DAC Status update

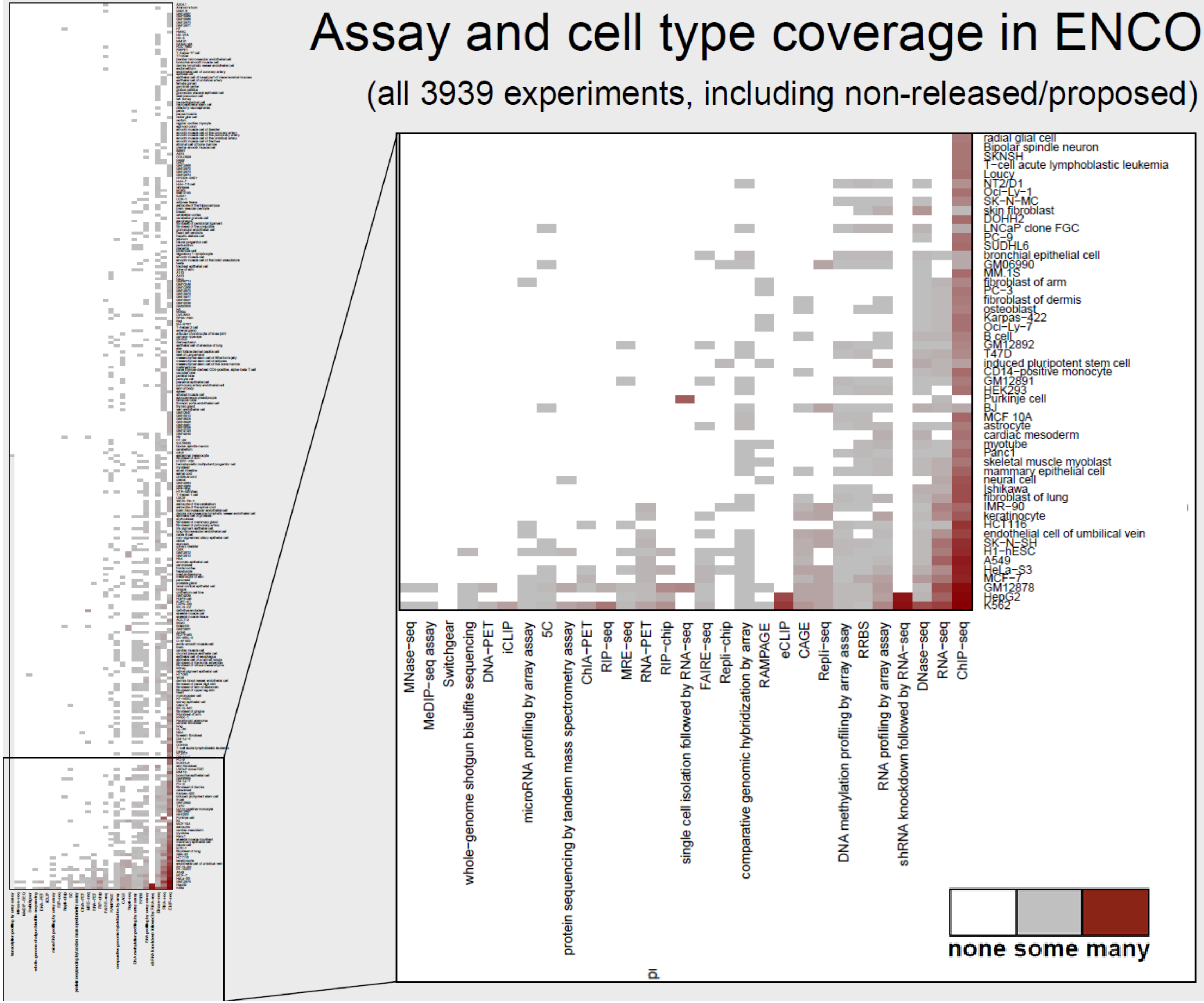
4. **Analysis of ENCODE portfolio by cell type**
   - Action: The DAC will analyze the ENCODE Portfolio by cell type and determine what space ENCODE has and has not covered (5/21)

5. **Tracking ENCODE Element Identification Over Time (NHGRI and DAC)**
   - Action: The DAC will provide the NHGRI team with a plan for tracking ENCODE element identification. (4/16)

6. **Cell identity testing (DAC)**
   - Action: The DAC will develop and apply methods for automatically testing the identity of cell types

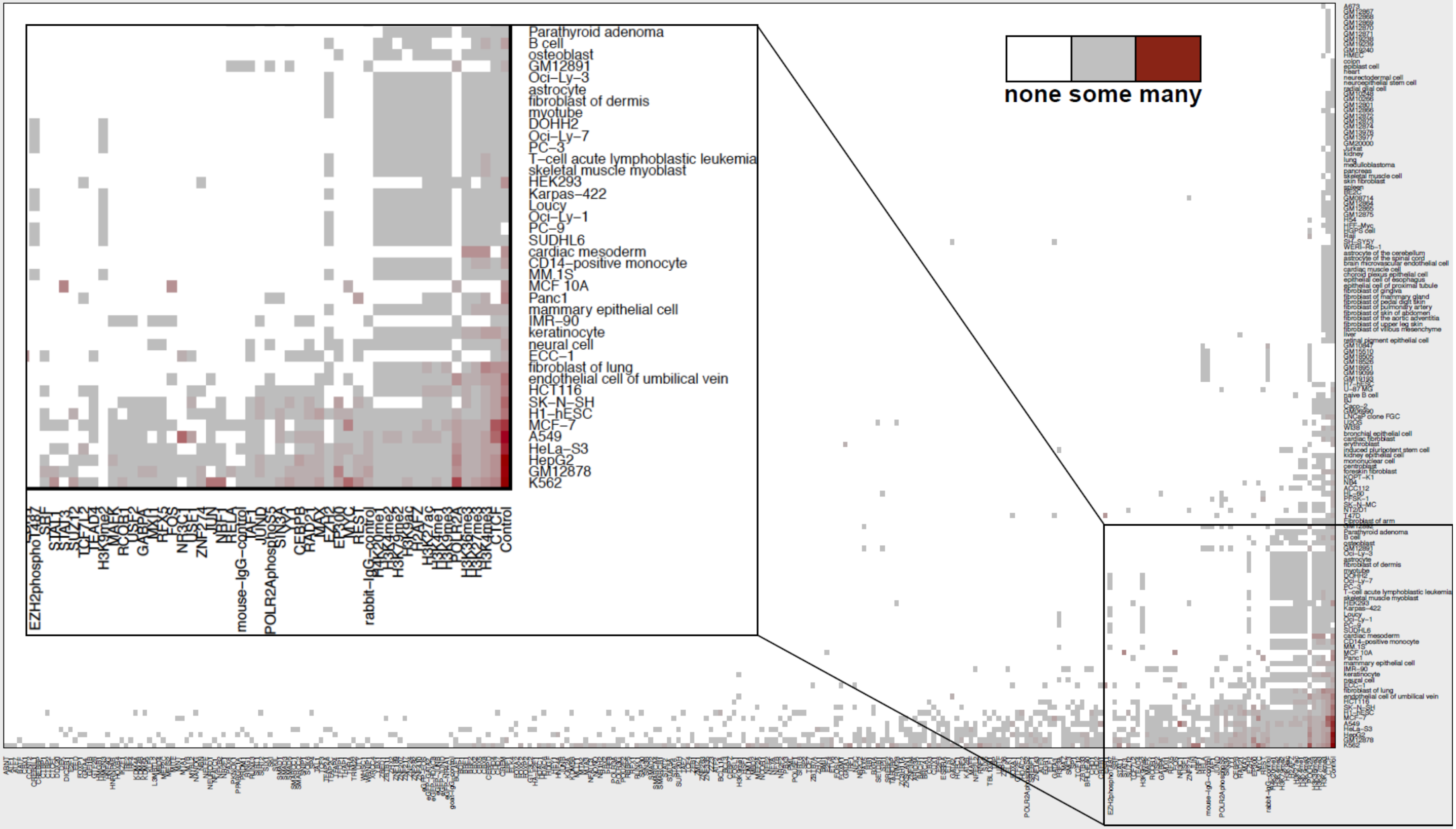
Assay and cell type coverage in ENCODE

(all 3939 experiments, including non-released/proposed)



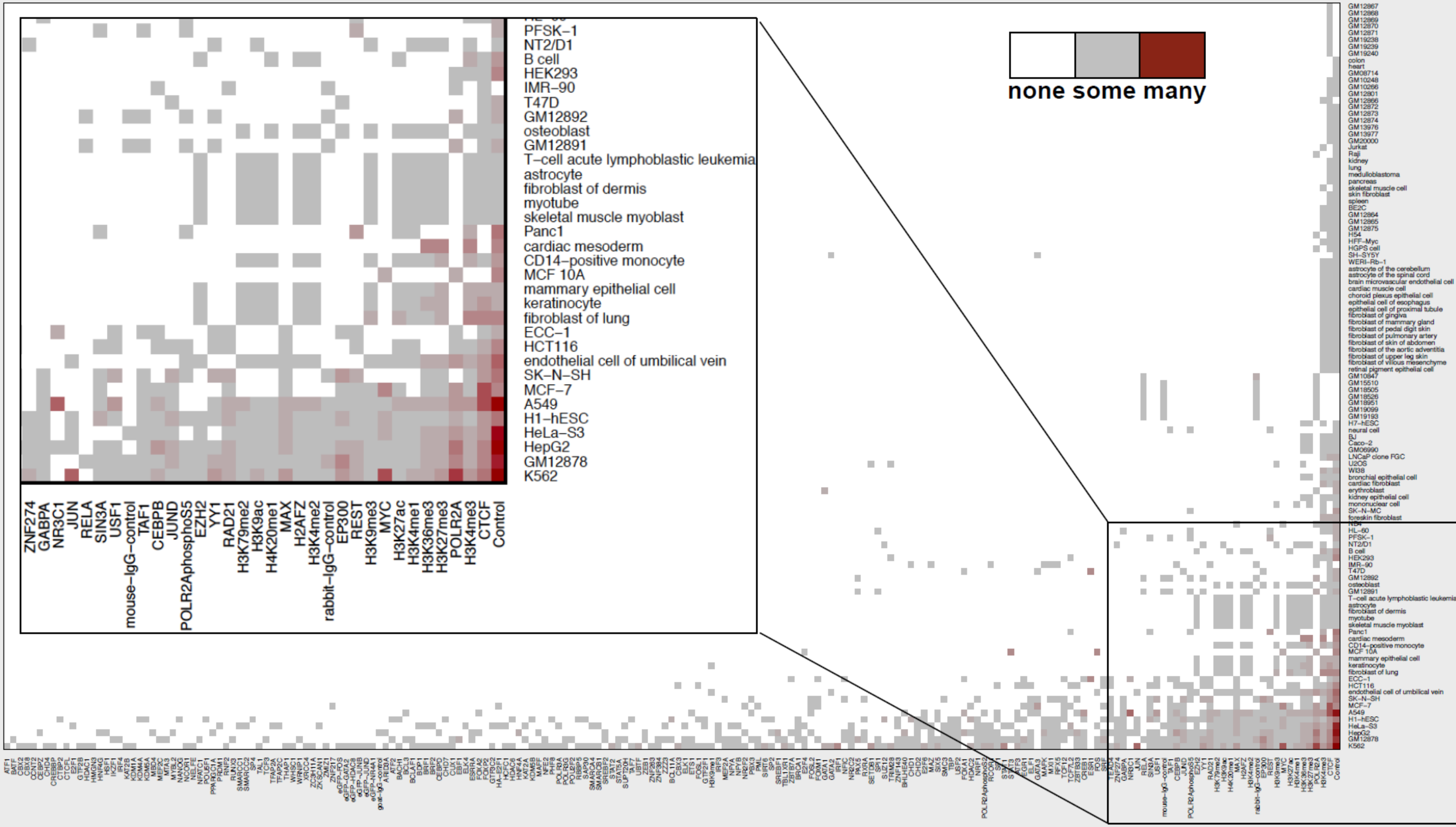
Cell type coverage of ChIP-seq experiments

(out of all 3939 experiments, including non-released/proposed)

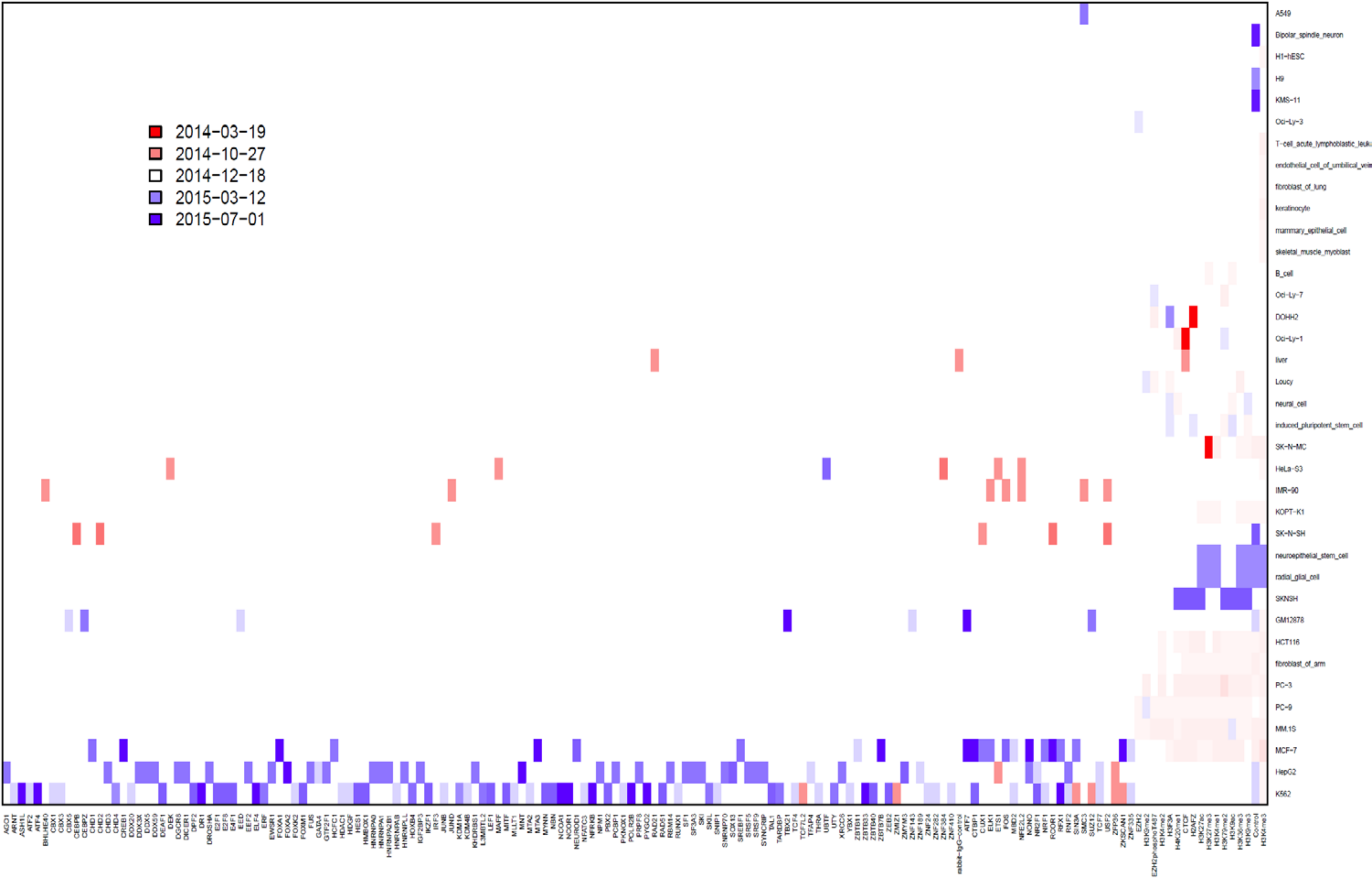


Cell type coverage of ChIP-seq experiments

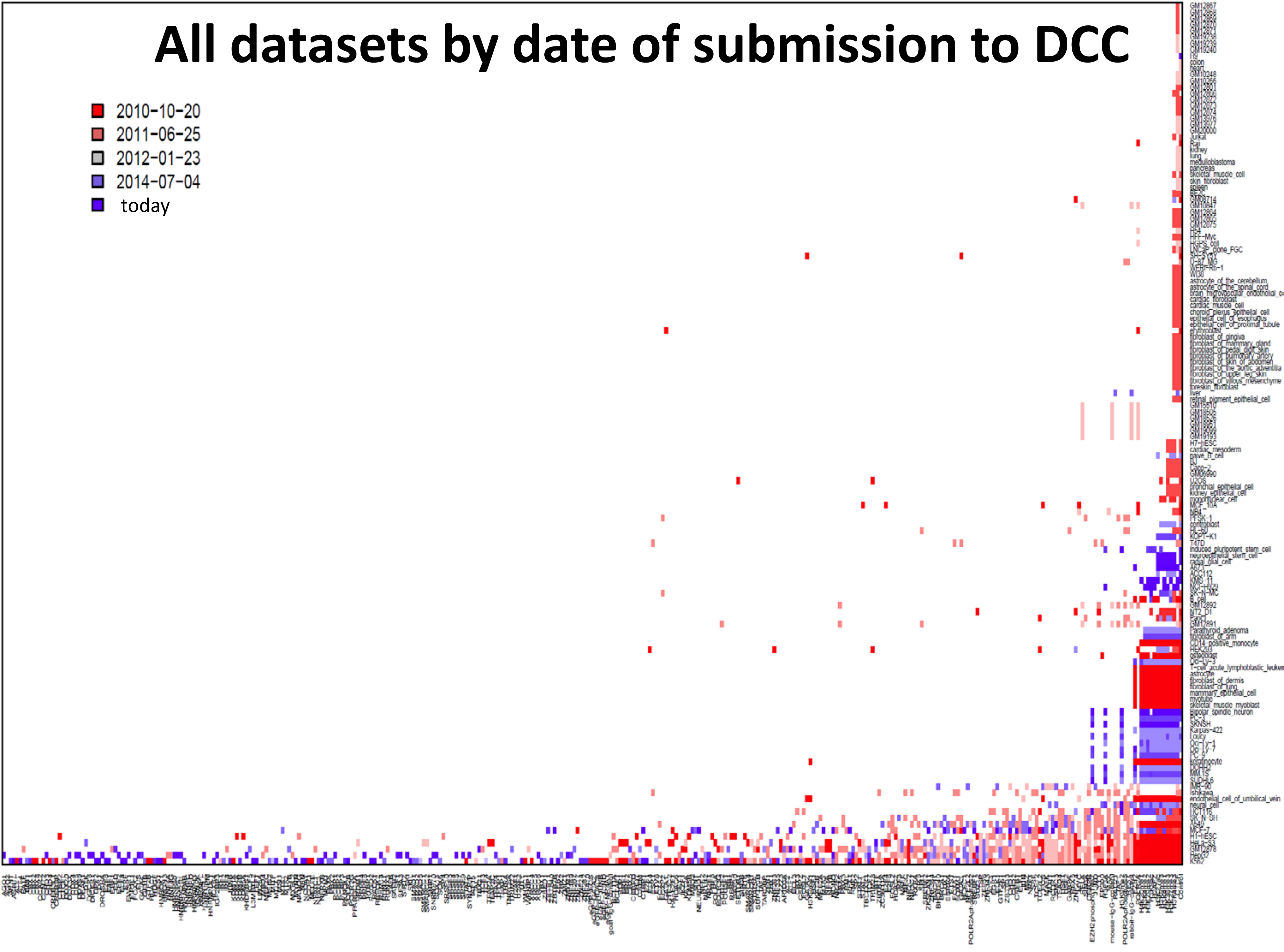
(out of 2618 released experiments only)



