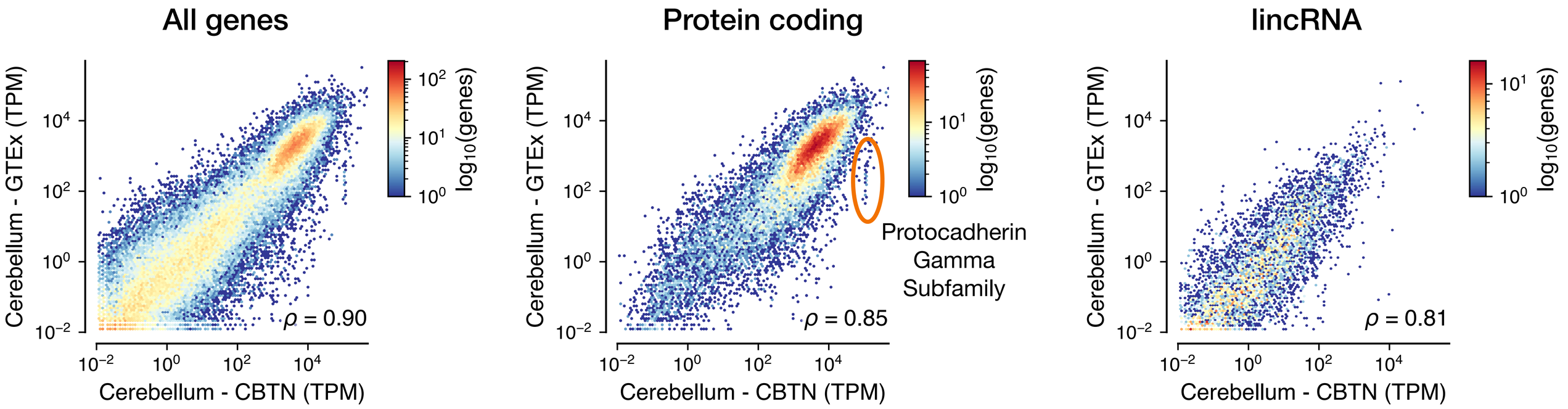
- Identical core methods (STAR, RSEM)
- Largely similar parameter settings, with some differences:
  - More stringent settings for alignment of spliced reads at annotated junctions in KF (GTEx uses same settings as ENCODE)
  - Chimeric alignment settings; these depend on read length etc. and may need to be adjusted for individual cohorts (e.g., tumor vs. normal)
- Results are essentially functionally equivalent in gene-level comparisons for samples processed with identical GENCODE versions using the two pipelines

