ENCODE variant effects

Thousands of samples, hundreds of unique individuals

- **~1,489** DNase I datasets (ENCODE2/3/4) (more to come…)
    - Single-end, paired-end, Solexa GA1, NovaSeq data
- **326** cell types (not including stimulation states); most from primary cells/tissues
- **~496** distinct individuals (genotypes) (including the canonical ENCODE cell lines; K562, GM12878, HepG2, etc.)
- Combining ‘personalized’ genomes with biochemical ENCODE assays provides insight into how individual regulatory variants impact chromatin and gene regulation
- We don’t have full genome sequencing for all of these individuals in ENCODE, but ENCODE assay provide high depth sequence coverage at regulatory DNA (i.e., DNase I data can be mined for regulatory alleles → ‘regulotyping’)

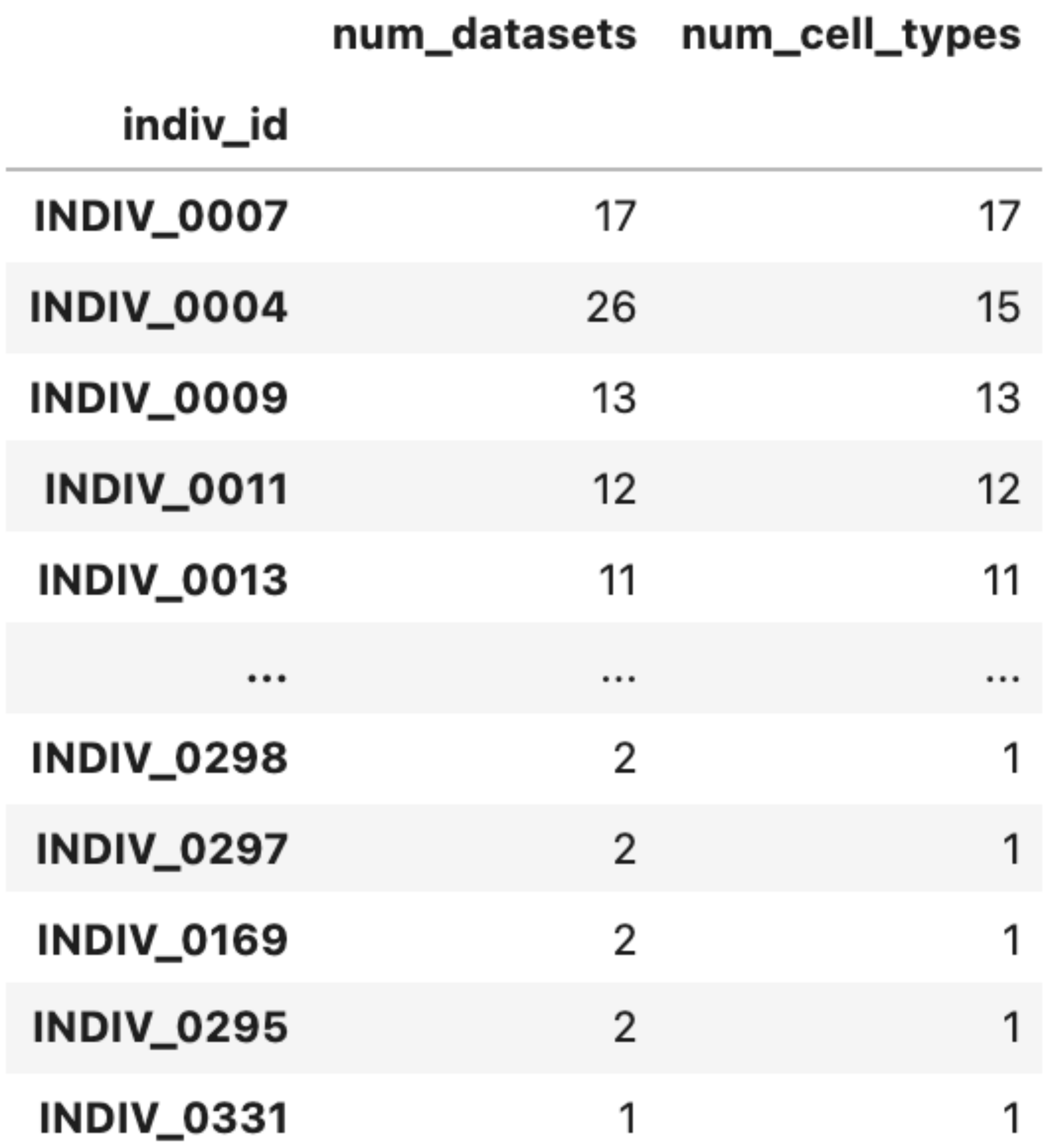
Genotyping from DNase I data

- We implemented a ‘bcftools’ based genotyping pipeline
  - http://github.com/jvierstra/nf-genotyping
  - We start by genotyping all datasets (n=1,489) individually
  - From these rough genotypes we determine kinship using the KING method
- We then merge datasets (BAM files) derived from same individuals and perform a more comprehensive genotyping run using the same pipeline.
- At least 12 reads to call genotype; heterozygous call require at least 4 on alternative allele
- Indels are not considered

Genotyping from DNase I data

- ***3.1 million SNVs genotyped in DHS***
    - ***~50K per individual***
    - ***~28K per dataset***
- Median read depth per sample: ~100 million
- Ts/Tv ratio ~2.17
- Genotypes call per individual, hets/homs
- In progress: Compare to WGS or Hi-C approaches (in another project we have compare to SNP array/imputation with >99% concordance; though less sensitivity as expected)