Upcoming datasets by date of submission to DCC (submitted, but not yet released)



All datasets by date of submission to DCC



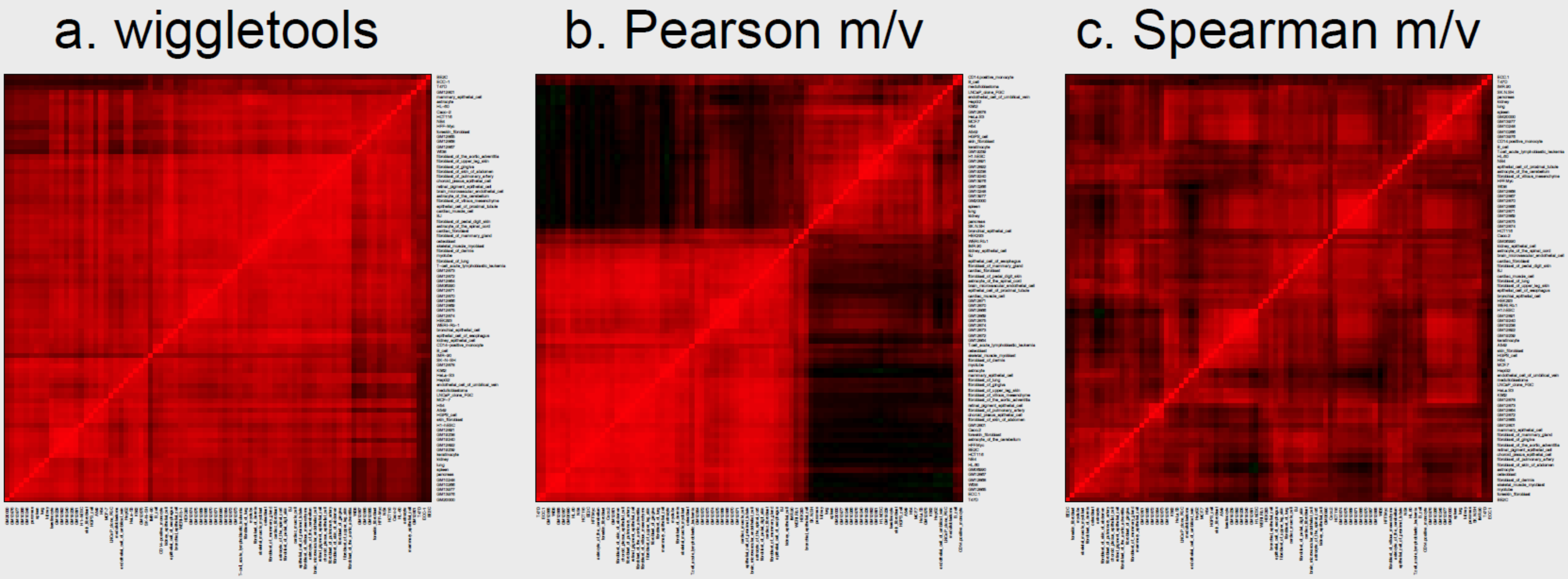
Beyond metadata: correlating datasets

Focusing on 1449 ChIP-seq experiments (released, hg19)

1. Download all ENCODE ChIP-seq experiment data
   - Using the DCC's REST API: http://wiki.encodedcc.org/index.php/The_ENCODE_REST_API
2. For each cell-type/target combination, select the largest file (as a weak proxy for data quality / sequencing depth)
   - This results in 1100 data files for ChIP-seq alone
3. Calculate pairwise correlations between cell types
   - Pearson correlation on full bigWig files (wiggletools)
   - Pearson/Spearman correlation on 10,000 regions that are most variable across cell types

For example: CTCF

🟩 ■ 🟥
-1 0 +1



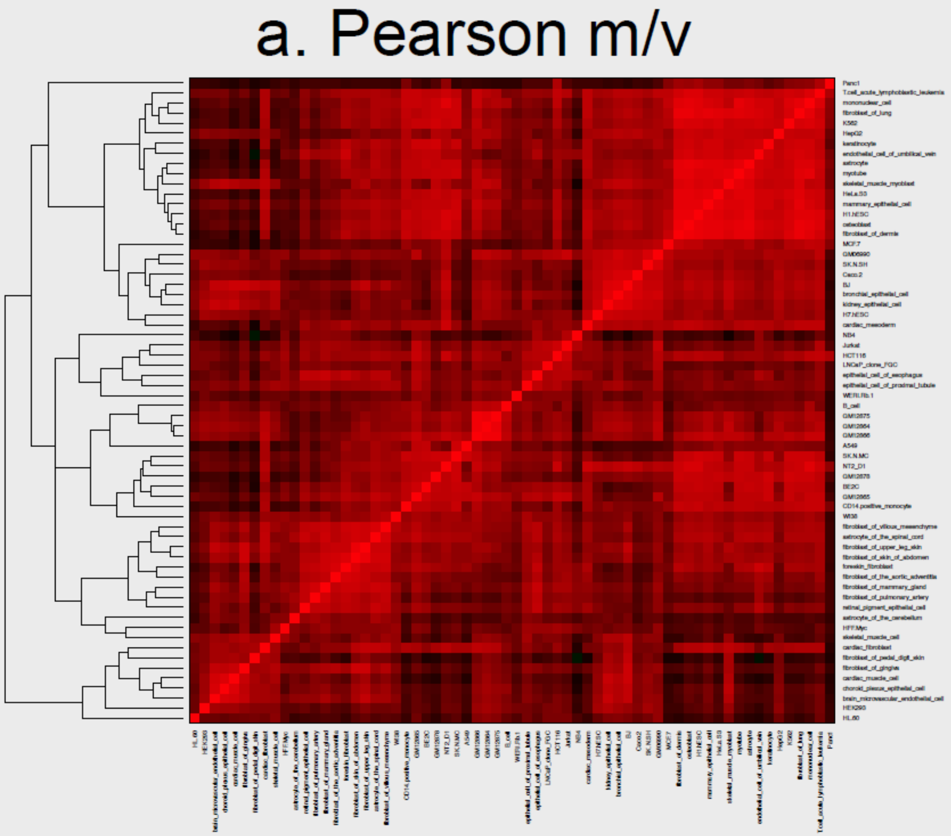
- Shown are pairwise correlation coefficients across 86 cell types with CTCF ChIP-seq data.
- Pearson correlations based on full bigWig files (a) may be too sensitive to noise.

m/v: across 10,000 ‘most variable’ regions

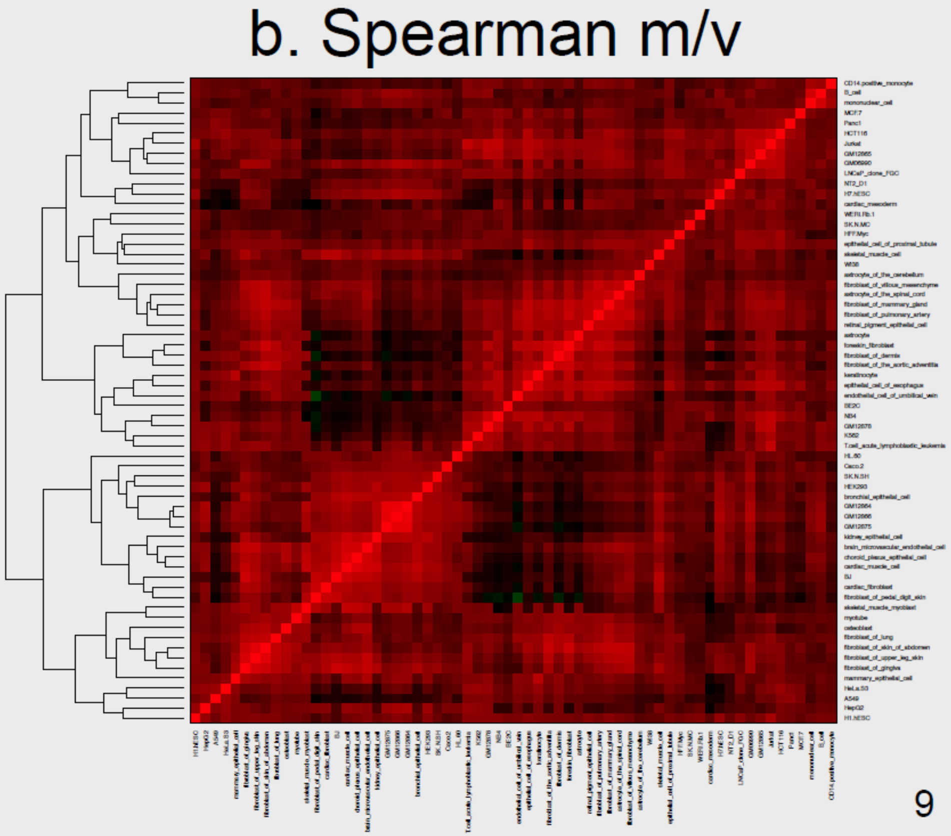
Multiple target proteins

Because we have multiple ChIP-seq datasets per target protein, we can average correlation matrices across targets in an attempt to reduce the effect of individual targets and noise.

Histones: H3K4me3, H3K27me3, H3K36me3, H3K27ac, H3K4me1, H3K9me3, H2AFZ, H3K4me2, H4K20me1, H3K79me2, H3K9ac, H3K9me1

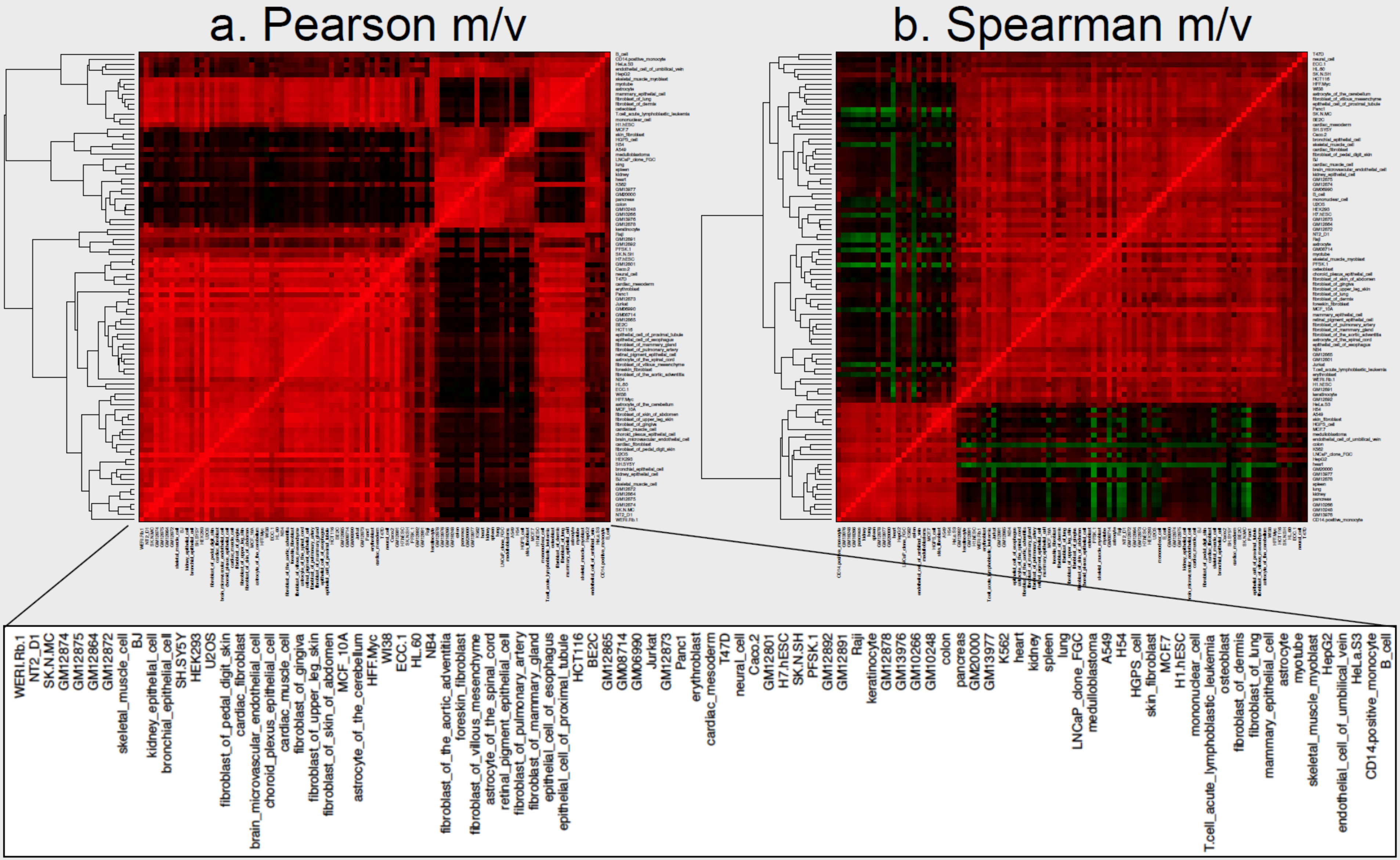


a. Pearson m/v



b. Spearman m/v

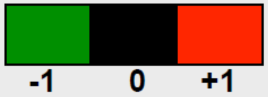
All ChIP-seq data combined (201 target proteins)

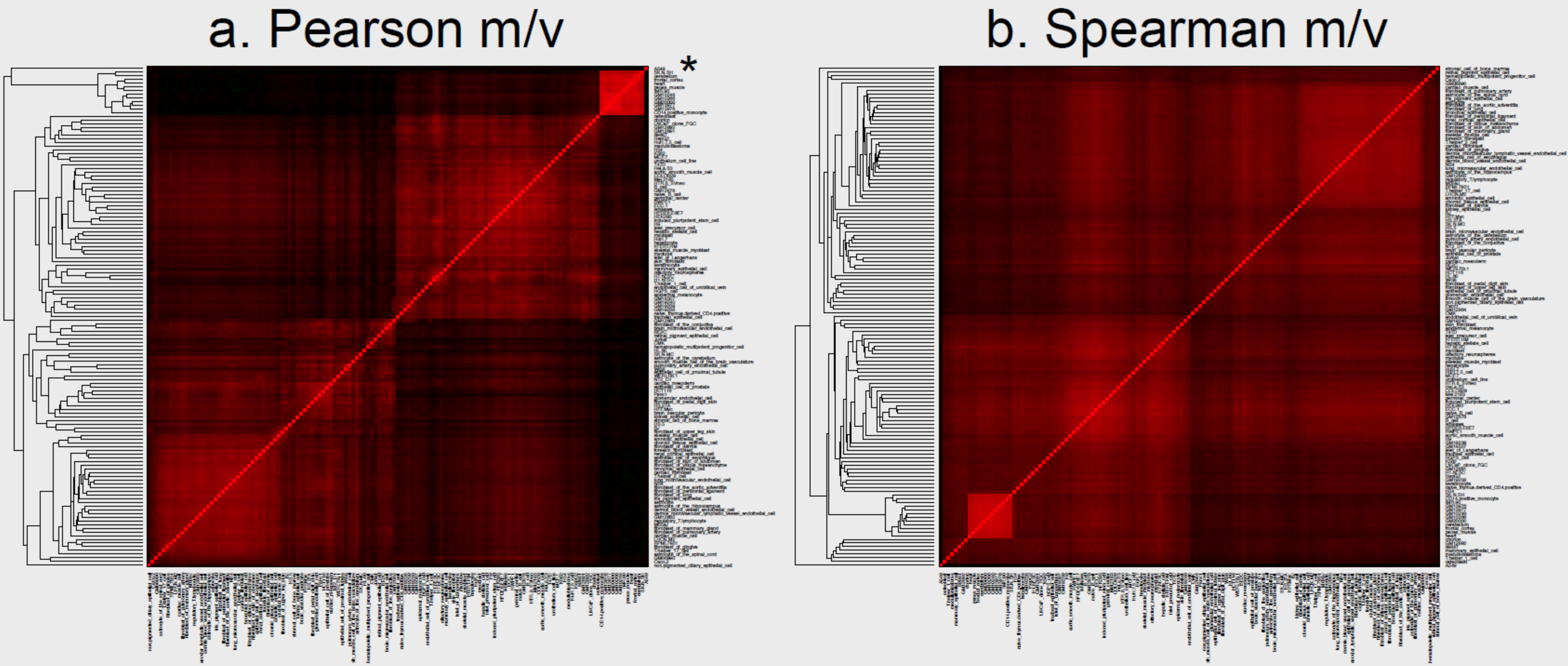

a. Pearson m/v

b. Spearman m/v

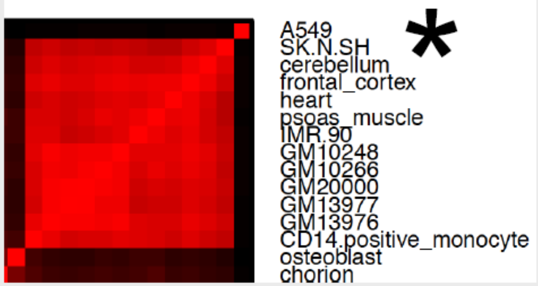
Local grouping makes some sense, global grouping not really

