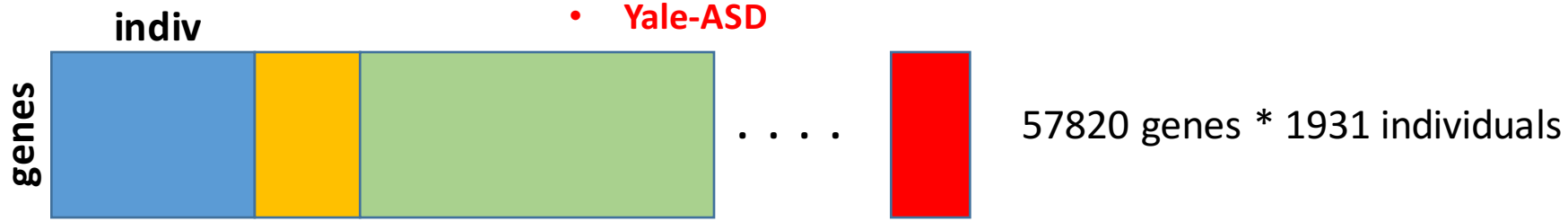
1) Sample_IDs → individual IDs [Bipseq, CMC_HBCC, LIBD]

2) Remove redundancy by selecting same individual from one dataset (prunes # samples from ~2200 to 1931)

3) Merge 9 gene expression datasets from 9 studies (very different dimensions for each!)

- **Bipseq**
- **BrainGVEX**
- **Brainspan**
- **NIMH Human Brain Collection Core**
- **CommonMind**
- **GTEx**
- **Lieber Institute for Brain Development**
- **UCLA-ASD**
- **Yale-ASD**



57820 genes * 1931 individuals

4) Gene expression normalization from FPKM data – applied to the ***conglomerated*** matrix

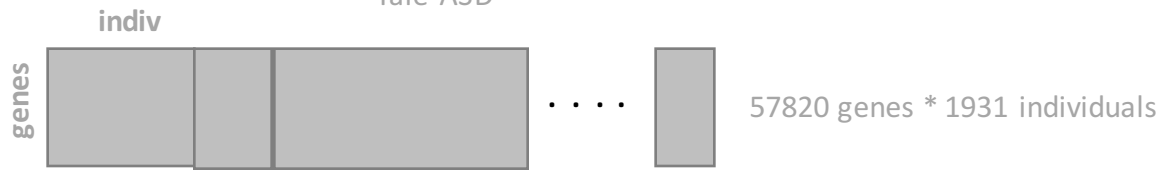
5) Generate covariate files (both known and hidden)

1) Sample_IDs → individual IDs [Bipseq, CMC_HBCC, LIBD]

2) Remove redundancy by selecting same individual from one dataset
(prunes # samples from ~2200 to 1931)

3) Merge 9 gene expression datasets from 9 studies (very different dimensions for each!)

- Bipseq
- BrainGVEX
- Brainspan
- NIMH Human Brain Collection Core
- CommonMind
- GTEx
- Lieber Institute for Brain Development
- UCLA-ASD
- Yale-ASD



57820 genes * 1931 individuals

threshold for # of individuals w/ FPKM > 0.1 for that gene	# genes
min_10	43886
min_50	38150
min_100	35416
min_150	33694

* 57820 gene IDs (pre-filtering)

4) Gene expression normalization from FPKM data – applied to the *conglomerated* matrix

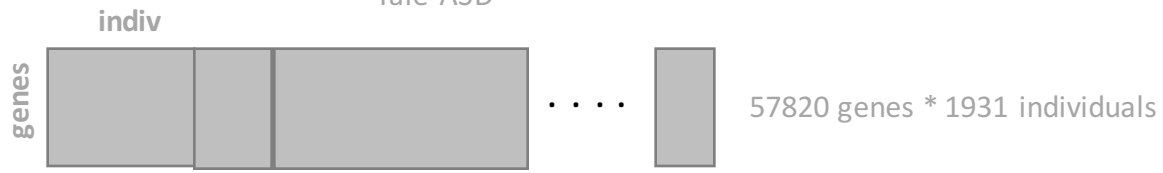
5) Generate covariate files (both known and hidden)

1) Sample_IDs → individual IDs [Bipseq, CMC_HBCC, LIBD]

2) Remove redundancy by selecting same individual from one dataset
(prunes # samples from ~2200 to 1931)

3) Merge 9 gene expression datasets from 9 studies (very different dimensions for each!)

- Bipseq
- BrainGVEX
- Brainspan
- NIMH Human Brain Collection Core
- CommonMind
- GTEx
- Lieber Institute for Brain Development
- UCLA-ASD
- Yale-ASD



57820 genes * 1931 individuals

4) Gene expression normalization from FPKM data – applied to the ***conglomerated*** matrix

5) Generate covariate files (both known and hidden)

Bipseq	63
BrainGVEX	429
Brainspan	6
CMC	603
CMC_HBCC	349
GTEx_DFC	116
LIBD	257
UCLA-ASD_DFC	84
Yale-ASD_DFC	24

Affective	8
Autism	44
Bipolar	217
Control	1104
Schizophrenia	558

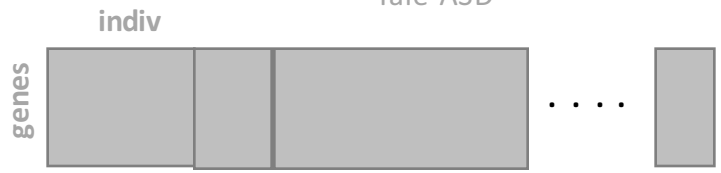
Female	685
Male	1246

1) Sample_IDs → individual IDs [Bipseq, CMC_HBCC, LIBD]

2) Remove redundancy by selecting same individual from one dataset
(prunes # samples from ~2200 to 1931)

3) Merge 9 gene expression datasets from 9 studies (very different dimensions for each!)

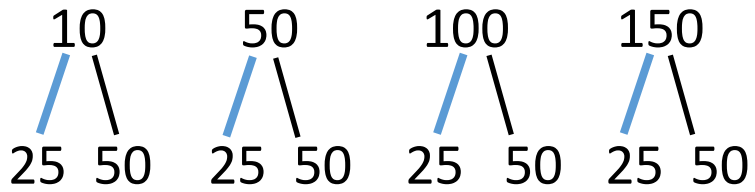
- Bipseq
- BrainGVEX
- Brainspan
- NIMH Human Brain Collection Core
- CommonMind
- GTEx
- Lieber Institute for Brain Development
- UCLA-ASD
- Yale-ASD



57820 genes * 1931 individuals

4 different normalized gene expression profiles (by threshold for # of individuals w/FPKM > 0.1 for that gene)

PEER Factors



4) Gene expression normalization from FPKM data – applied to the ***conglomerated*** matrix

5) Generate covariate files (both known and hidden)