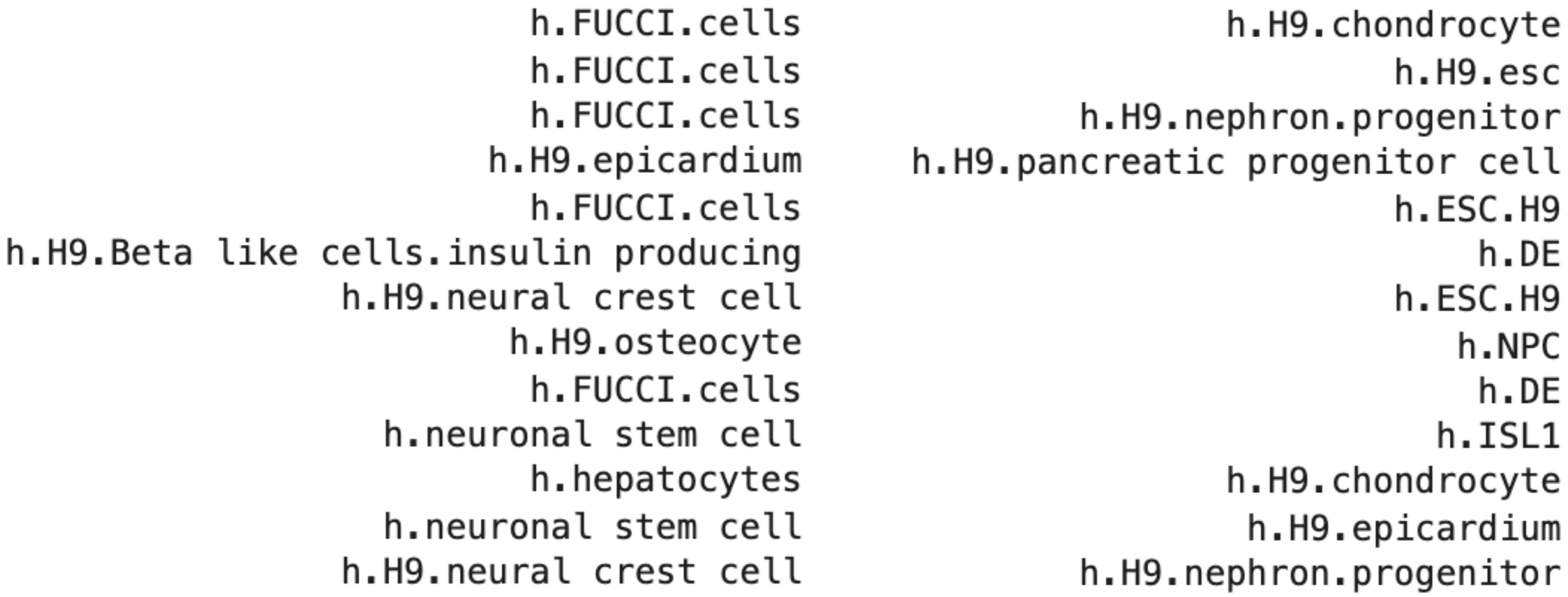
Genotyping from DNase I data



	num_datasets	num_cell_types
indiv_id		
INDIV_0007	17	17
INDIV_0004	26	15
INDIV_0009	13	13
INDIV_0011	12	12
INDIV_0013	11	11
...	...	...
INDIV_0298	2	1
INDIV_0297	2	1
INDIV_0169	2	1
INDIV_0295	2	1
INDIV_0331	1	1