Encore: DNaseI





- Shown are pairwise correlation coefficients across 135 cell types with DNaseI-seq data.



m/v: across 10,000 ‘most variable’ regions

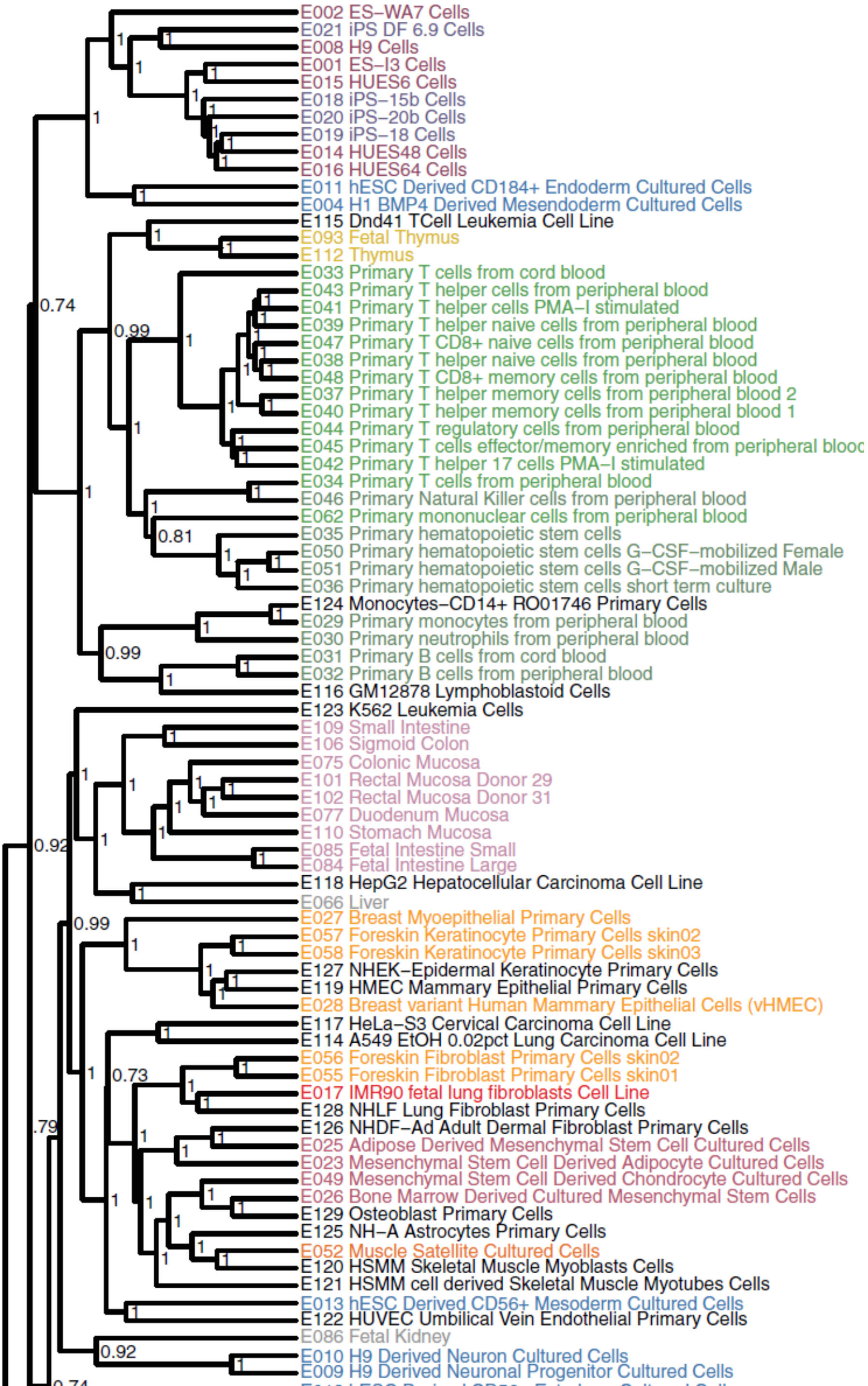
Take-home messages

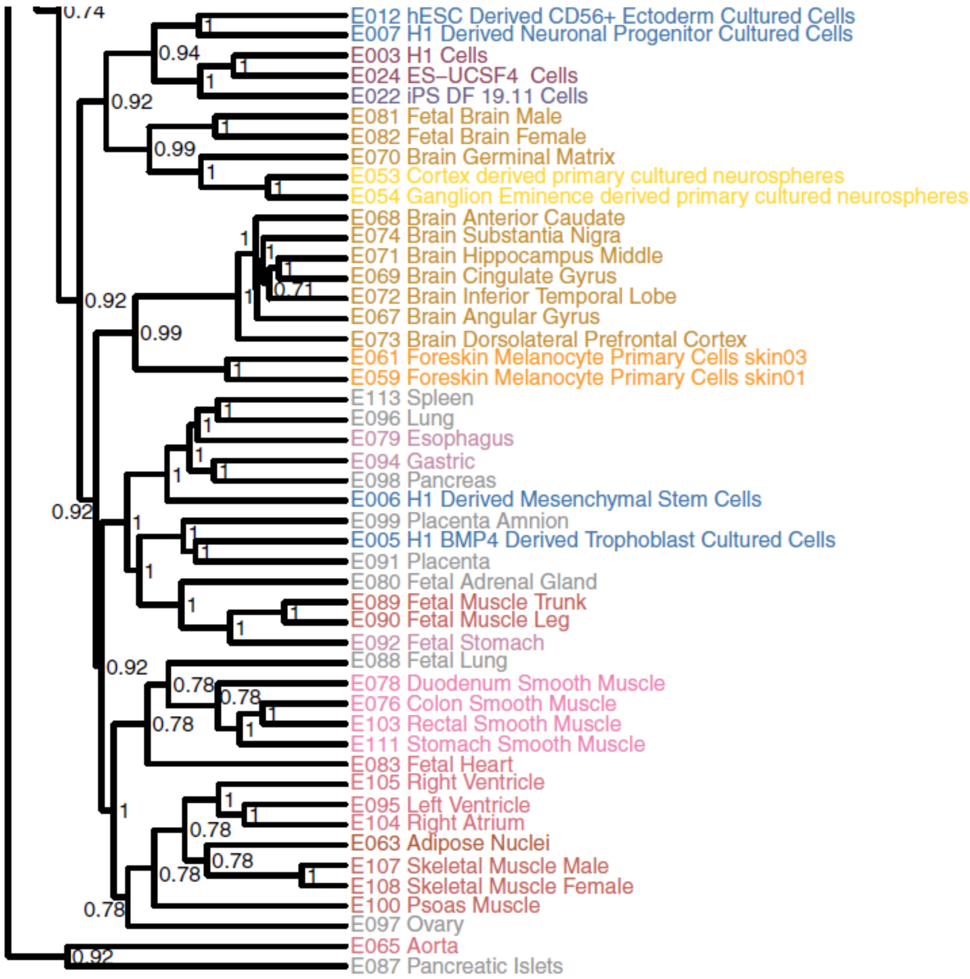
- The meta-data matrix in the first few slides can serve as a reference for deciding which factors to profile in which cell types

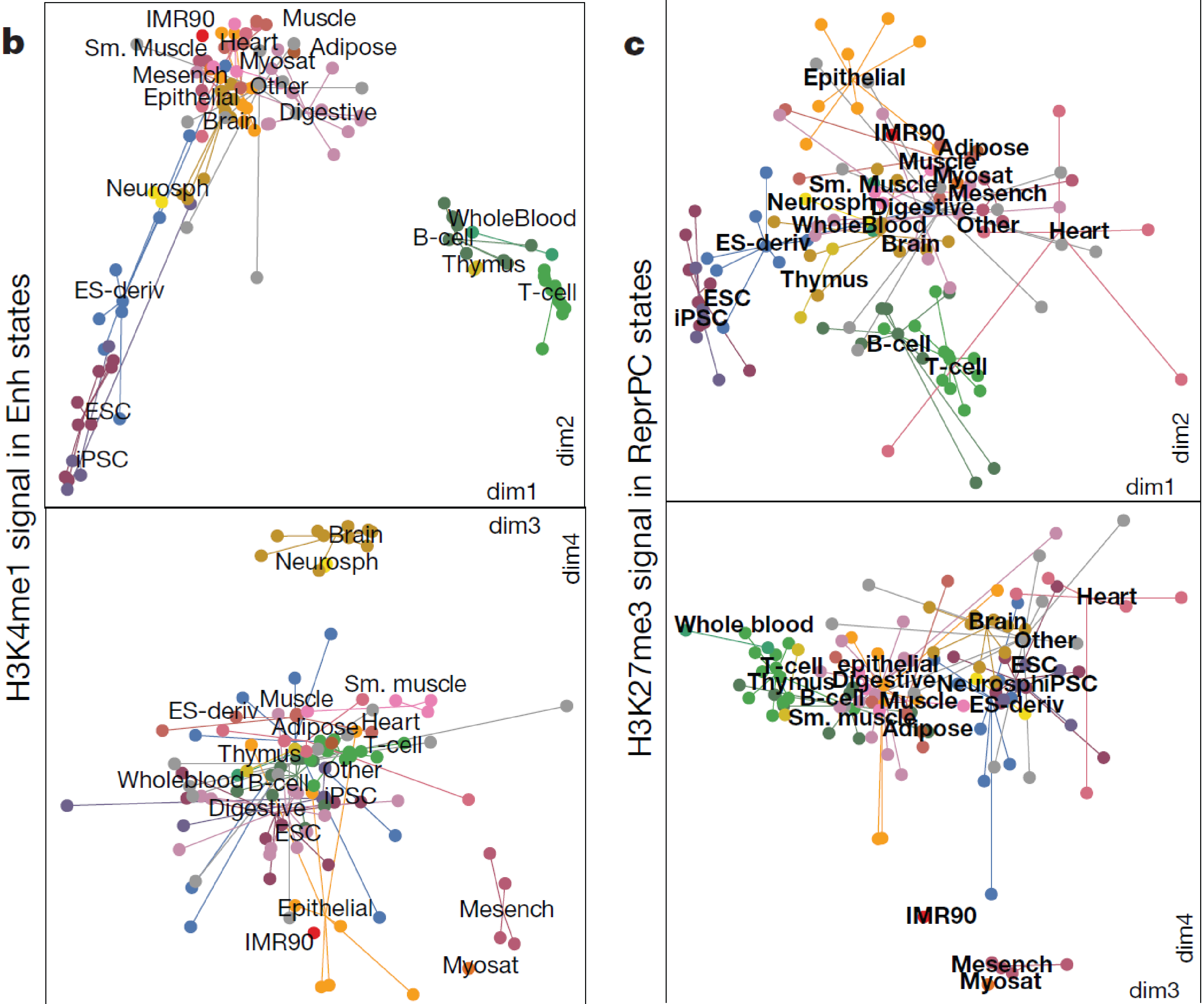
- The correlation matrices may give some idea on which cell types are outliers based on the data available, after which these cell types, or cell types like it, can be profiled in more depth.

- Caveats of current approach:
  - Selecting the largest bigWig files may bias towards certain labs and/or periods in time
  - Data in the ENCODE repository can not be assumed to be uniformly processed. Most often it isn't, with signals being on various scales (e.g., -log10(p-val), fold-change, enrichment, read counts, etc).
  - Selecting 10,000 most variable regions may not be sufficient to reduce the effects of noise

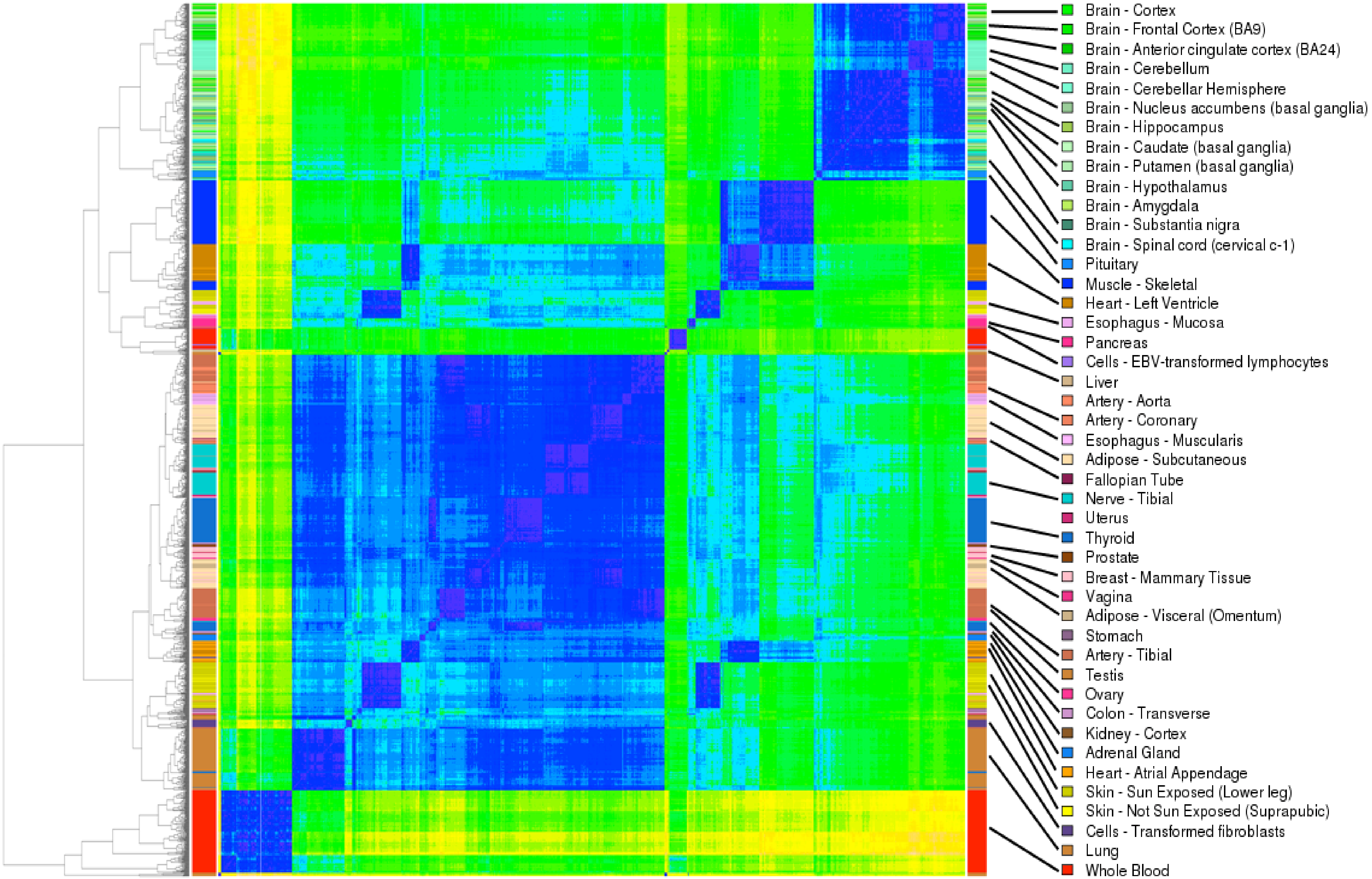
H3K4me1 beyond ENCODE: Cluster with Roadmap


E002 ES–WA7 Cells
E021 iPS DF 6.9 Cells
E008 H9 Cells
E001 ES–I3 Cells
E015 HUES6 Cells
E018 iPS–15b Cells
E020 iPS–20b Cells
E019 iPS–18 Cells
E014 HUES48 Cells
E016 HUES64 Cells
E011 hESC Derived CD184+ Endoderm Cultured Cells
E004 H1 BMP4 Derived Mesendoderm Cultured Cells
E115 Dnd41 TCell Leukemia Cell Line
E093 Fetal Thymus
E112 Thymus
E033 Primary T cells from cord blood
E043 Primary T helper cells from peripheral blood
E041 Primary T helper cells PMA–I stimulated
E039 Primary T helper naive cells from peripheral blood
E047 Primary T CD8+ naive cells from peripheral blood
E038 Primary T helper naive cells from peripheral blood
E048 Primary T CD8+ memory cells from peripheral blood
E037 Primary T helper memory cells from peripheral blood 2
E040 Primary T helper memory cells from peripheral blood 1
E044 Primary T regulatory cells from peripheral blood
E045 Primary T cells effector/memory enriched from peripheral blooc
E042 Primary T helper 17 cells PMA–I stimulated
E034 Primary T cells from peripheral blood
E046 Primary Natural Killer cells from peripheral blood
E062 Primary mononuclear cells from peripheral blood
E035 Primary hematopoietic stem cells
E050 Primary hematopoietic stem cells G–CSF–mobilized Female
E051 Primary hematopoietic stem cells G–CSF–mobilized Male
E036 Primary hematopoietic stem cells short term culture
E124 Monocytes–CD14+ RO01746 Primary Cells
E029 Primary monocytes from peripheral blood
E030 Primary neutrophils from peripheral blood
E031 Primary B cells from cord blood
E032 Primary B cells from peripheral blood
E116 GM12878 Lymphoblastoid Cells
E123 K562 Leukemia Cells
E109 Small Intestine
E106 Sigmoid Colon
E075 Colonic Mucosa
E101 Rectal Mucosa Donor 29
E102 Rectal Mucosa Donor 31
E077 Duodenum Mucosa
E110 Stomach Mucosa
E085 Fetal Intestine Small
E084 Fetal Intestine Large
E118 HepG2 Hepatocellular Carcinoma Cell Line
E066 Liver
E027 Breast Myoepithelial Primary Cells
E057 Foreskin Keratinocyte Primary Cells skin02
E058 Foreskin Keratinocyte Primary Cells skin03
E127 NHEK–Epidermal Keratinocyte Primary Cells
E119 HMEC Mammary Epithelial Primary Cells
E028 Breast variant Human Mammary Epithelial Cells (vHMEC)
E117 HeLa–S3 Cervical Carcinoma Cell Line
E114 A549 EtOH 0.02pct Lung Carcinoma Cell Line
E056 Foreskin Fibroblast Primary Cells skin02
E055 Foreskin Fibroblast Primary Cells skin01
E017 IMR90 fetal lung fibroblasts Cell Line
E128 NHLF Lung Fibroblast Primary Cells
E126 NHDF–Ad Adult Dermal Fibroblast Primary Cells
E025 Adipose Derived Mesenchymal Stem Cell Cultured Cells
E023 Mesenchymal Stem Cell Derived Adipocyte Cultured Cells
E049 Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells
E026 Bone Marrow Derived Cultured Mesenchymal Stem Cells
E129 Osteoblast Primary Cells
E125 NH–A Astrocytes Primary Cells
E052 Muscle Satellite Cultured Cells
E120 HSMM Skeletal Muscle Myoblasts Cells
E121 HSMM cell derived Skeletal Muscle Myotubes Cells
E013 hESC Derived CD56+ Mesoderm Cultured Cells
E122 HUVEC Umbilical Vein Endothelial Primary Cells
E086 Fetal Kidney
E010 H9 Derived Neuron Cultured Cells
E009 H9 Derived Neuronal Progenitor Cultured Cells
0.74
0.99
0.81
0.99
0.92
0.99
0.73
.79
0.92
0.74