Comparison of KF tumor samples with ‘matched’ GTEx samples



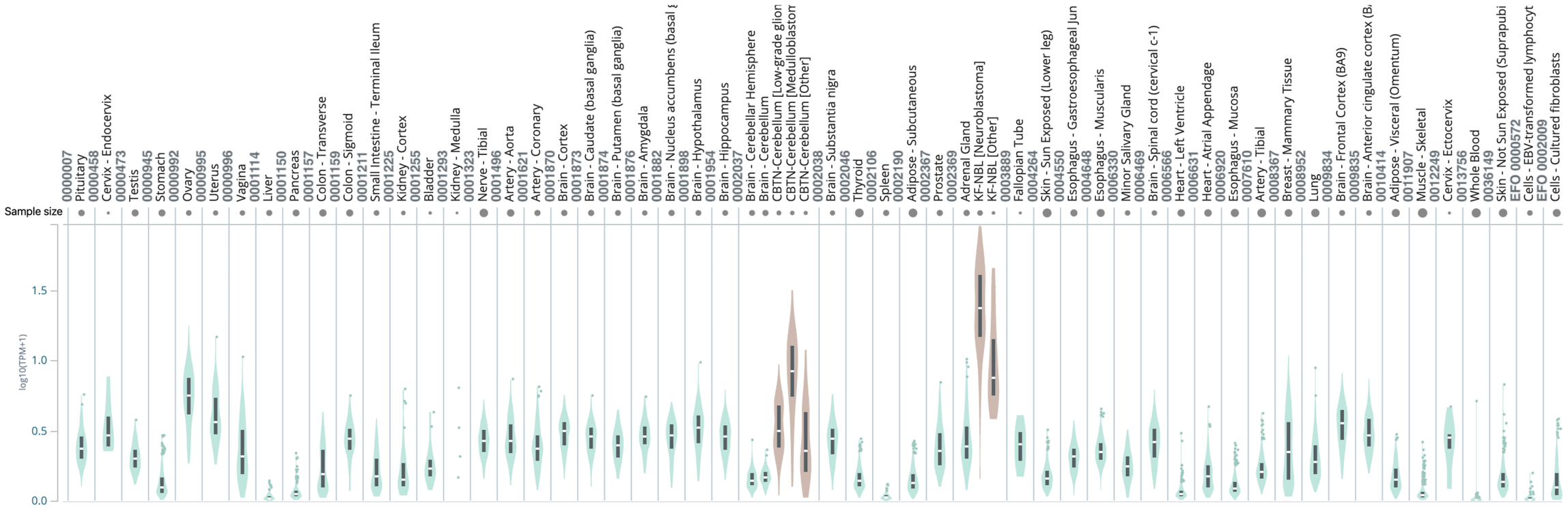
- Benchmarking dataset:
  - CBTN Cerebellum (n = 247)
  - GTEx Cerebellum (n = 241)
  - KF Neuroblastoma (n = 209)
  - GTEx Adrenal Gland (n = 258)
- To minimize processing batch effects, we reprocessed KF data with the GTEx V8 pipeline for these comparisons
- PCA shows expected differences between tumor and reference normal tissues
  - Part of this difference is compounded with differences in LC protocols (total RNA vs. polyA+)

Comparison of KF tumor samples with GTEx



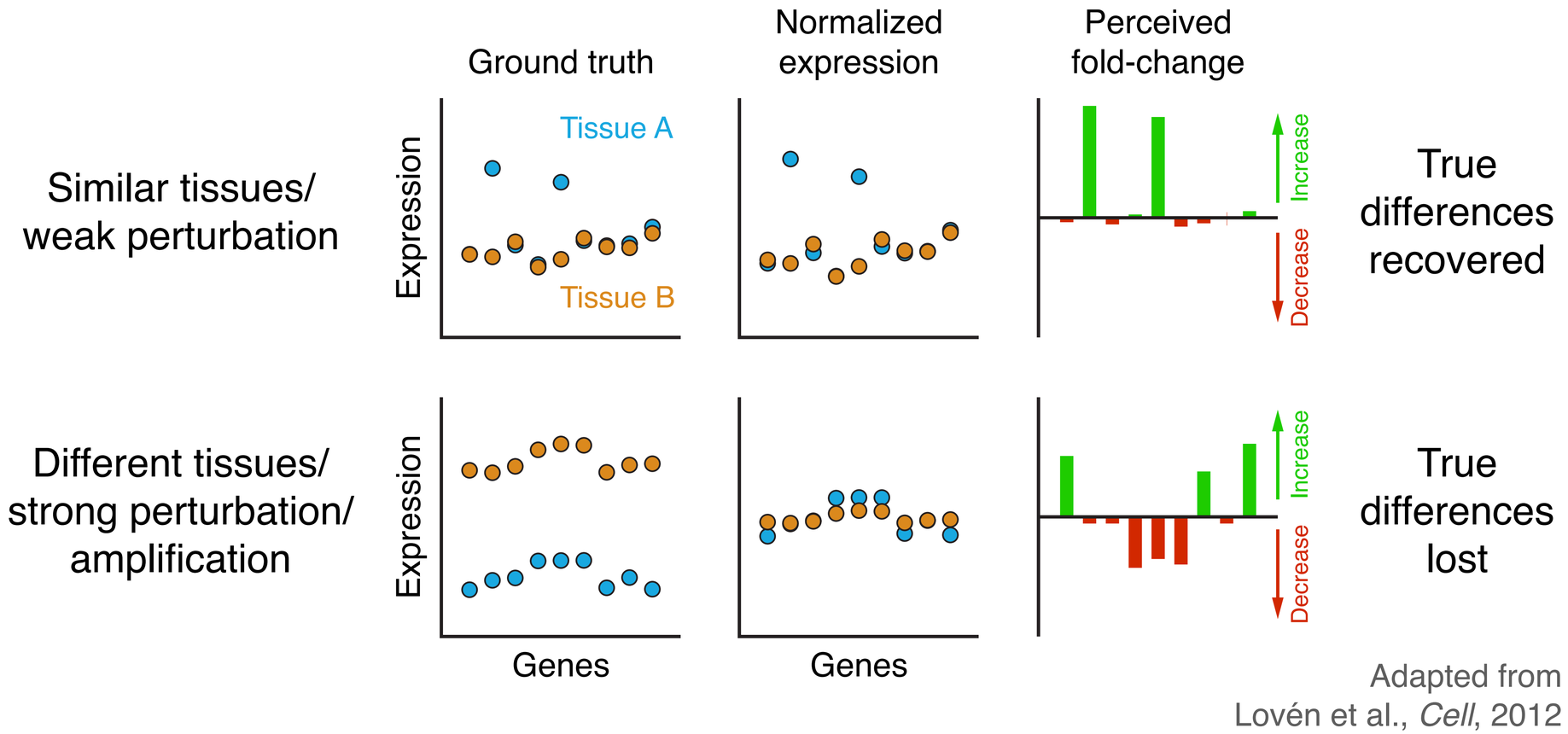
- Comparison of individual samples suggests that a majority of genes are differentially expressed in tumor vs normal comparisons
  - Quantifying these effects is challenging since the samples (and their constituting cell types/states) are fundamentally different
  - A subset of genes are clear outliers (e.g. over expression of protocadherin gamma genes in tumor samples)