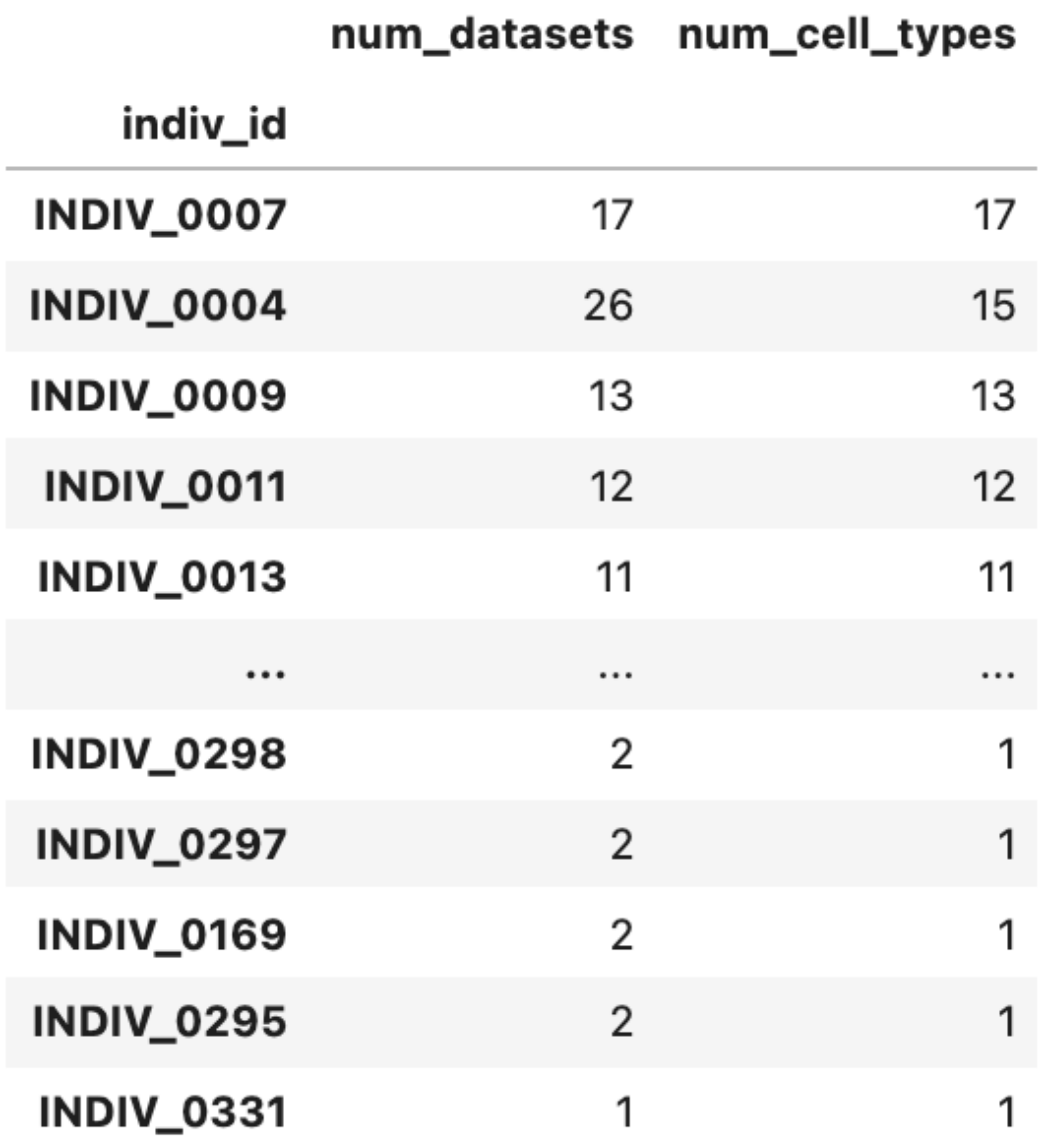
- 95 individuals with 3+ unique cell types
- 64 cell types with 3+ unique individuals

ES cell differentiations



```
                       h.FUCCI.cells                h.H9.chondrocyte
                       h.FUCCI.cells                        h.H9.esc
                       h.FUCCI.cells         h.H9.nephron.progenitor
                     h.H9.epicardium  h.H9.pancreatic progenitor cell
                       h.FUCCI.cells                        h.ESC.H9
h.H9.Beta like cells.insulin producing                          h.DE
                  h.H9.neural crest cell                    h.ESC.H9
                     h.H9.osteocyte                            h.NPC
                       h.FUCCI.cells                            h.DE
                 h.neuronal stem cell                         h.ISL1
                       h.hepatocytes                h.H9.chondrocyte
                 h.neuronal stem cell                h.H9.epicardium
                  h.H9.neural crest cell     h.H9.nephron.progenitor
```