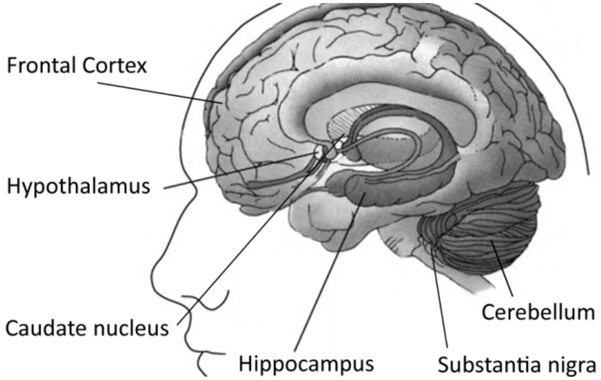

0.74
E012 hESC Derived CD56+ Ectoderm Cultured Cells
E007 H1 Derived Neuronal Progenitor Cultured Cells
E003 H1 Cells
E024 ES–UCSF4 Cells
E022 iPS DF 19.11 Cells
E081 Fetal Brain Male
E082 Fetal Brain Female
E070 Brain Germinal Matrix
E053 Cortex derived primary cultured neurospheres
E054 Ganglion Eminence derived primary cultured neurospheres
E068 Brain Anterior Caudate
E074 Brain Substantia Nigra
E071 Brain Hippocampus Middle
E069 Brain Cingulate Gyrus
E072 Brain Inferior Temporal Lobe
E067 Brain Angular Gyrus
E073 Brain Dorsolateral Prefrontal Cortex
E061 Foreskin Melanocyte Primary Cells skin03
E059 Foreskin Melanocyte Primary Cells skin01
E113 Spleen
E096 Lung
E079 Esophagus
E094 Gastric
E098 Pancreas
E006 H1 Derived Mesenchymal Stem Cells
E099 Placenta Amnion
E005 H1 BMP4 Derived Trophoblast Cultured Cells
E091 Placenta
E080 Fetal Adrenal Gland
E089 Fetal Muscle Trunk
E090 Fetal Muscle Leg
E092 Fetal Stomach
E088 Fetal Lung
E078 Duodenum Smooth Muscle
E076 Colon Smooth Muscle
E103 Rectal Smooth Muscle
E111 Stomach Smooth Muscle
E083 Fetal Heart
E105 Right Ventricle
E095 Left Ventricle
E104 Right Atrium
E063 Adipose Nuclei
E107 Skeletal Muscle Male
E108 Skeletal Muscle Female
E100 Psoas Muscle
E097 Ovary
E065 Aorta
E087 Pancreatic Islets
0.94
0.92
0.99
0.71
0.92
0.99
0.92
0.92
0.78
0.78
0.78
0.78
0.78
0.78
0.78
0.92

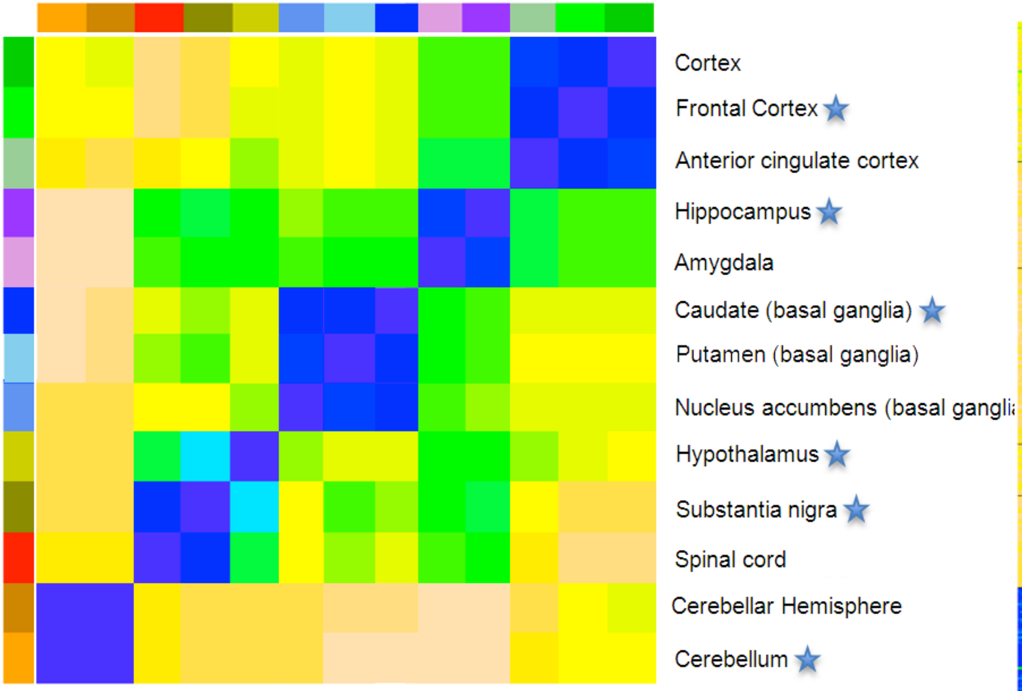

b
H3K4me1 signal in Enh states
IMR90
Muscle
Sm. Muscle
Heart
Adipose
Myosat
Mesench
Other
Epithelial
Digestive
Brain
Neurosph
WholeBlood
B-cell
Thymus
T-cell
ES-deriv
ESC
iPSC
dim1
dim2
dim3
dim4
Brain
Neurosph
Muscle
Sm. muscle
ES-deriv
Adipose
Heart
Thymus
T-cell
Wholeblood
B-cell
Other
iPSC
Digestive
ESC
Epithelial
IMR90
Mesench
Myosat
c
H3K27me3 signal in ReprPC states
Epithelial
IMR90
Adipose
Muscle
Myosat
Sm. Muscle
Mesench
Neurosph
Digestive
WholeBlood
Other
Heart
ES-deriv
Brain
Thymus
ESC
iPSC
B-cell
T-cell
dim1
dim2
Whole blood
Heart
Brain
Other
T-cell
epithelial
ESC
Thymus
Digestive
Neurosph
iPSC
B-cell
Muscle
ES-deriv
Sm. muscle
Adipose
IMR90
Mesench
Myosat
dim3
dim4

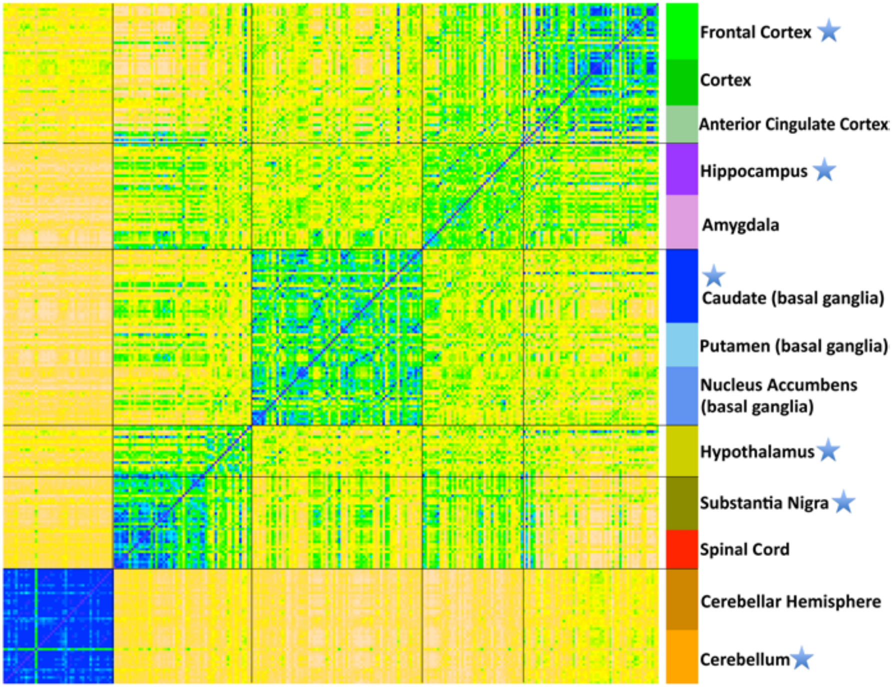
RNA-Seq expression clustering: GTEx



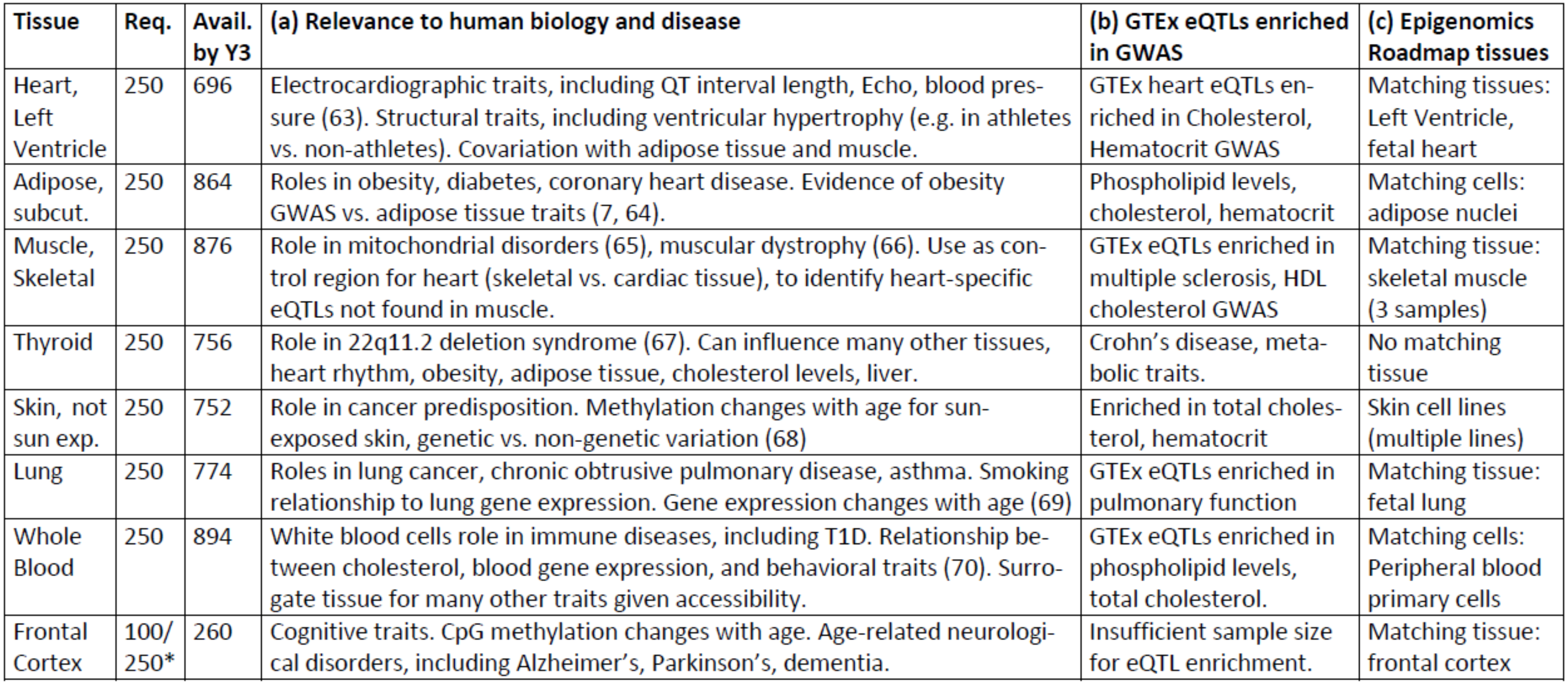
GTEx: Focus on brain sub-regions



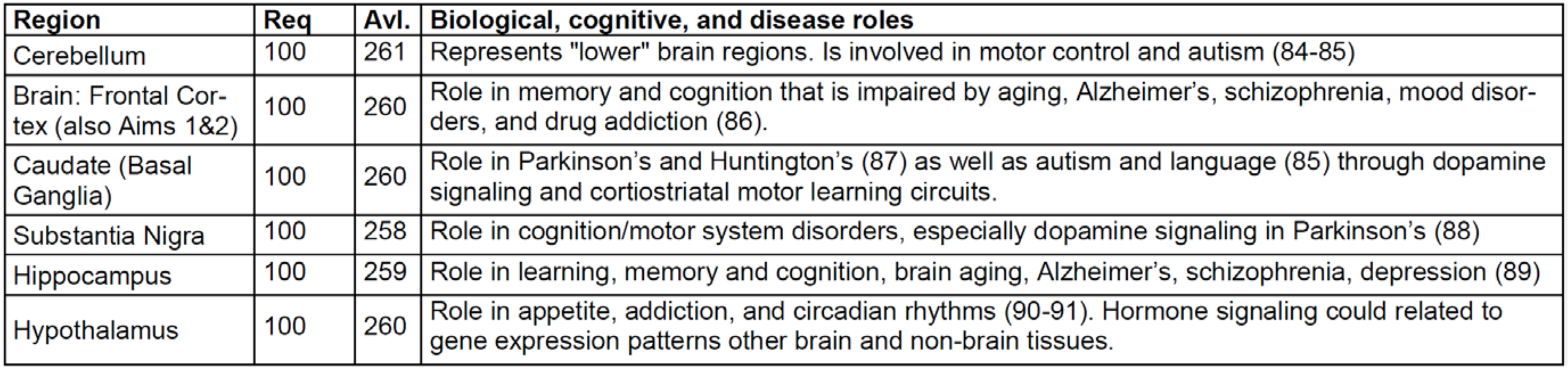




Prioritization for human disease relevance

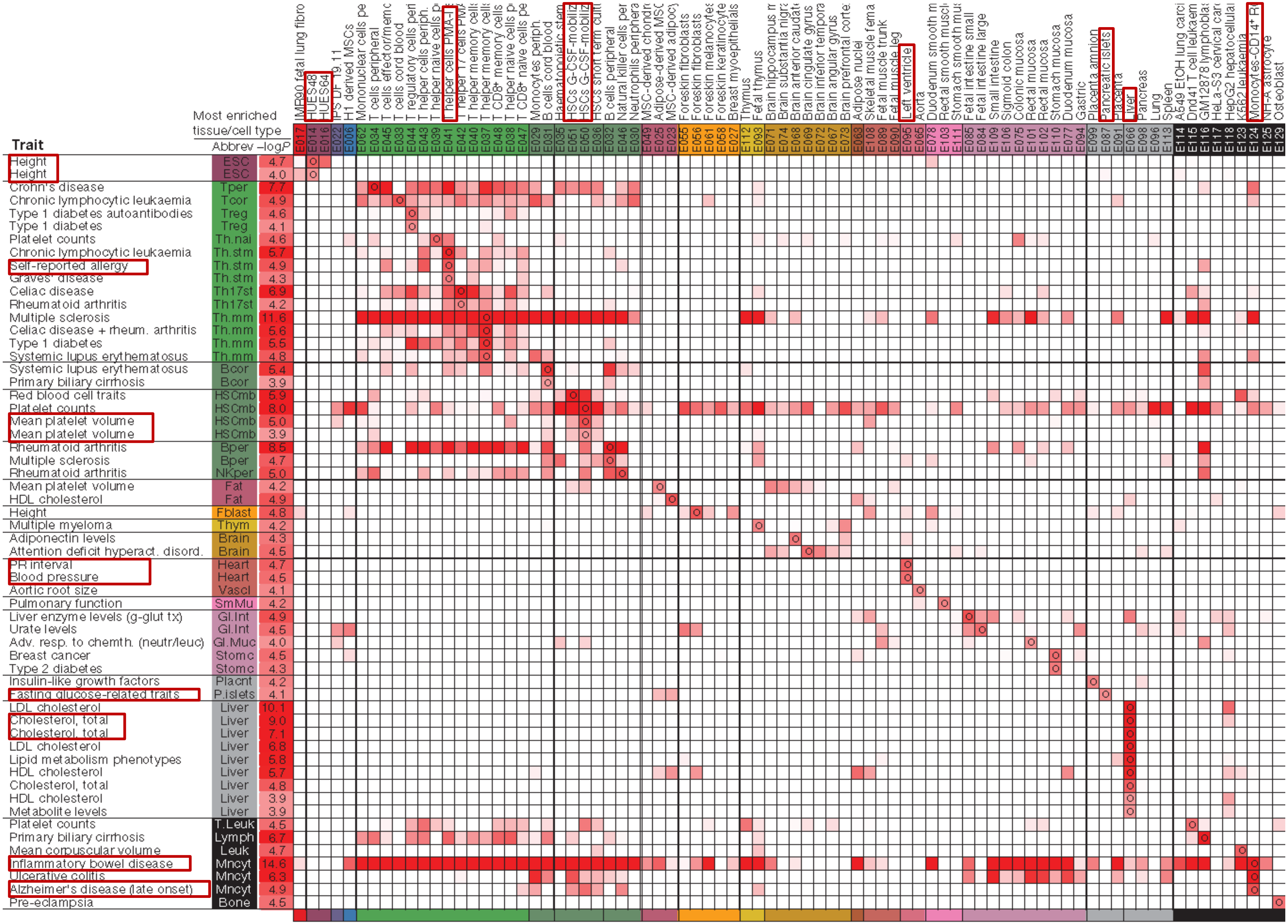


Tissue	Req.	Avail. by Y3	(a) Relevance to human biology and disease	(b) GTEx eQTLs enriched in GWAS	(c) Epigenomics Roadmap tissues
Heart, Left Ventricle	250	696	Electrocardiographic traits, including QT interval length, Echo, blood pressure (63). Structural traits, including ventricular hypertrophy (e.g. in athletes vs. non-athletes). Covariation with adipose tissue and muscle.	GTEx heart eQTLs enriched in Cholesterol, Hematocrit GWAS	Matching tissues: Left Ventricle, fetal heart
Adipose, subcut.	250	864	Roles in obesity, diabetes, coronary heart disease. Evidence of obesity GWAS vs. adipose tissue traits (7, 64).	Phospholipid levels, cholesterol, hematocrit	Matching cells: adipose nuclei
Muscle, Skeletal	250	876	Role in mitochondrial disorders (65), muscular dystrophy (66). Use as control region for heart (skeletal vs. cardiac tissue), to identify heart-specific eQTLs not found in muscle.	GTEx eQTLs enriched in multiple sclerosis, HDL cholesterol GWAS	Matching tissue: skeletal muscle (3 samples)
Thyroid	250	756	Role in 22q11.2 deletion syndrome (67). Can influence many other tissues, heart rhythm, obesity, adipose tissue, cholesterol levels, liver.	Crohn's disease, metabolic traits.	No matching tissue
Skin, not sun exp.	250	752	Role in cancer predisposition. Methylation changes with age for sun-exposed skin, genetic vs. non-genetic variation (68)	Enriched in total cholesterol, hematocrit	Skin cell lines (multiple lines)
Lung	250	774	Roles in lung cancer, chronic obtrusive pulmonary disease, asthma. Smoking relationship to lung gene expression. Gene expression changes with age (69)	GTEx eQTLs enriched in pulmonary function	Matching tissue: fetal lung
Whole Blood	250	894	White blood cells role in immune diseases, including T1D. Relationship between cholesterol, blood gene expression, and behavioral traits (70). Surrogate tissue for many other traits given accessibility.	GTEx eQTLs enriched in phospholipid levels, total cholesterol.	Matching cells: Peripheral blood primary cells
Frontal Cortex	100/ 250*	260	Cognitive traits. CpG methylation changes with age. Age-related neurological disorders, including Alzheimer's, Parkinson's, dementia.	Insufficient sample size for eQTL enrichment.	Matching tissue: frontal cortex