Integrated visualizations of GTEx and KF datasets



- We enhanced GTEx Portal visualizations to jointly display and enable comparisons between datasets
- Interpretation is challenging: which expression differences between tumor and ‘matched’ normal samples are meaningful?

Limits of between-sample normalization



Adapted from Lovén et al., *Cell*, 2012

- Assumptions for between-sample normalization:
  - Majority of genes unchanged between samples/tissues/conditions
  - Comparable cell types and numbers, transcriptional output, etc. in each sample

Summary

- RNA-seq datasets can be highly variable between studies (differences in sequencing protocol, tissue sampling, tumor/normal, analysis pipelines, etc.)
  - Integrative analysis of resources requires minimizing technical sources of variability to enable biologically meaningful comparisons
  - Harmonized computational pipelines eliminate variability introduced by analyses
- Resources should be updated as reference annotations (e.g., GENCODE) continue to be improved (e.g., with long read sequencing)
  - The ideal frequency of such updates depends on analyses: isoform-level quantifications are most sensitive to annotation changes and will most benefit from annotation improvements
- Even with fully harmonized datasets, comparisons between different tissues or sample types (e.g., tumor vs normal) remain challenging due to differences in cell type composition, cell state, etc.
  - No ‘black box’ solutions

Outlook and recommendations

- Harmonized 'best practices' pipeline (with parameter settings based on GTEx/TOPMed etc.) and identical GENCODE version to process GTEx, Kids First, and other CFDE RNA-seq datasets
  - Make pipelines available to the community on a variety of platforms, including Terra and CAVATICA
- Coordinate plans for reprocessing data based on the upcoming GTEx V10 release (which will use the latest available GENCODE version)
- Provide guidelines for reprocessing datasets at intervals guided by significant changes in the GENCODE annotation
  - If possible, coordinate with GENCODE
  - Re-processing is mainly limited by costs:
    - The current implementation of the KF pipeline costs ~$2/sample (~$3000 for all samples)
    - In GTEx, the current cost is ~$1/sample (~$25,000 for all samples)
- Coordinate on future pipeline changes, including alignment to personal genomes for cohorts with WGS

Acknowledgments

Allison Heath
Adam Resnick
Deanne Taylor
Eric Wenger
Kai Wang
Chris Liu
Yu Hu
Adam Kraya
Komal Rathi
Run Jin

Kristin Ardlie
François Aguet
Jared Nedzel
Katherine Huang
Duyen Nguyen
Lan Nguyen
Joseph Okondo

George Papanicolaou