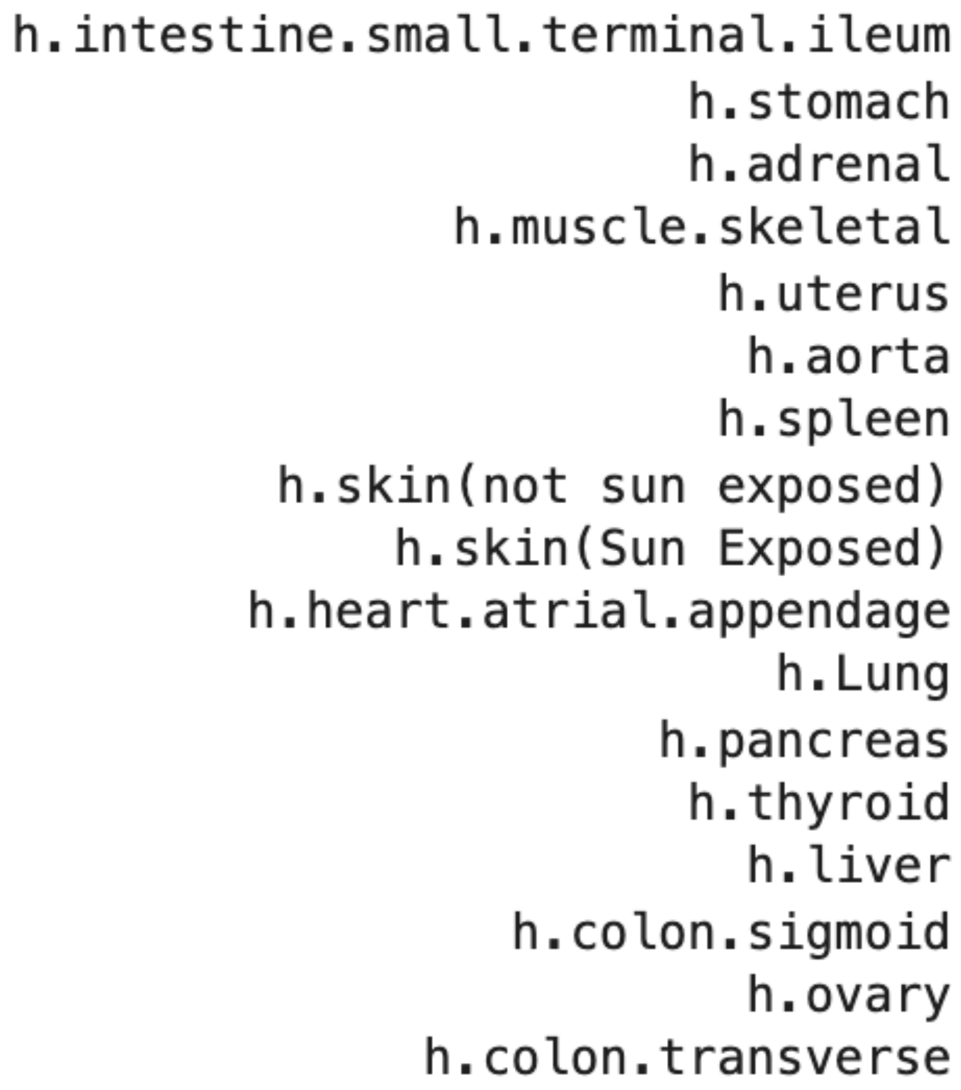
Genotyping from DNase I data



	num_datasets	num_cell_types
indiv_id		
INDIV_0007	17	17
INDIV_0004	26	15
INDIV_0009	13	13
INDIV_0011	12	12
INDIV_0013	11	11
...	...	...
INDIV_0298	2	1
INDIV_0297	2	1
INDIV_0169	2	1
INDIV_0295	2	1
INDIV_0331	1	1

- 95 individuals with 3+ unique cell types
- 64 cell types with 3+ unique individuals