Region	Req	Avl.	Biological, cognitive, and disease roles
Cerebellum	100	261	Represents "lower" brain regions. Is involved in motor control and autism (84-85)
Brain: Frontal Cortex (also Aims 1&2)	100	260	Role in memory and cognition that is impaired by aging, Alzheimer's, schizophrenia, mood disorders, and drug addiction (86).
Caudate (Basal Ganglia)	100	260	Role in Parkinson's and Huntington's (87) as well as autism and language (85) through dopamine signaling and cortiostriatal motor learning circuits.
Substantia Nigra	100	258	Role in cognition/motor system disorders, especially dopamine signaling in Parkinson's (88)
Hippocampus	100	259	Role in learning, memory and cognition, brain aging, Alzheimer's, schizophrenia, depression (89)
Hypothalamus	100	260	Role in appetite, addiction, and circadian rhythms (90-91). Hormone signaling could related to gene expression patterns other brain and non-brain tissues.

Prioritization based on observed GWAS enrichments



	Most enriched tissue/cell type	
Trait	Abbrev	−log*P*
Height	ESC	4.7
Height	ESC	4.0
Crohn's disease	Tper	7.7
Chronic lymphocytic leukaemia	Tcor	4.9
Type 1 diabetes autoantibodies	Treg	4.6
Type 1 diabetes	Treg	4.1
Platelet counts	Th.nai	4.6
Chronic lymphocytic leukaemia	Th.stm	5.7
Self-reported allergy	Th.stm	4.9
Graves' disease	Th.stm	4.3
Celiac disease	Th17st	6.9
Rheumatoid arthritis	Th17st	4.2
Multiple sclerosis	Th.mm	11.6
Celiac disease + rheum. arthritis	Th.mm	5.6
Type 1 diabetes	Th.mm	5.5
Systemic lupus erythematosus	Th.mm	4.8
Systemic lupus erythematosus	Bcor	5.4
Primary biliary cirrhosis	Bcor	3.9
Red blood cell traits	HSCmb	5.9
Platelet counts	HSCmb	8.0
Mean platelet volume	HSCmb	5.0
Mean platelet volume	HSCmb	3.9
Rheumatoid arthritis	Bper	8.5
Multiple sclerosis	Bper	4.7
Rheumatoid arthritis	NKper	5.0
Mean platelet volume	Fat	4.2
HDL cholesterol	Fat	4.9
Height	Fblast	4.8
Multiple myeloma	Thym	4.2
Adiponectin levels	Brain	4.3
Attention deficit hyperact. disord.	Brain	4.5
PR interval	Heart	4.7
Blood pressure	Heart	4.5
Aortic root size	Vascl	4.1
Pulmonary function	SmMu	4.2
Liver enzyme levels (g-glut tx)	GI.Int	4.9
Urate levels	GI.Int	4.5
Adv. resp. to chemth. (neutr/leuc)	GI.Muc	4.0
Breast cancer	Stomc	4.5
Type 2 diabetes	Stomc	4.3
Insulin-like growth factors	Placnt	4.2
Fasting glucose-related traits	P.islets	4.1
LDL cholesterol	Liver	10.1
Cholesterol, total	Liver	9.0
Cholesterol, total	Liver	7.1
LDL cholesterol	Liver	6.8
Lipid metabolism phenotypes	Liver	5.8
HDL cholesterol	Liver	5.7
Cholesterol, total	Liver	4.8
HDL cholesterol	Liver	3.9
Metabolite levels	Liver	3.9
Platelet counts	T.Leuk	4.5
Primary biliary cirrhosis	Lymph	6.7
Mean corpuscular volume	Leuk	4.7
Inflammatory bowel disease	Mncyt	14.6
Ulcerative colitis	Mncyt	6.3
Alzheimer's disease (late onset)	Mncyt	4.9
Pre-eclampsia	Bone	4.5

DAC Status update

4. **Analysis of ENCODE portfolio by cell type**
   - Action: The DAC will analyze the ENCODE Portfolio by cell type and determine what space ENCODE has and has not covered (5/21)

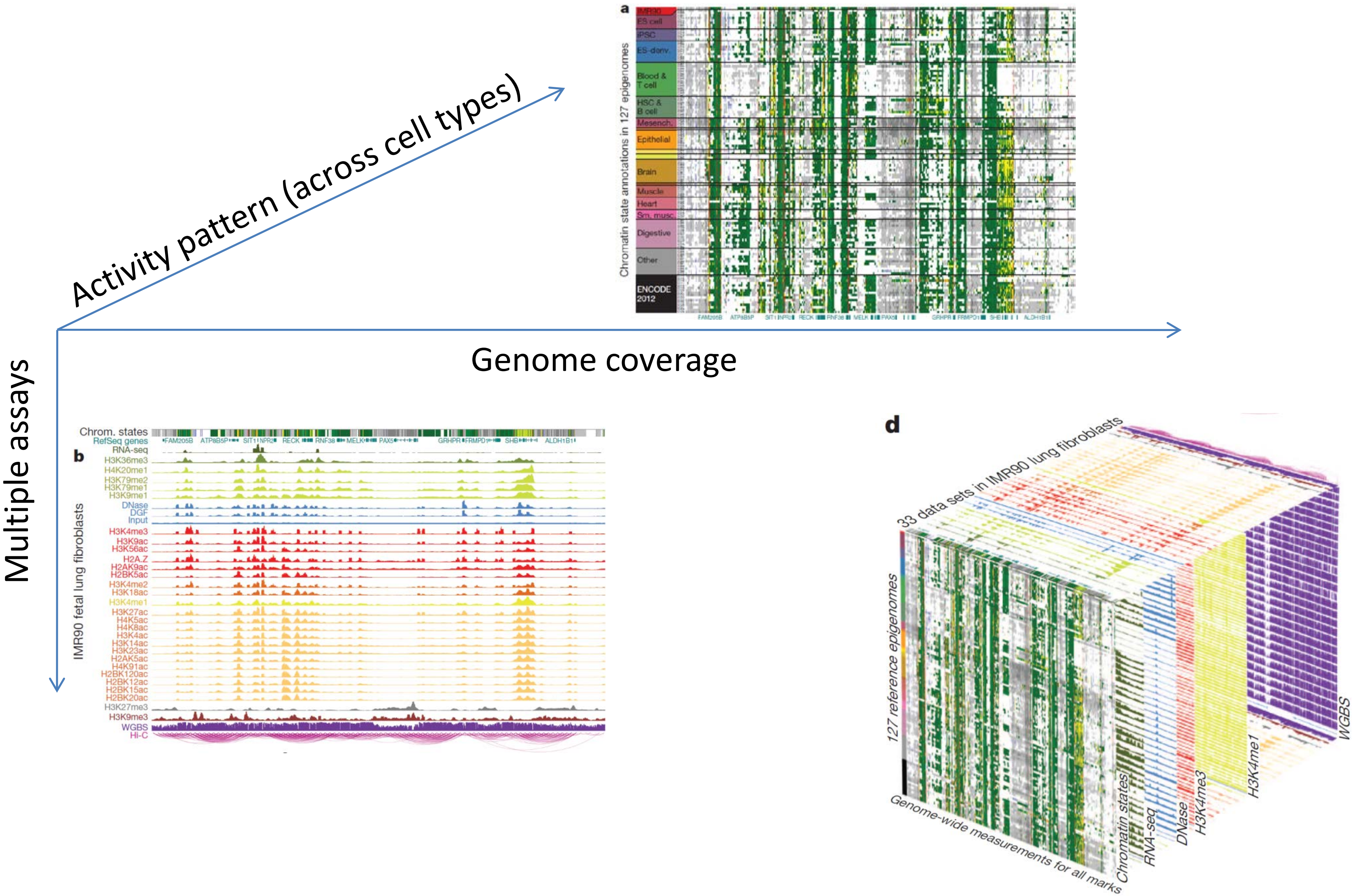
5. **Tracking ENCODE Element Identification Over Time (NHGRI and DAC)**
   - Action: The DAC will provide the NHGRI team with a plan for tracking ENCODE element identification. (4/16)

6. **Cell identity testing (DAC)**
   - Action: The DAC will develop and apply methods for automatically testing the identity of cell types

Tracking ENCODE element identification over time

- **An information-based approach for evaluating the usefulness of existing and planned experiments in ENCODE.**
- Our goal is to develop **formal methods** for assessing the **information gained** from **additional experiments** in the **context of the compendium** of existing ENCODE experiments

Assessing ENCODE progress



Activity pattern (across cell types)

Genome coverage

Multiple assays

Evaluating information content of experiments

Quantify the unique information each experiment provide in the context of the compendium using **information-theoretic approaches.**

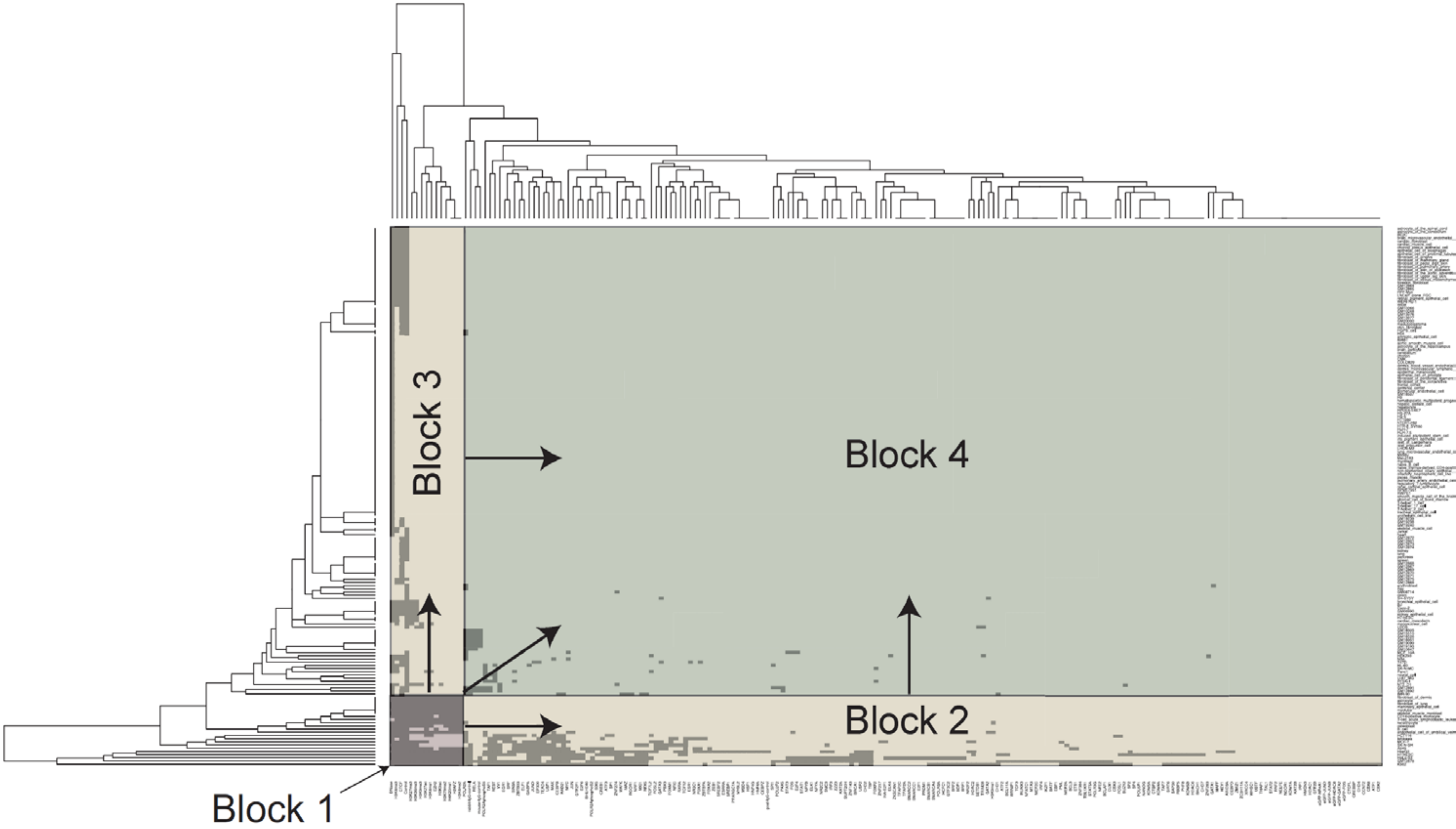
Several factors need to be taken into account :

1. the **reproducibility** of an assay between replicate experiments
2. the **resolution** of the assay
3. the **robustness** of the experiment to variation in experimental conditions
4. the **rarity** of the element type
5. the ability to **predict** of a given assay from other assays in the **same cell type**
6. the ability to **predict** a given assay from same/other assays in **other cell types**
7. the increase in **enrichments for independent datasets** e.g. GWAS variants, regulatory motif matches, evolutionary conservation. resulting from the incorporation of a given experiment to an existing compendium
8. the increased **ability to predict known regulatory motifs** by incorporation of the additional experiments
9. the increase in the **ability to predict the activity pattern** of a given element resulting from incorporation of the additional experiment in an existing data compendium

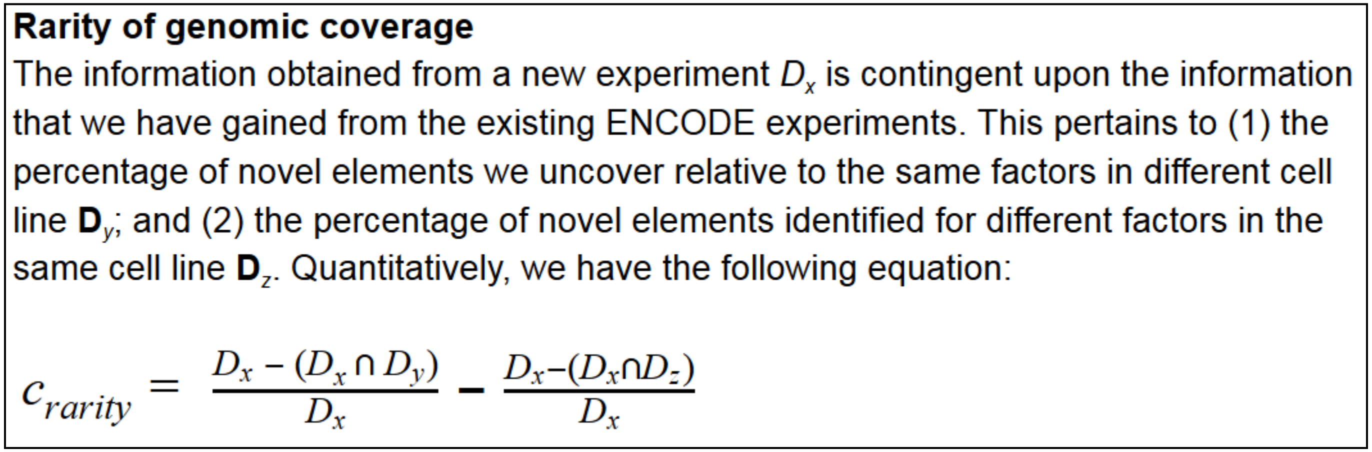
Factors influencing these properties

a) the type of assay;
b) the specific cell type selected;
c) the experimental conditions used;
d) the quality of antibodies (when applicable);
e) the cell type heterogeneity of the sample;
f) the sequencing depth at which the experiment is carried out;
g) the amount of DNA extracted (and thus effective depth of the library).

