Gene expression datasets processed

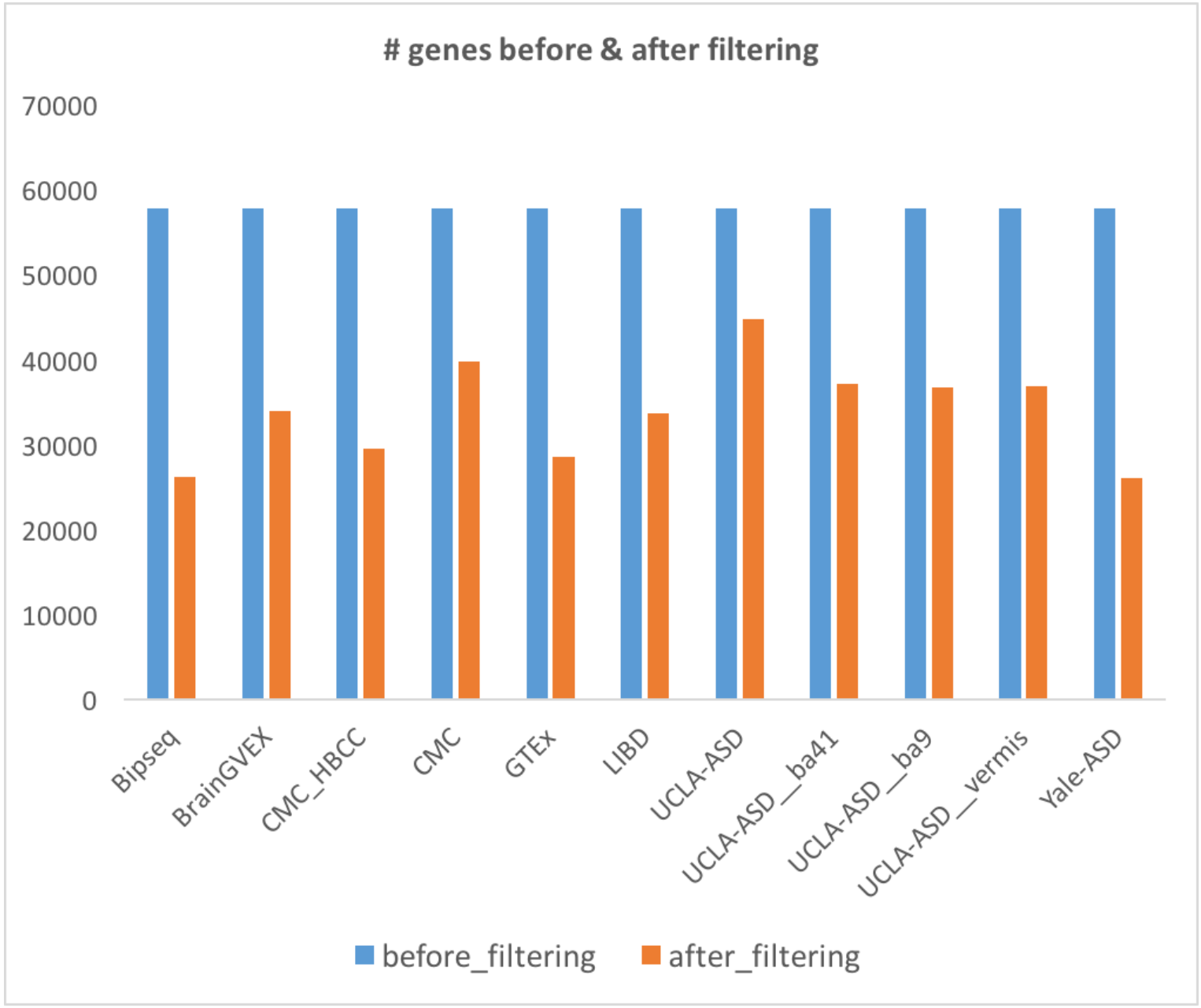
- CommonMind (CMC)
- NIMH Human Brain Collection Core (CMC HBCC collection)
- Lieber Institute for Brain Development (LIBD)
- Brain GVEX
- GTEx
- UCLA-ASD
  - UCLA-ASD__ba9
  - UCLA-ASD__ba41
  - UCLA-ASD__vermis
- Yale-ASD
- Bipseq (bipolar disorder)

For each dataset -- how much data do we lose by filtering out lowly-expressed genes?