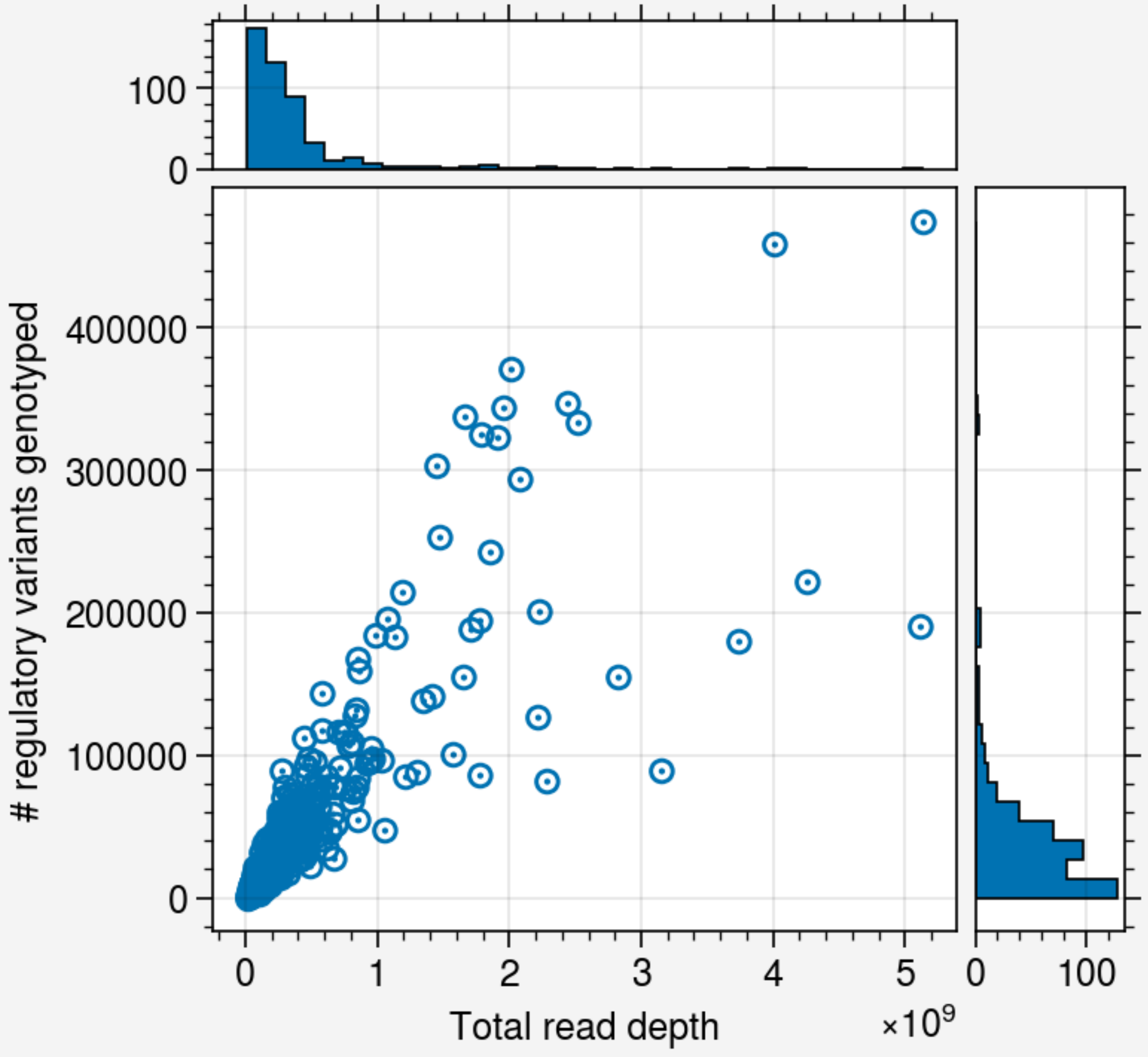
ENTex



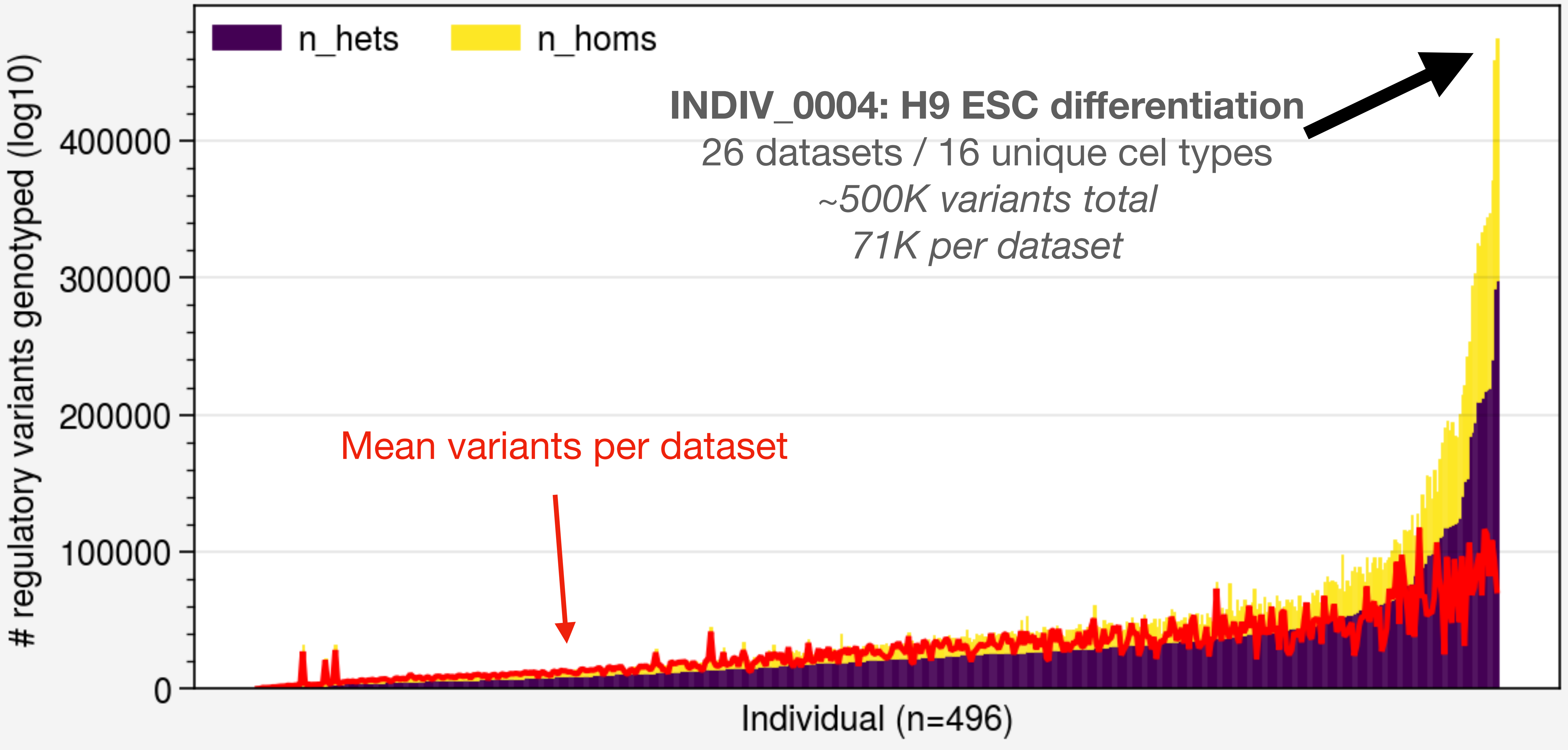
```
h.intestine.small.terminal.ileum
                       h.stomach
                       h.adrenal
               h.muscle.skeletal
                        h.uterus
                         h.aorta
                        h.spleen
         h.skin(not sun exposed)
             h.skin(Sun Exposed)
        h.heart.atrial.appendage
                          h.Lung
                      h.pancreas
                       h.thyroid
                         h.liver
                 h.colon.sigmoid
                         h.ovary
              h.colon.transverse
```