ENCODE Imputation strategy: 4 stages



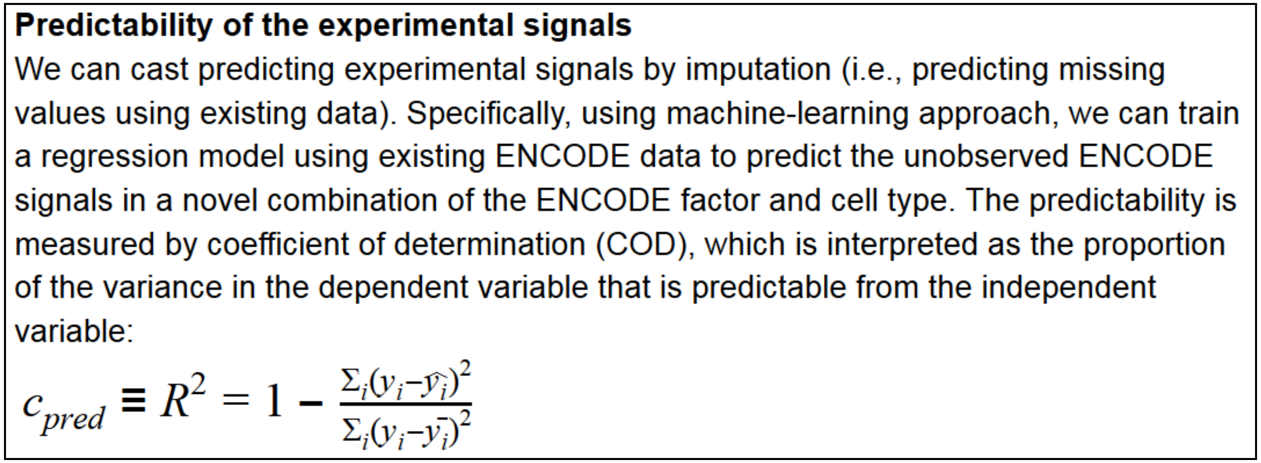
More concretely…



Rarity of genomic coverage

The information obtained from a new experiment D_x is contingent upon the information that we have gained from the existing ENCODE experiments. This pertains to (1) the percentage of novel elements we uncover relative to the same factors in different cell line $\mathbf{D}_y$; and (2) the percentage of novel elements identified for different factors in the same cell line $\mathbf{D}_z$. Quantitatively, we have the following equation:

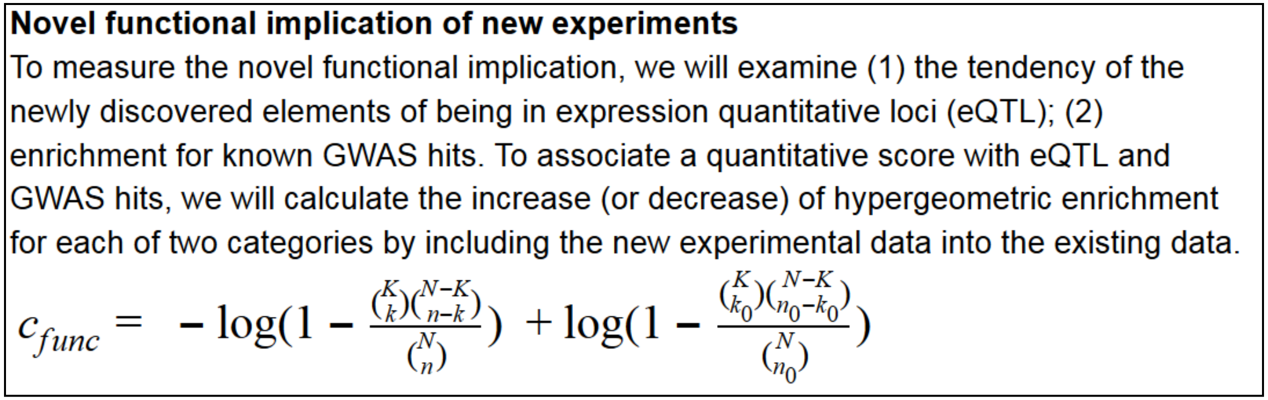
$$c_{rarity} = \frac{D_x - (D_x \cap D_y)}{D_x} - \frac{D_x - (D_x \cap D_z)}{D_x}$$



Predictability of the experimental signals

We can cast predicting experimental signals by imputation (i.e., predicting missing values using existing data). Specifically, using machine-learning approach, we can train a regression model using existing ENCODE data to predict the unobserved ENCODE signals in a novel combination of the ENCODE factor and cell type. The predictability is measured by coefficient of determination (COD), which is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable:

$$c_{pred} \equiv R^2 = 1 - \frac{\Sigma_i(y_i - \hat{y_i})^2}{\Sigma_i(y_i - \bar{y_i})^2}$$



Novel functional implication of new experiments

To measure the novel functional implication, we will examine (1) the tendency of the newly discovered elements of being in expression quantitative loci (eQTL); (2) enrichment for known GWAS hits. To associate a quantitative score with eQTL and GWAS hits, we will calculate the increase (or decrease) of hypergeometric enrichment for each of two categories by including the new experimental data into the existing data.

$$c_{func} = -\log\left(1 - \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}\right) + \log\left(1 - \frac{\binom{K}{k_0}\binom{N-K}{n_0-k_0}}{\binom{N}{n_0}}\right)$$

(…)

Deliverables proposed

1. present a framework that incorporates each of these metrics in a formal information-theoretic framework;
2. systematically apply these metrics to the ENCODE 2 and ENCODE 3 compendiums to evaluate the information gained by each dataset;
3. summarize the lessons learned from this systematic application on the value of different experiment types and different cell types;
4. make predictions for the most informative experiments to carry out going forward, including assays, cell types, and sequencing depth;
5. provide a series of tools for enabling such analyses more broadly.

DAC Status update

4. Analysis of ENCODE portfolio by cell type

- Action: The DAC will analyze the ENCODE Portfolio by cell type and determine what space ENCODE has and has not covered (5/21)

5. Tracking ENCODE Element Identification Over Time (NHGRI and DAC)

- Action: The DAC will provide the NHGRI team with a plan for tracking ENCODE element identification. (4/16)

6. Cell identity testing (DAC)

- Action: The DAC will develop and apply methods for automatically testing the identity of cell types

Identifying mixups in NGS datasets

1. Larger datasets increase chance of swaps
2. Can have huge effect on conclusions – may manifest as “interesting” results
   - Has happened to me
3. Investigated for eQTL datasets
   - None (to my knowledge) for epigenetic data
4. May also be useful for identifying low quality datasets

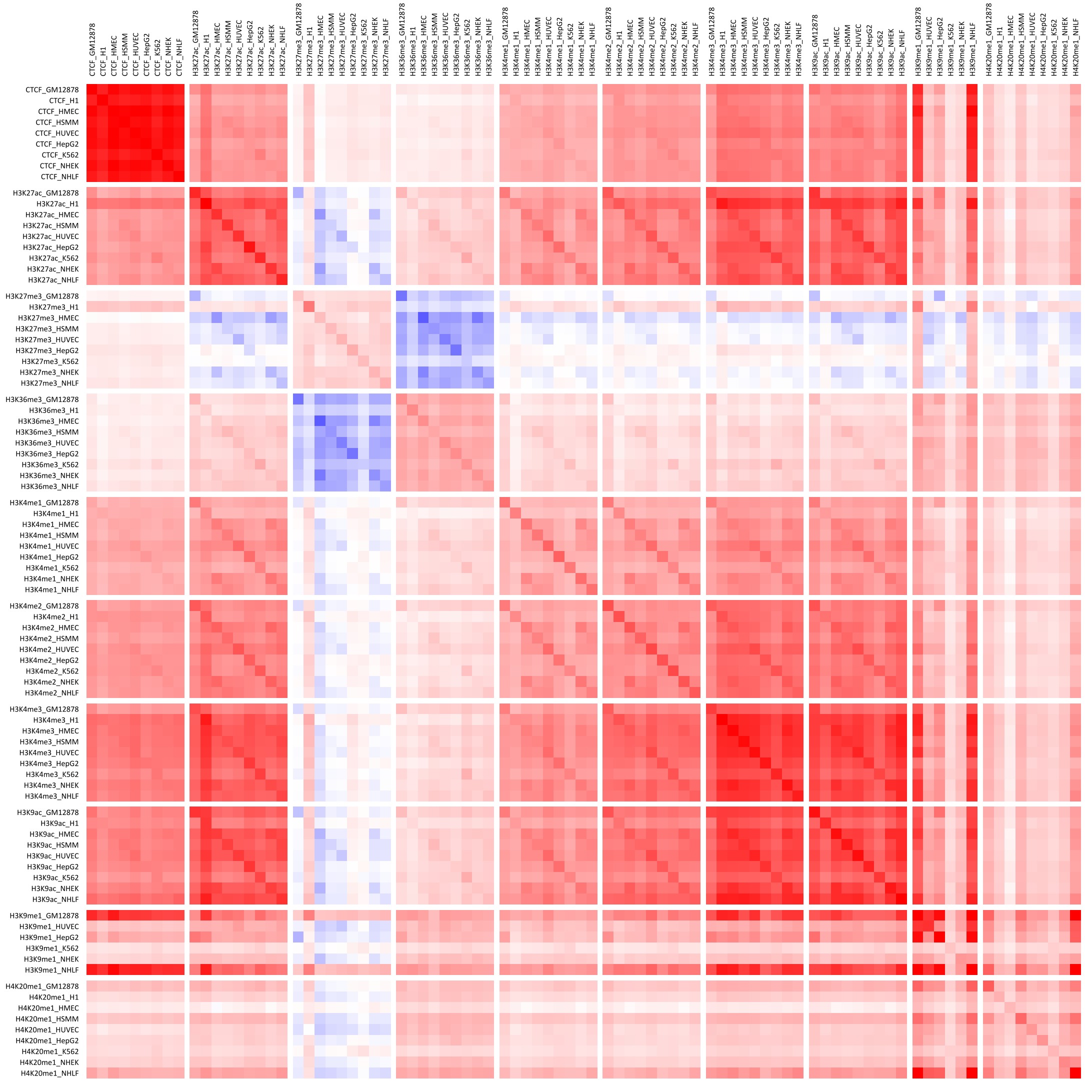
87 input epigenetic datasets

Cell Types
H1
K562
GM12878
HepG2
HUVEC
HSMM
NHLF
NHEK
HMEC

Marks
CTCF
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K9me1
H4K20me1

- Epigenetic data from ENCODE2 (Ernst, et al. 2011)
- Complete matrix except H3K9me1_H1, H3K9me1_HSMM, H3K9me1_HMEC

Score #1: Peak overlap enrichment

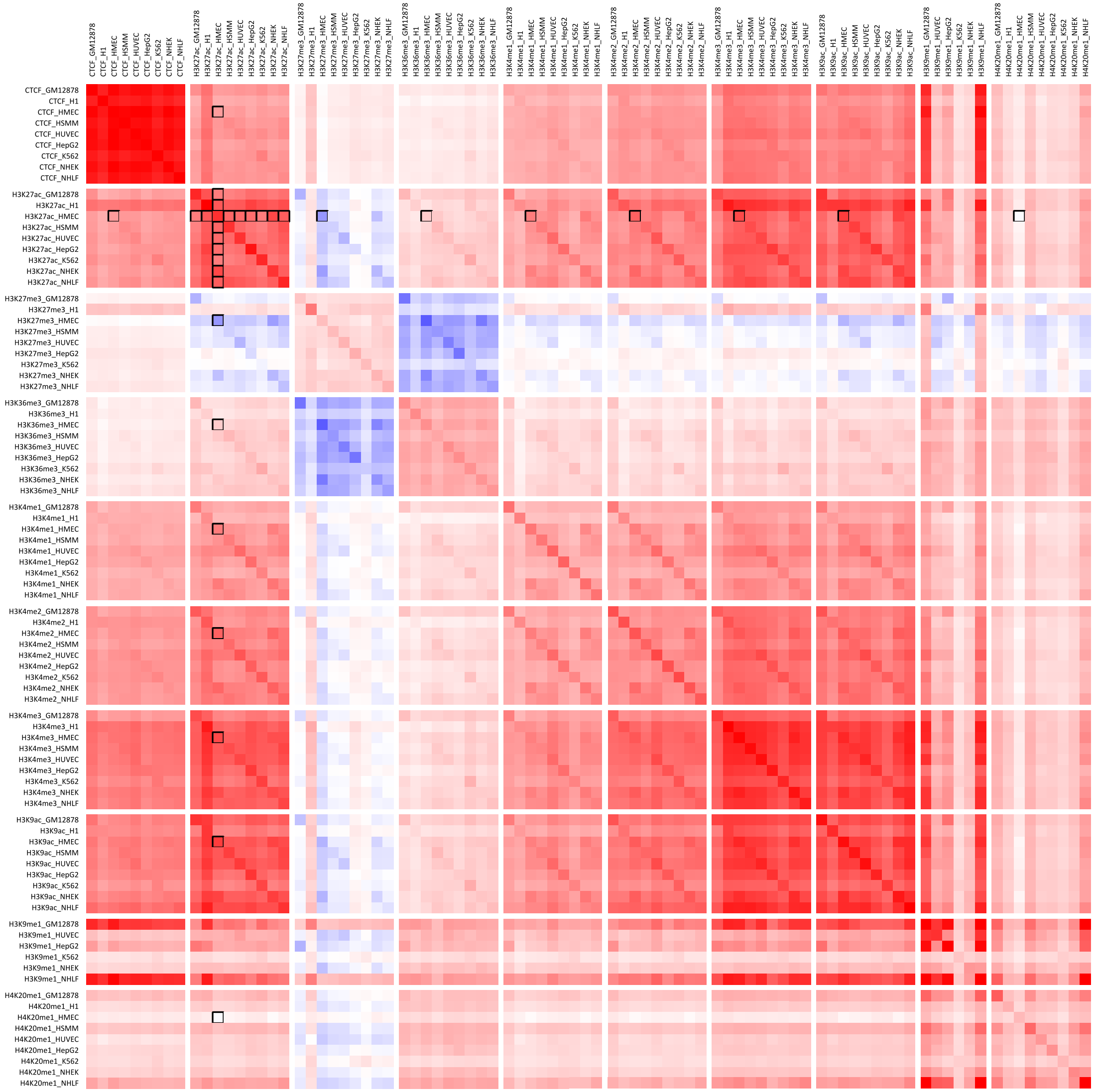


- Log enrichment of peaks of one dataset in the peaks of another
- Each group of rows/columns is a specific mark
- Clear increase in enrichment for matching mark, and more subtly for matching cell type
- Can this be used to identify sample swaps?

Strategy for identifying sample swaps

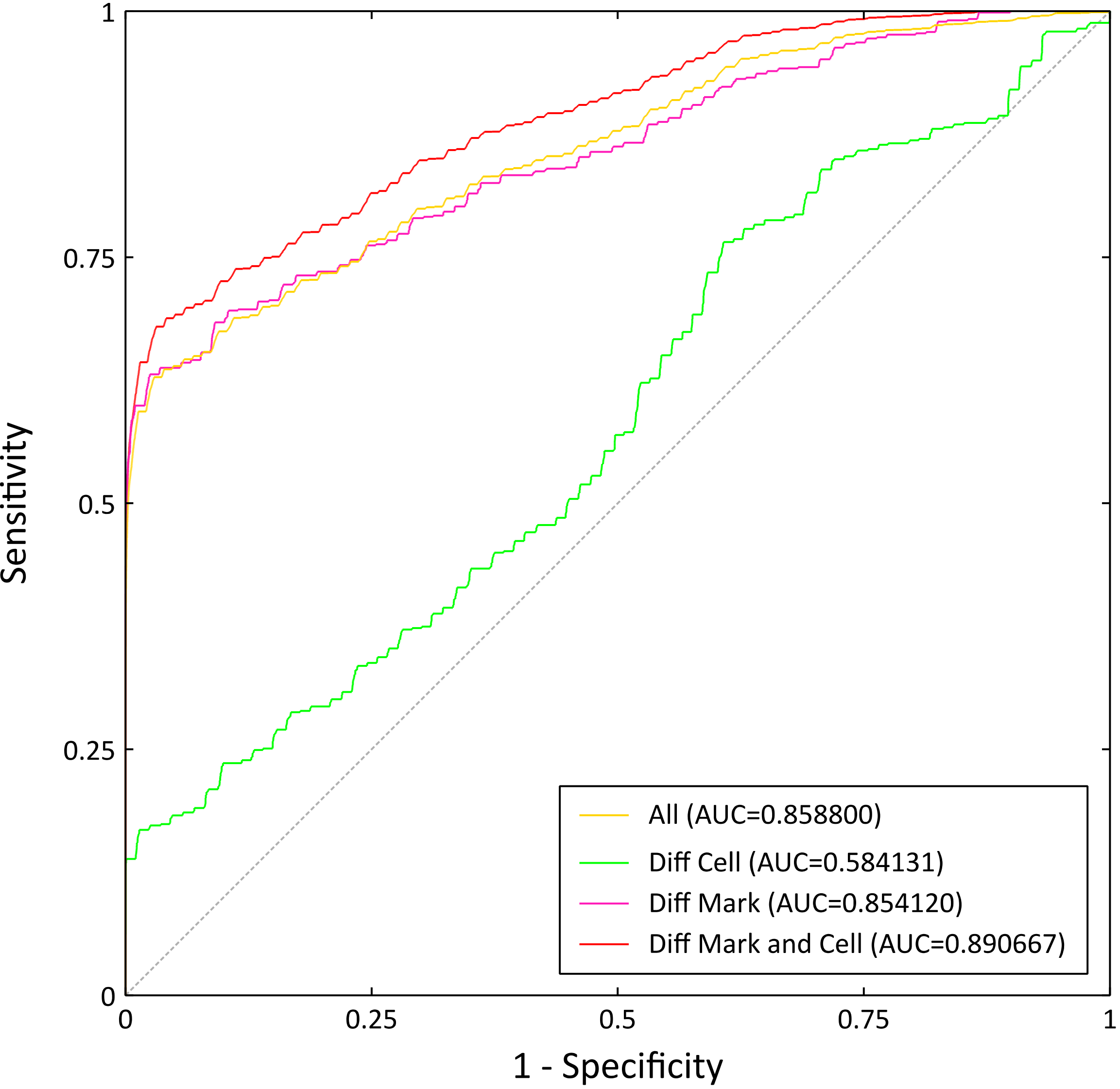
- Compute similarity for every pair of datasets
  - will discuss four today
- Produce dataset’s **consistency score**
  - Compute average similarity to all datasets in the same set
  - Set can be datasets with the same mark, cell type, or either
  - Subtract average similarity to all other datasets
- Artificially swap all pairs to measure performance (AUC)
  - Note: each swap can effect the score of other datasets