Is such filtering even justified at all?

Standard GTEx filtering:
Genes must have at least
10 samples with:

- **RPKM > 0.1**
- raw **read counts > 6**



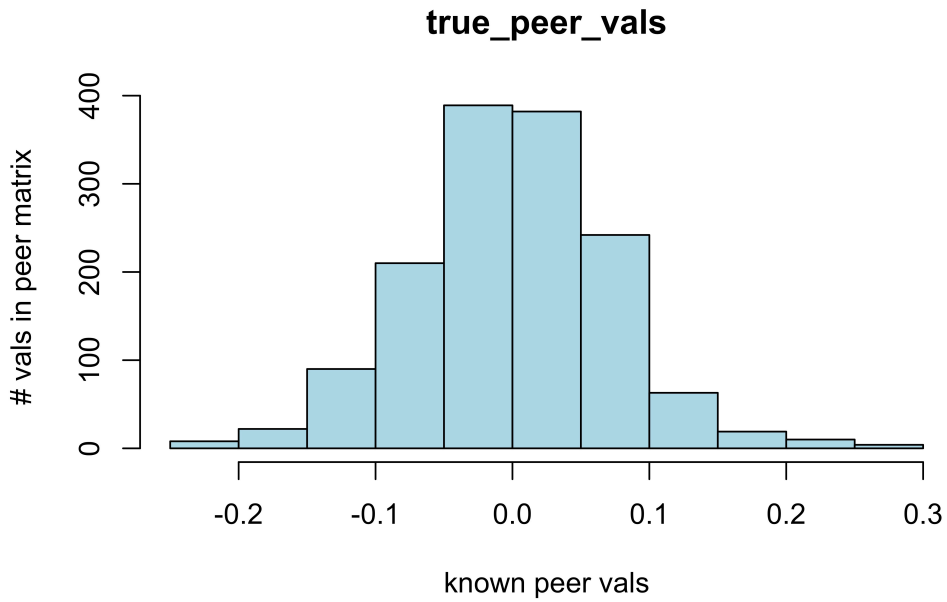
How well do our methods recapitulate those of established protocols?

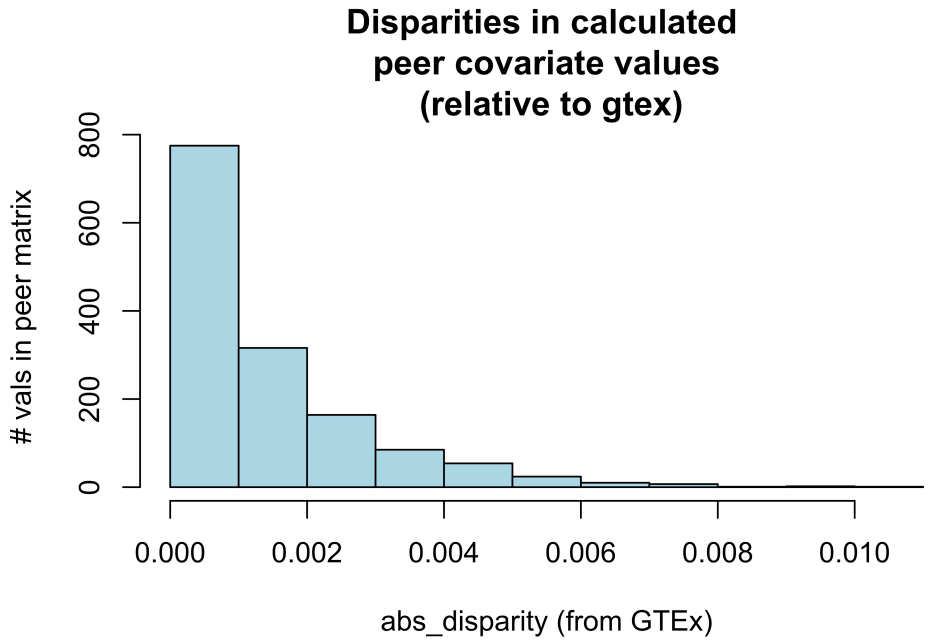
Perfect concordance for GTEx **normalized gene expression values**:

- min error: -4.4408920985e-15
- max error: 4.88498130835e-15

Very strong concordance for associated **PEER covariate values**:

- min error: -0.015
- max error: 0.019

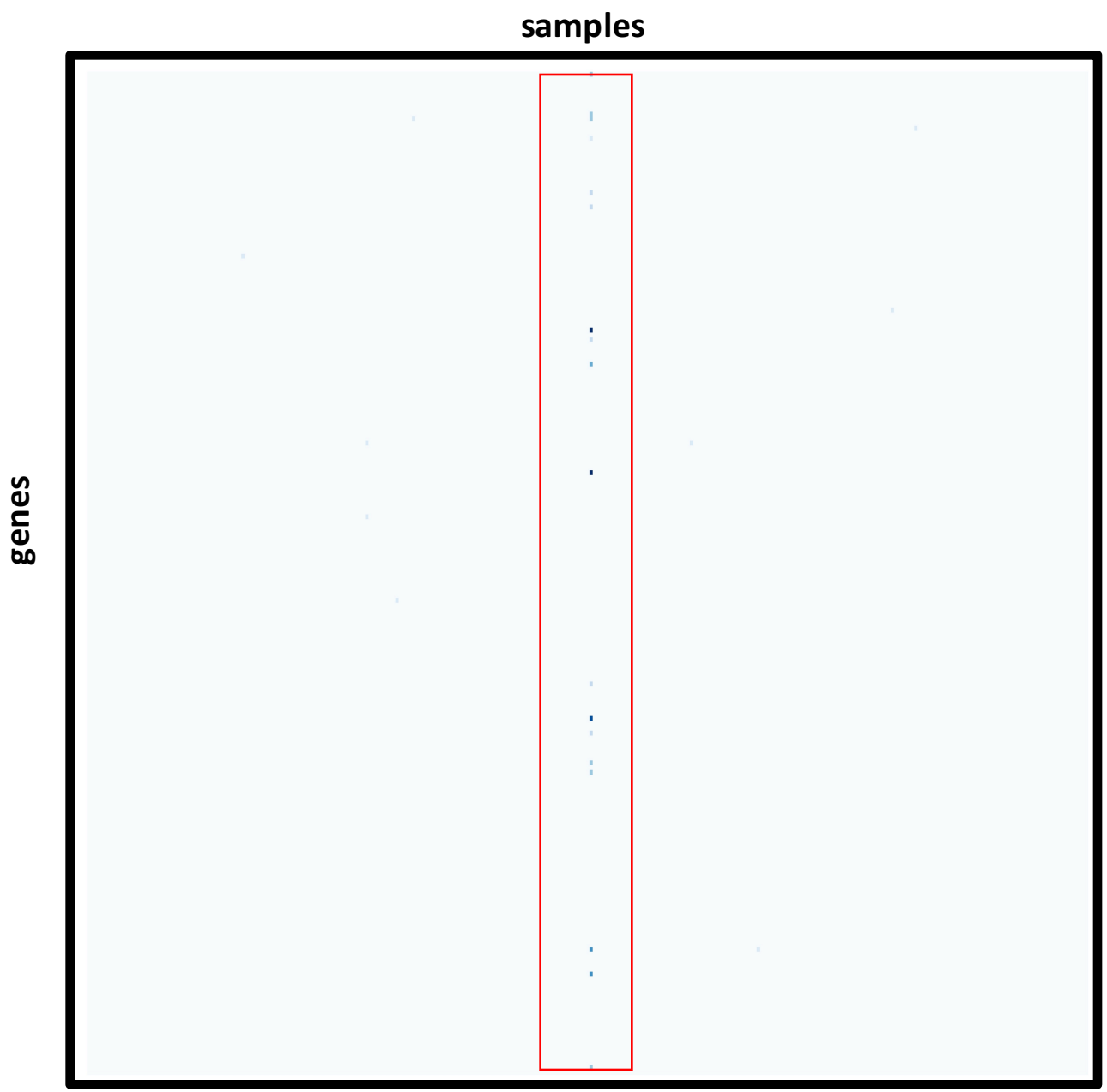




How does outlier removal affect normalized gene expression values?

For cmc -- compare final results to what happens w/outlier removal (SL has filtered the dataset to remove outliers)

603 (QCed) vs 613 (original) samples



Additional considerations:

- thresholds to use = ? (note GTEx parameters & heterogeneity in dataset sizes)
- no read count thresholds imposed
- fpkm vs rpkm
- # covariates to use = ? (15 -vs- 20 -vs- 25 factor covariate sets calculated)
- remove mitochondrial genes? -- there are very few of them