Genotyping from DNase I data

Variants by read depth



Variants called per individual



INDIV_0004: H9 ESC differentiation

26 datasets / 16 unique cel types

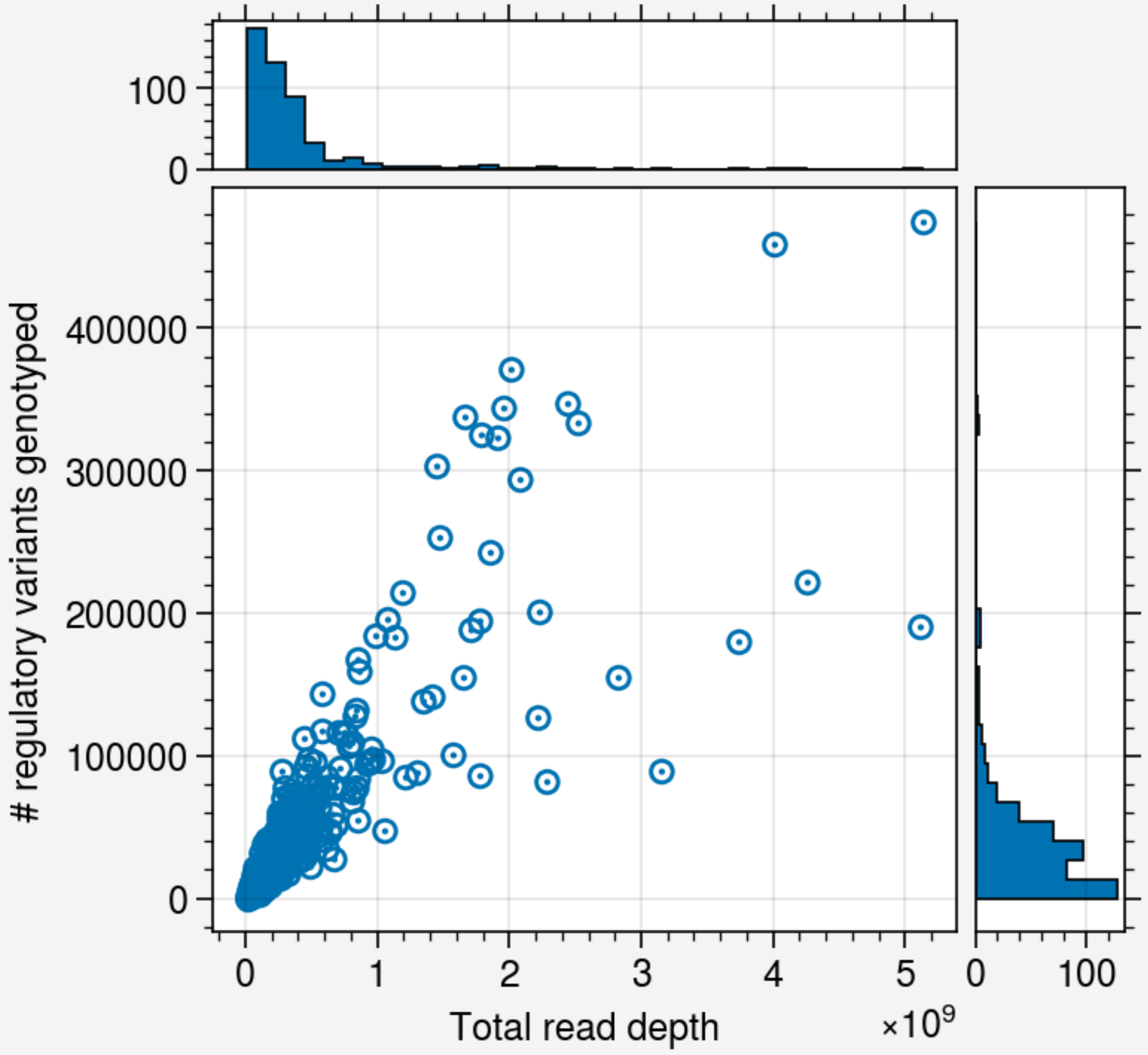
~500K variants total

71K per dataset

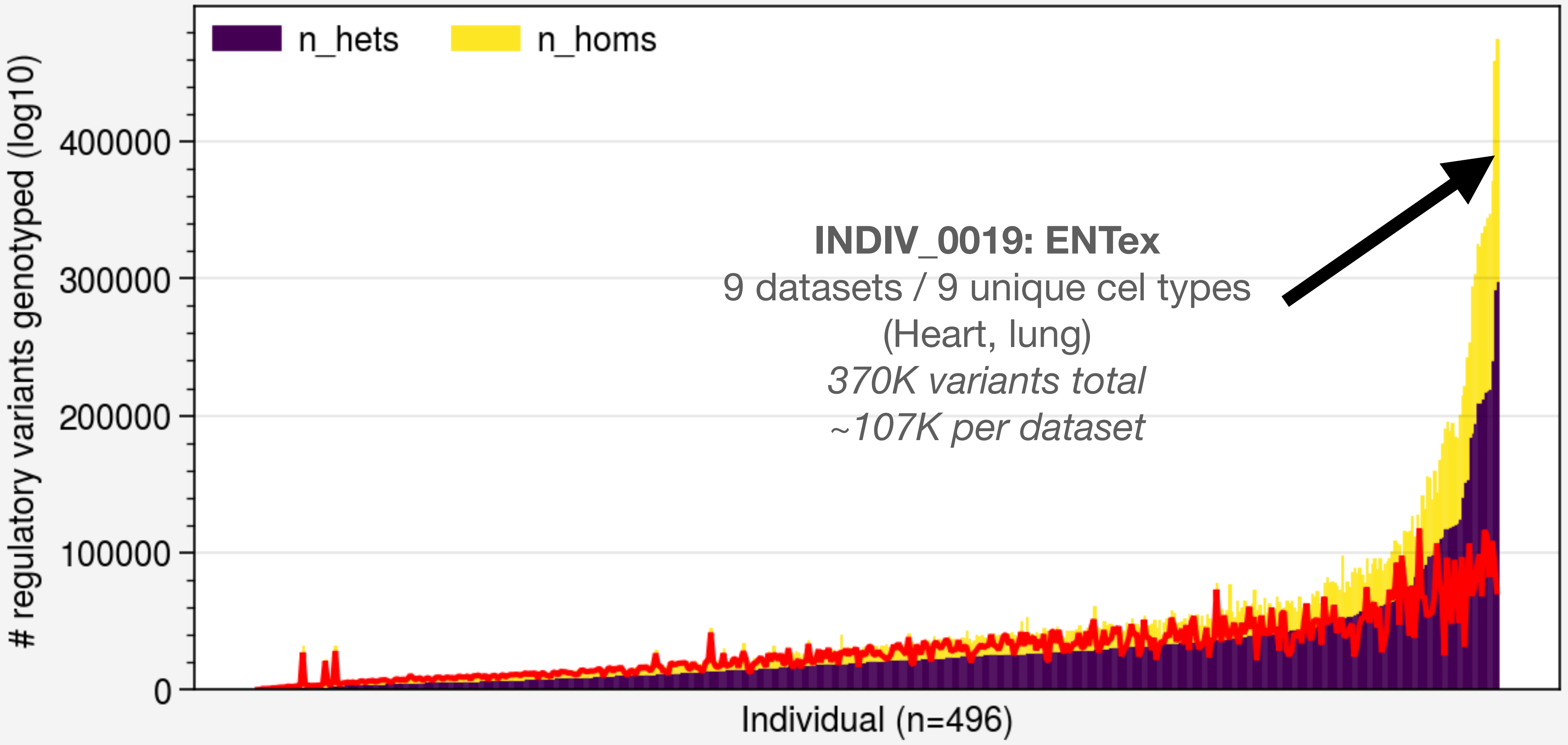
Mean variants per dataset

Genotyping from DNase I data

Variants by read depth



Variants called per individual



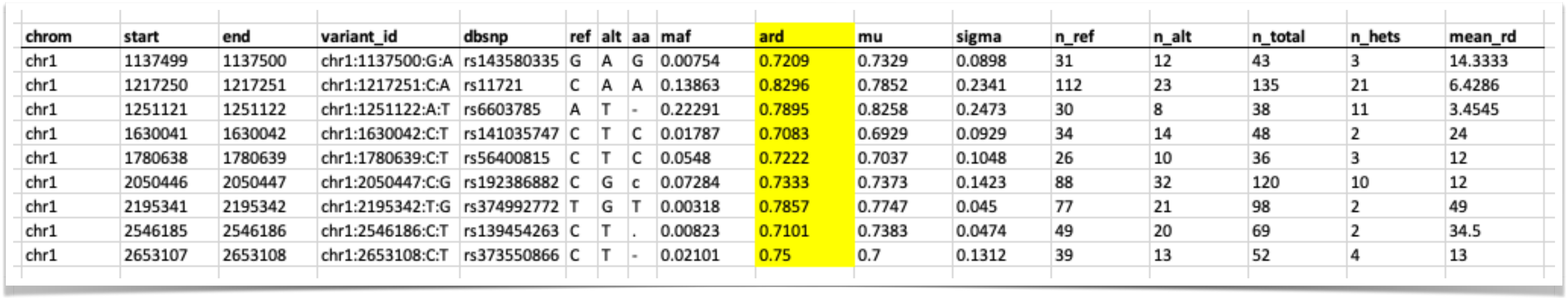
INDIV_0019: ENTex
9 datasets / 9 unique cel types
(Heart, lung)
370K variants total
~107K per dataset

Resolving read to individual alleles

- We have a pipeline that performs allele specific mapping of each dataset-individual pair.
- http://github.com/jvierstra/nf-genotyping
- Based on WASP (a method to remove mapping bias)
  - Finds reads that overlap a variant and creates ‘synthetic’ reads with containing reference and the alternate allele
  - These synthetic reads are then remapped to the genome (using bwa) and if an allele causes a mapping artifact for a particular read (i.e., maps to a new location) the reads is removed from downstream analysis.
- We additionally remove variants in which >10% of the reads are subject to mapping bias
- Data in VCF format

Resolving read to individual alleles

- For each heterozygous site we compute “ARD” (allelic read depth)
  - 1 = all reads reference allele
  - 0 = all reads alternate allele
- Table for each datasets, combined by cell type or individual (or overall)