Score #1: Peak overlap enrichment consistency



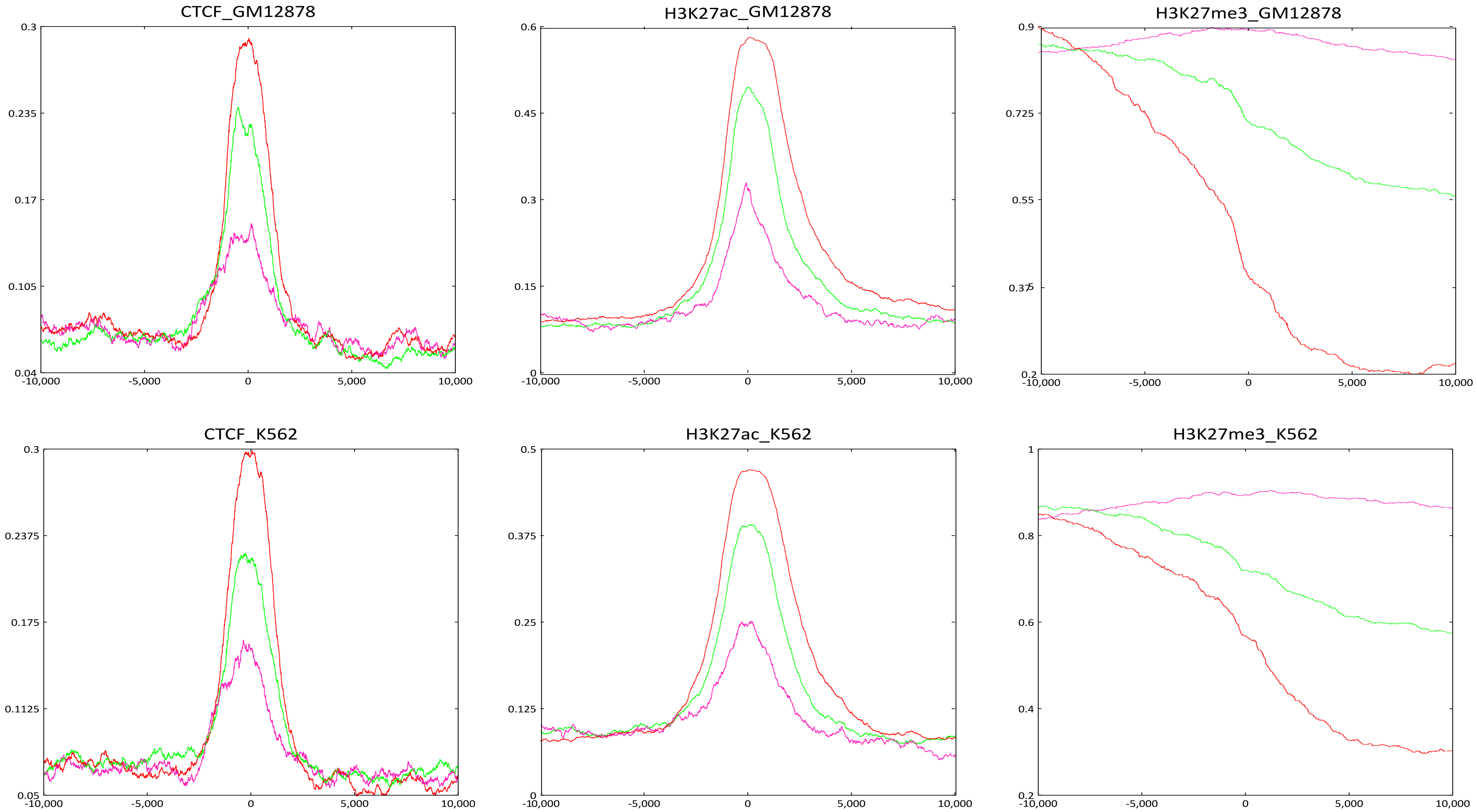
- Single number for each dataset
- Average score to all datasets with same mark or same cell minus average to all other datasets
- Easy to simulate sample swaps – how well can we find them?

Score #1: Peak overlap enrichment ROC



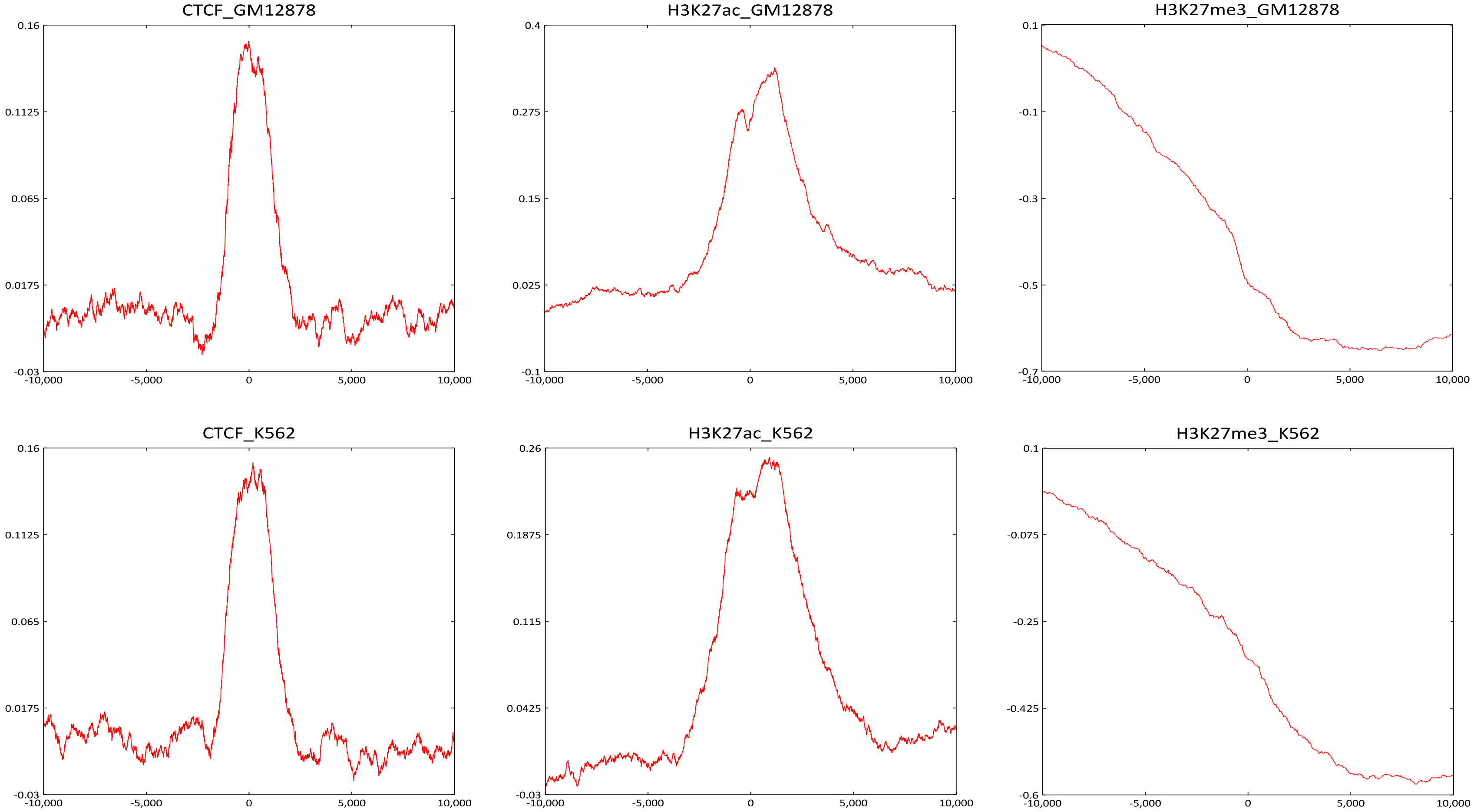
- Perform all 3741 = 87 choose 2 swaps
- Use consistency score to differentiate positive (swapped) to non-swapped datasets
- Overall AUC of 0.85 in identifying swaps
- Virtually no false positives at 50% sensitivity
- Poor performance in identifying swaps when mark does not change

Score #2: TSS profile of marks

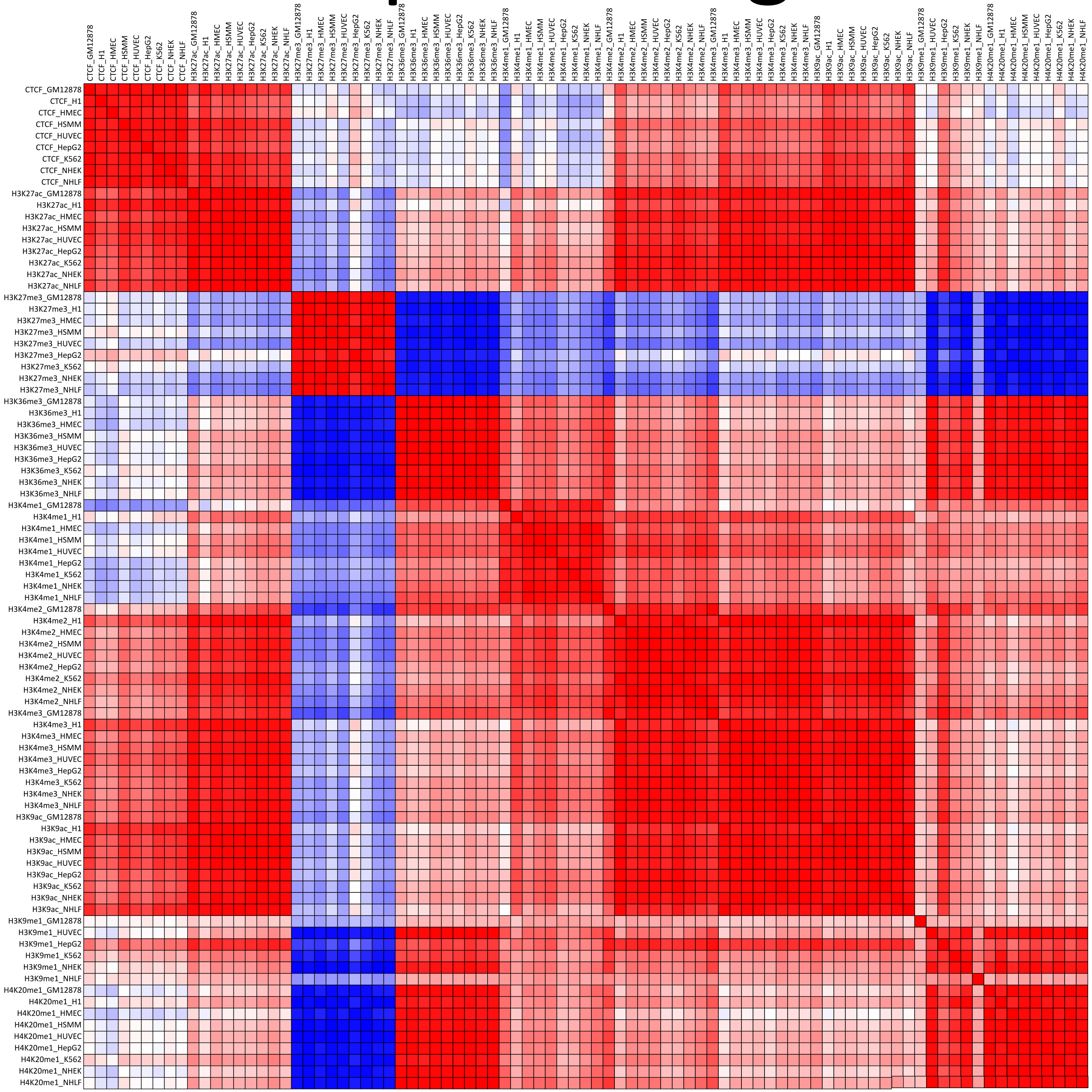


Mean peak density as function of distance from TSS for high (red), mid (green), and low (pink) expressed genes (expression is average across all cell types)

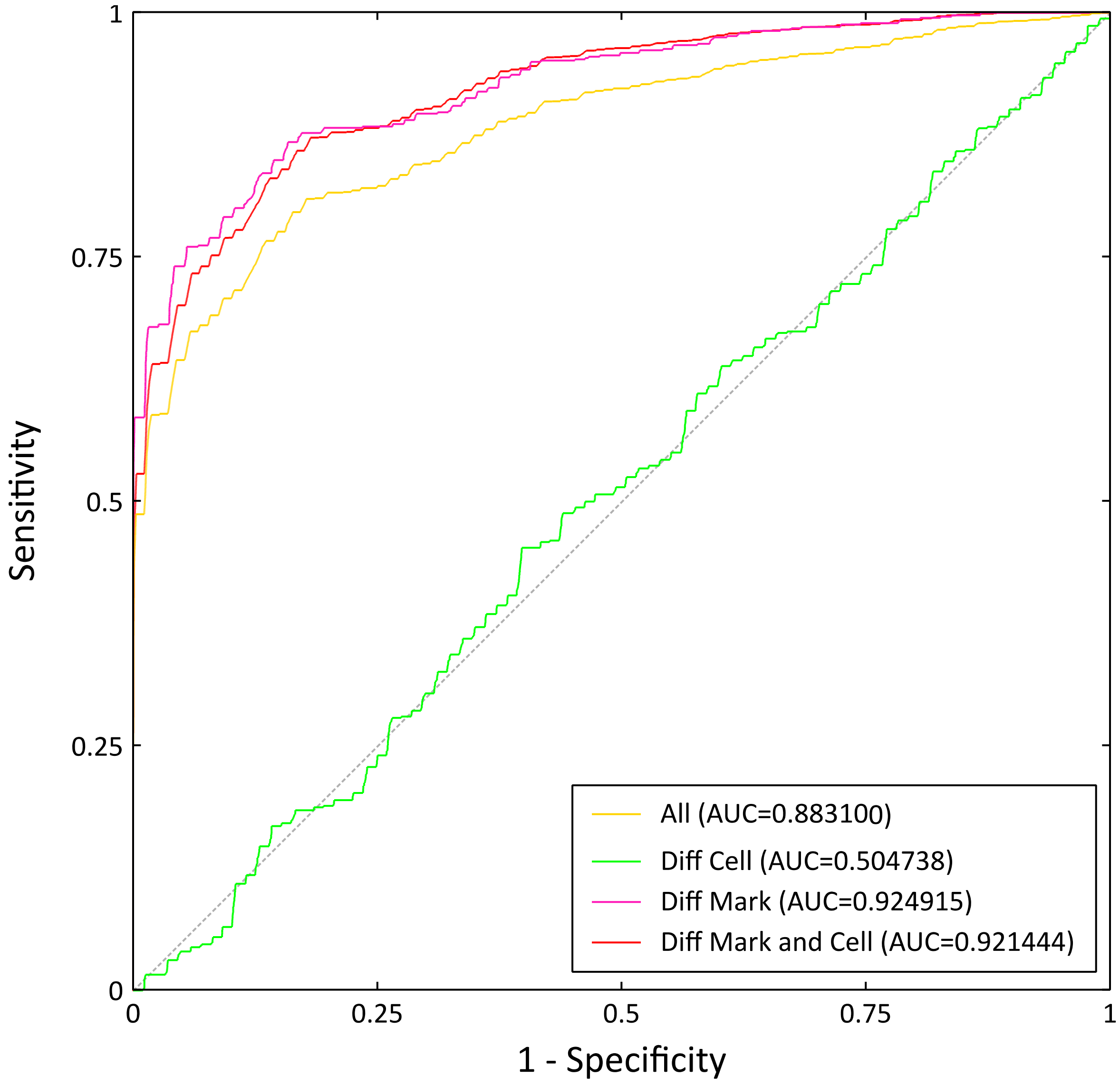
Score #2: TSS profile high minus low



Score #2: TSS profile high minus low



TSS profile can identify swapped marks

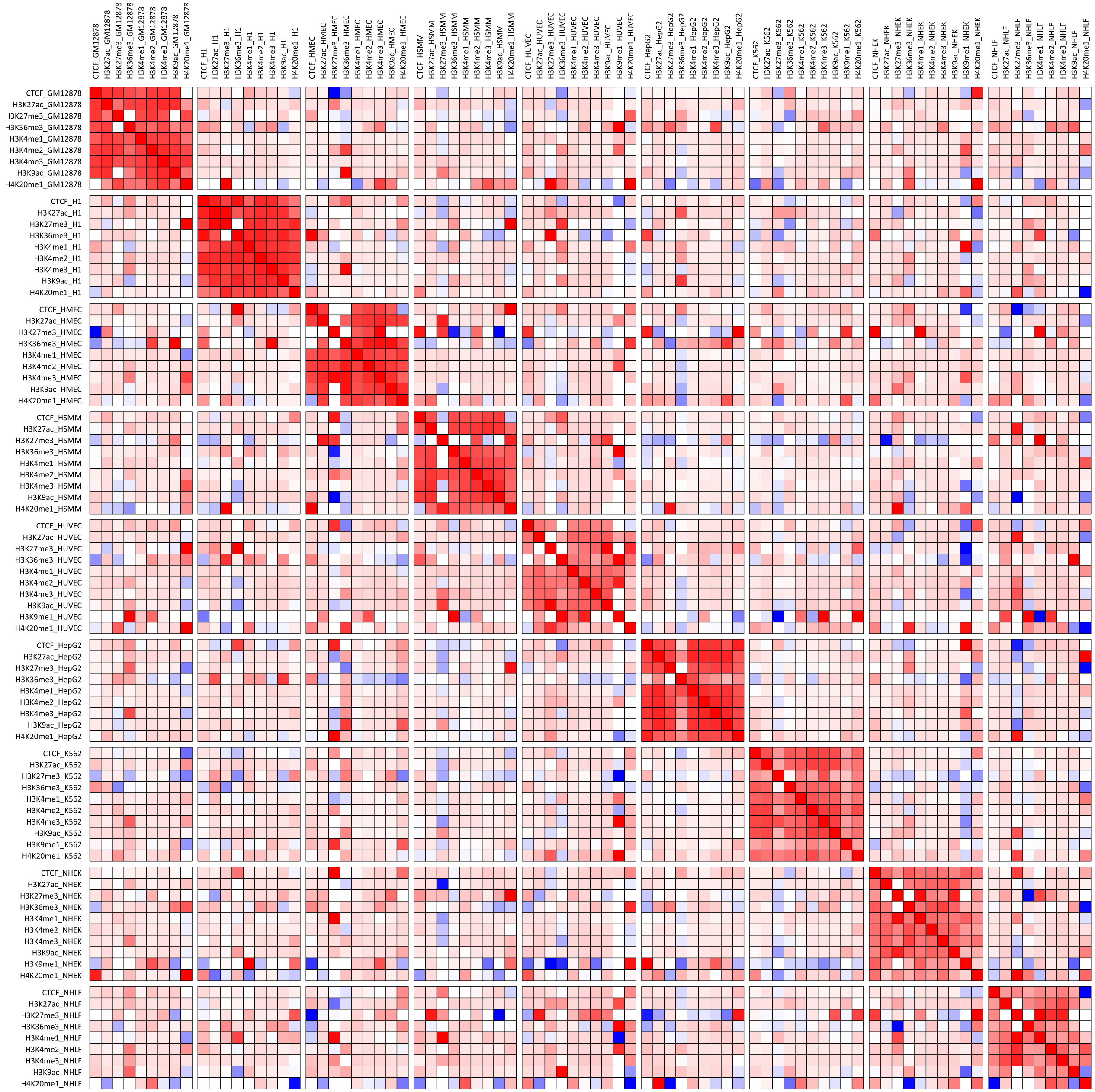


- Same procedure as with peak overlaps
- All 3741 swaps
- Roughly same overall AUC
- Better at distinguishing marks
- Cannot distinguish cell types