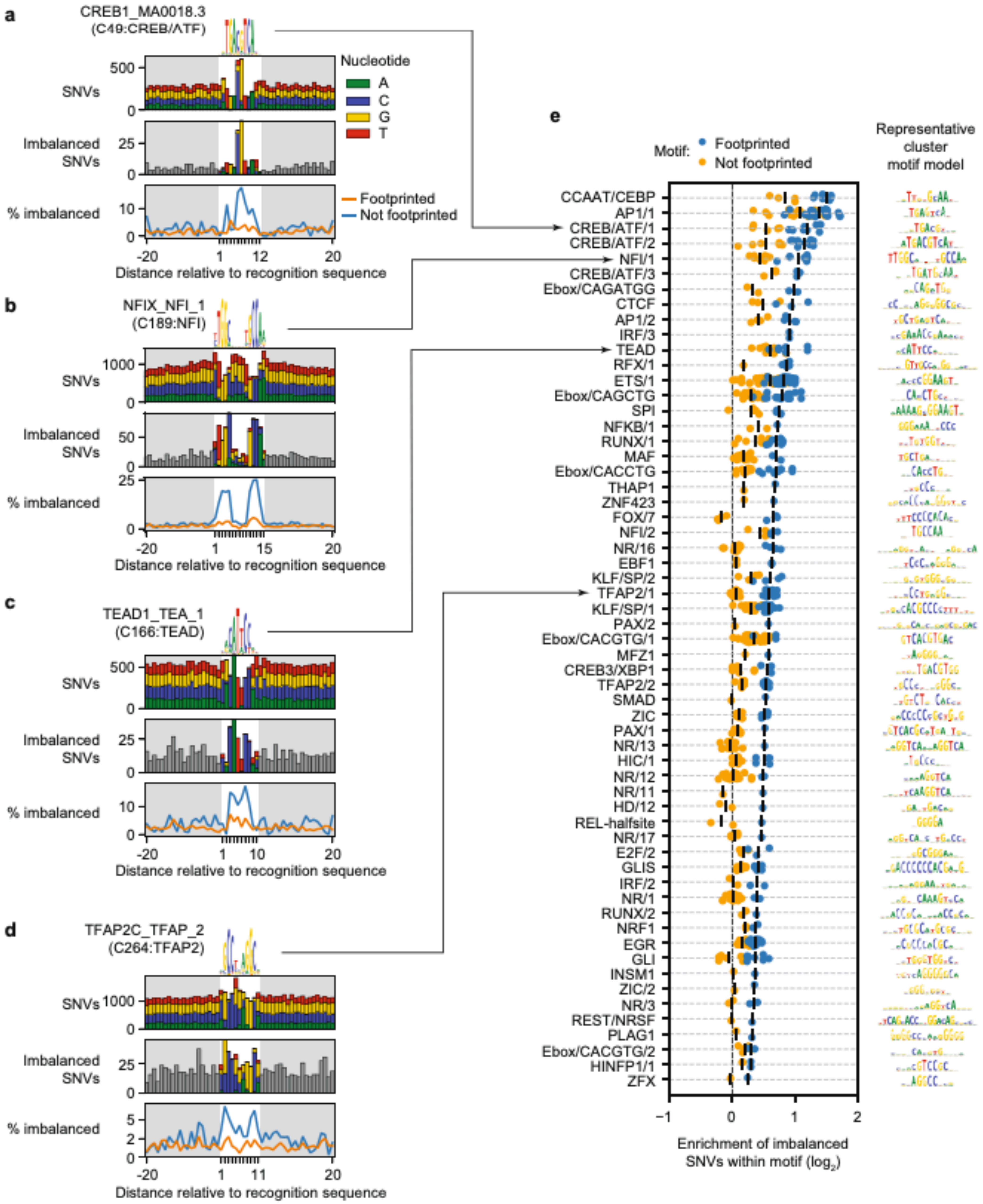
chrom	start	end	variant_id	dbsnp	ref	alt	aa	maf	ard	mu	sigma	n_ref	n_alt	n_total	n_hets	mean_rd
chr1	1137499	1137500	chr1:1137500:G:A	rs143580335	G	A	G	0.00754	0.7209	0.7329	0.0898	31	12	43	3	14.3333
chr1	1217250	1217251	chr1:1217251:C:A	rs11721	C	A	A	0.13863	0.8296	0.7852	0.2341	112	23	135	21	6.4286
chr1	1251121	1251122	chr1:1251122:A:T	rs6603785	A	T	-	0.22291	0.7895	0.8258	0.2473	30	8	38	11	3.4545
chr1	1630041	1630042	chr1:1630042:C:T	rs141035747	C	T	C	0.01787	0.7083	0.6929	0.0929	34	14	48	2	24
chr1	1780638	1780639	chr1:1780639:C:T	rs56400815	C	T	C	0.0548	0.7222	0.7037	0.1048	26	10	36	3	12
chr1	2050446	2050447	chr1:2050447:C:G	rs192386882	C	G	c	0.07284	0.7333	0.7373	0.1423	88	32	120	10	12
chr1	2195341	2195342	chr1:2195342:T:G	rs374992772	T	G	T	0.00318	0.7857	0.7747	0.045	77	21	98	2	49
chr1	2546185	2546186	chr1:2546186:C:T	rs139454263	C	T	.	0.00823	0.7101	0.7383	0.0474	49	20	69	2	34.5
chr1	2653107	2653108	chr1:2653108:C:T	rs373550866	C	T	-	0.02101	0.75	0.7	0.1312	39	13	52	4	13

Modeling variant effects

- SNVs with imbalance enriched in **motifs and footprints**

- Many SNVs are not imbalanced even when residing within critical positions of TF binding sites (~75% of SNVs at these positions have no measurable allelic skew)

- *Need models that include more context (cell type, chromatin state, sequence, etc.)*



Future plans/in progress

- Similarly processing ChIP-seq data (w/ Ryan Tewhey)
- Modeling/predicting variant effects using imbalance data
- Investigating cell context on ‘penetrance’ of variants
- QTL study using master DHS index
- Integrate data with disease/trait-associated variants
- Possibly look into indels; need to validate genoptying approach
- Building a useable atlas for ENCODE4

Data availability

- http://resources.altius.org/~jvierstra/projects/encode4-allelic-imbalance
  - /genotypes/[release date]/all.filtered.snps.annotated.vcf.gz <- genotypes
  - /allelic_mapping/[release data]/allele_counts.vcf.gz <- allelic counts for heterozygous sites
  - /allelic_mapping/[release data]/metadata.tsv <- contains ENCODE accession and individual ID (genotype ID) and basic dataset information
  - Will be password protected; do not share with non-ENCODE users