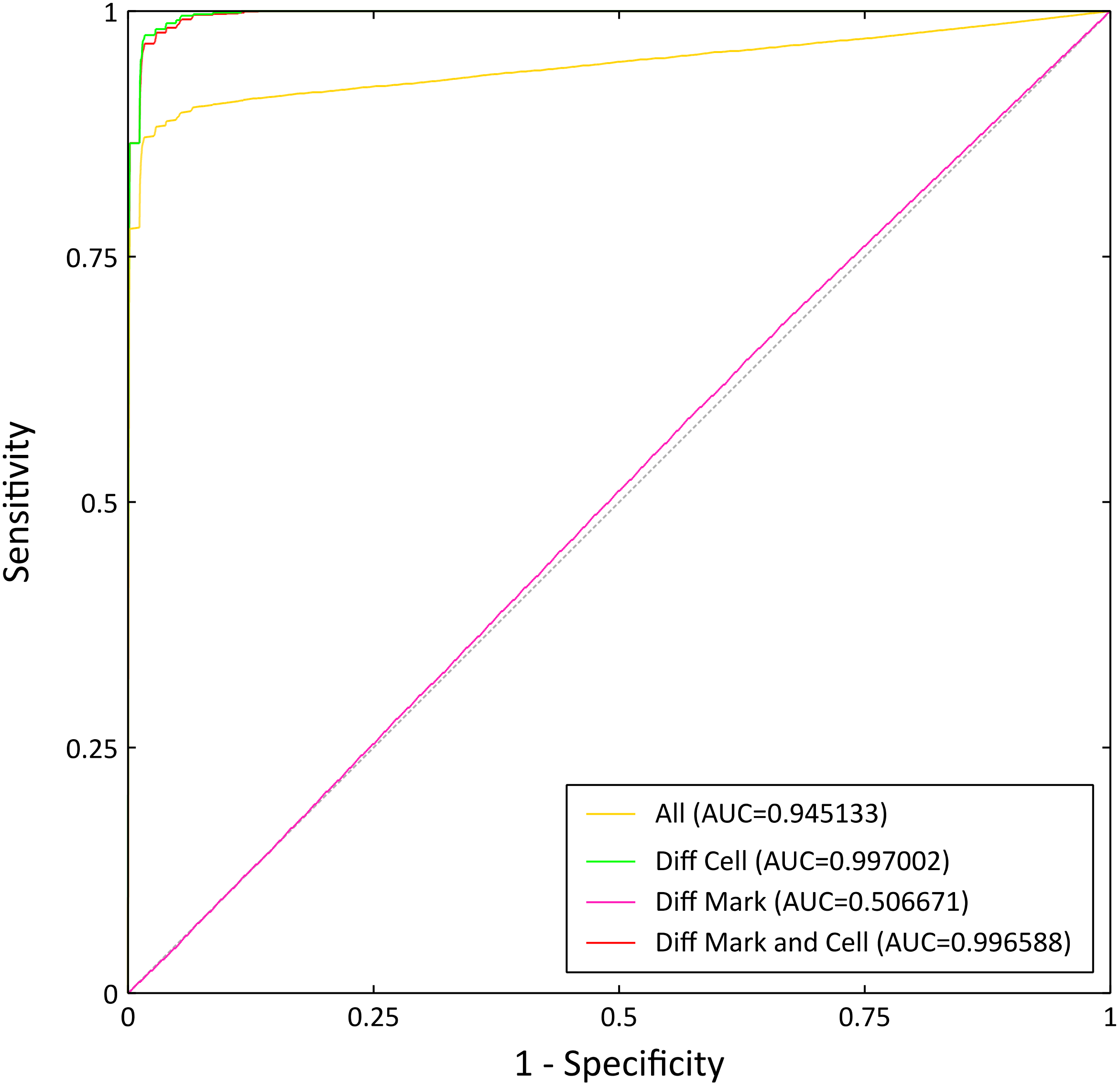
Score #3: Genetic evidence to identify cell swaps

1. Because we have raw reads, we can also look at SNPs to identify origin cell type
2. Count bases seen at reads for each of 660k snps on HapMap 650v3 array
3. Compute fraction of reads corresponding to the most observed base
   - Essentially building a vector of heterozygous vs. homozygous sites
4. For each pair of datasets, correlate all positions that have at least 5 reads in both

Score #3: Genetic consistency similarity

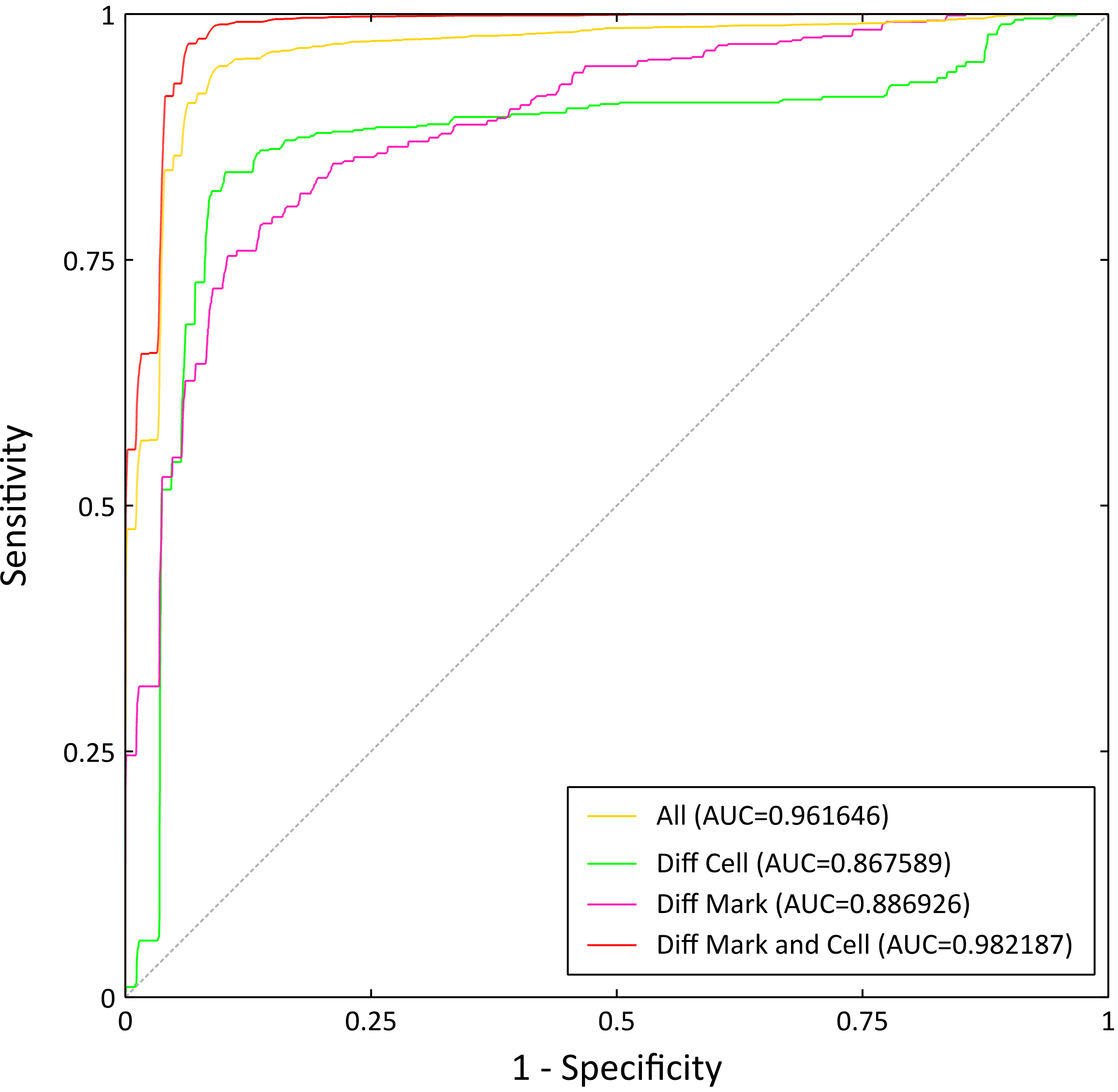


Score #3: Genetic evidence finds nearly all cell type swaps



- Consistency score produced using datasets of the same cell type
- Nearly perfect in identifying cell type swaps
- No power to identify mark swaps

Score #4: Genetic + TSS profile more balanced



- Simple mean of genetic and TSS profile similarity values
- Worse than genetic/tss at cell type/marks, but better overall

Challenges to using correlation for genetic evidence

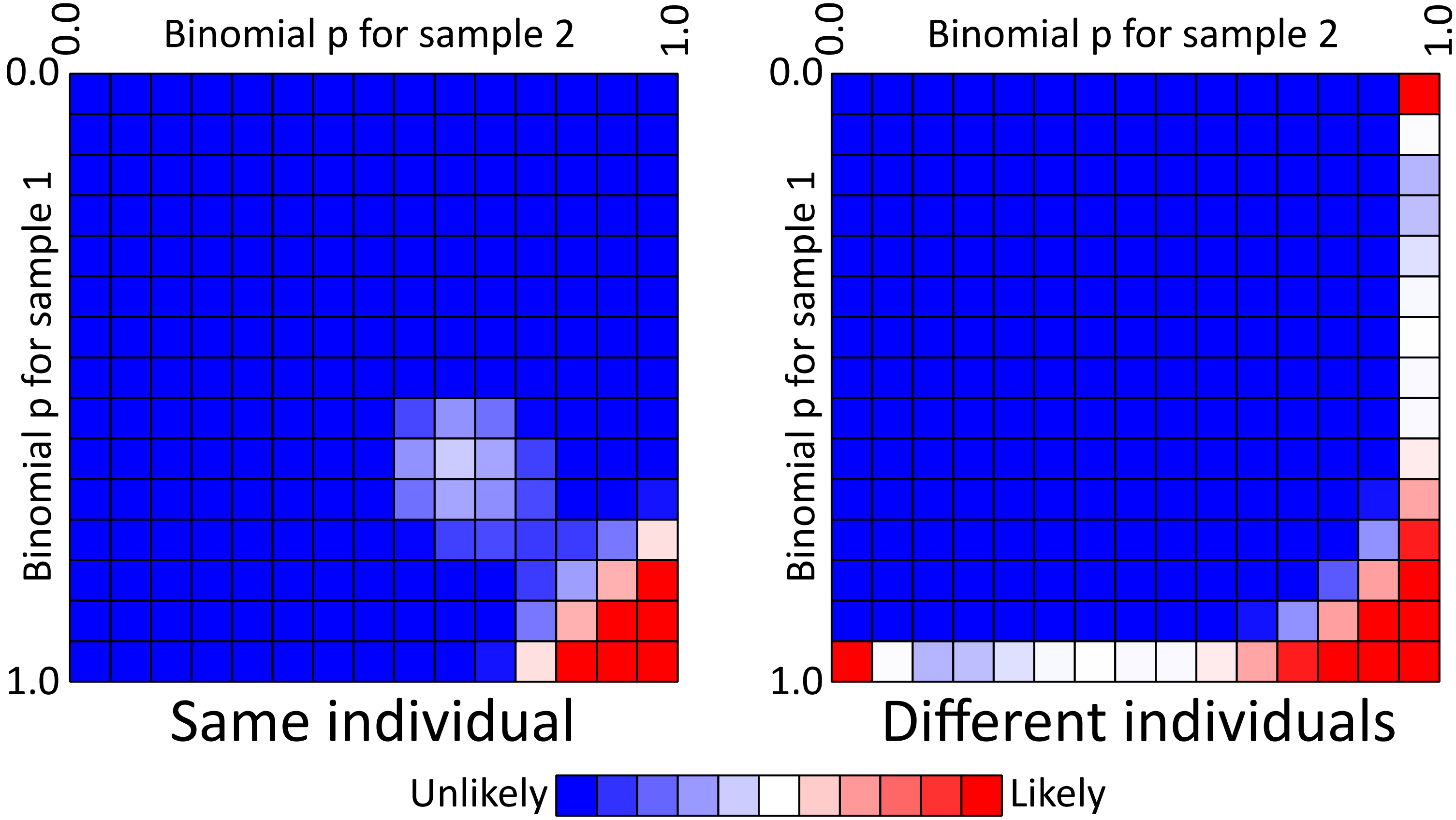
- Some pairs of samples only have a few positions in common
  - Makes correlations unreliable
- The positions that are in common are not comparable
  - Some have many reads, some have very few

→ Log likelihood ratio of trained models

EM Trained models for positions from (mis)matched individuals

- Consider a position with reads for two individuals (a,b)
- There are two alleles
  - Let 1 be the more observed allele
- We have two ratios: $a_1/(a_1+a_2)$, $b_1/(b_1+b_2)$
- Use EM to train mixture of a binomials to fit the observed ratios
  - Separately for matched, mismatched individuals

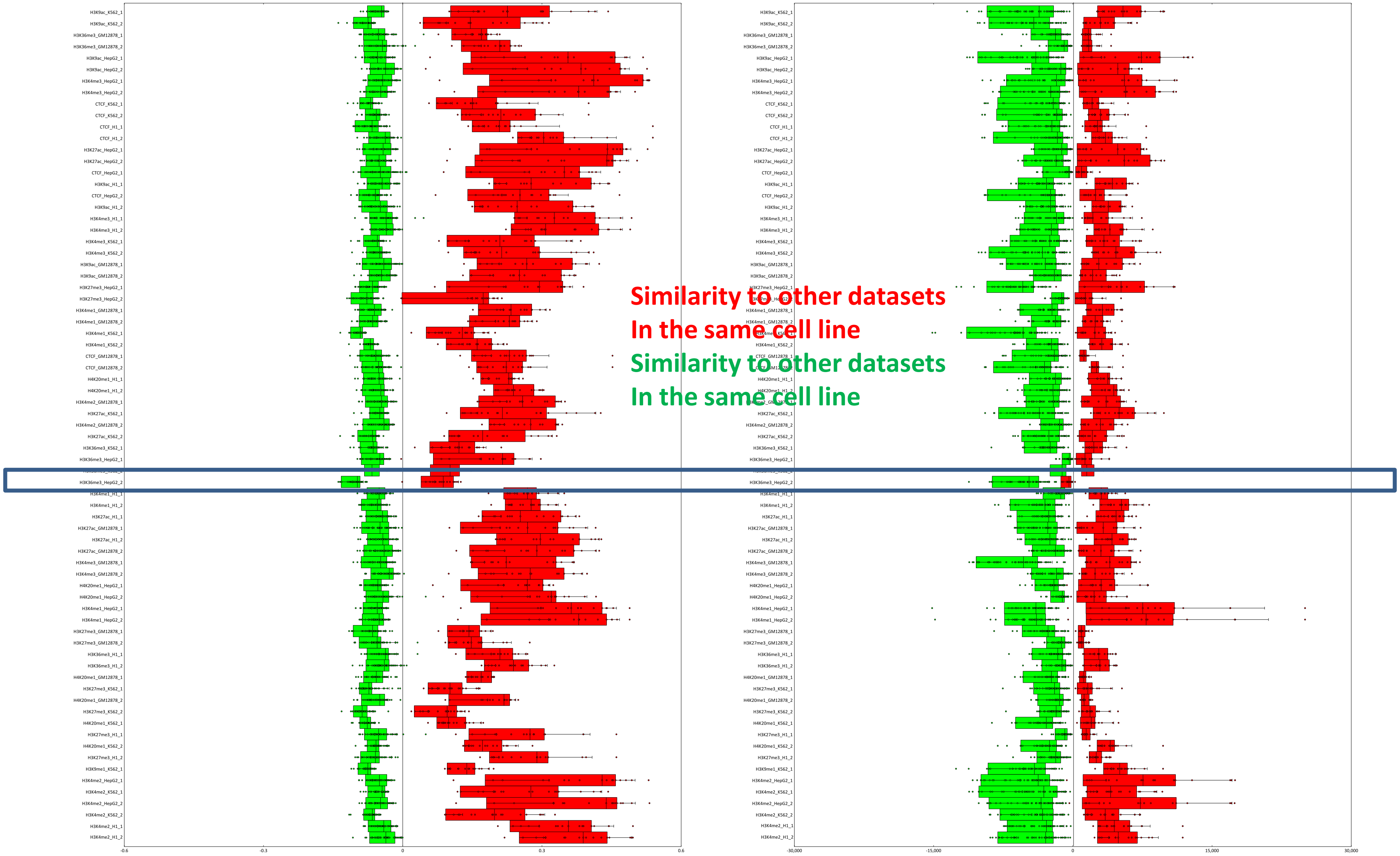
EM Trained models for positions from (mis)matched individuals



- Ratio summed across positions with reads for both samples
- Sign indicates same/different individual
- Magnitude indicates confidence

Worse performance with ENCODE data

(GM12878, H1, K562, HepG2 from ENCODE2)



Similarity to other datasets
In the same cell line

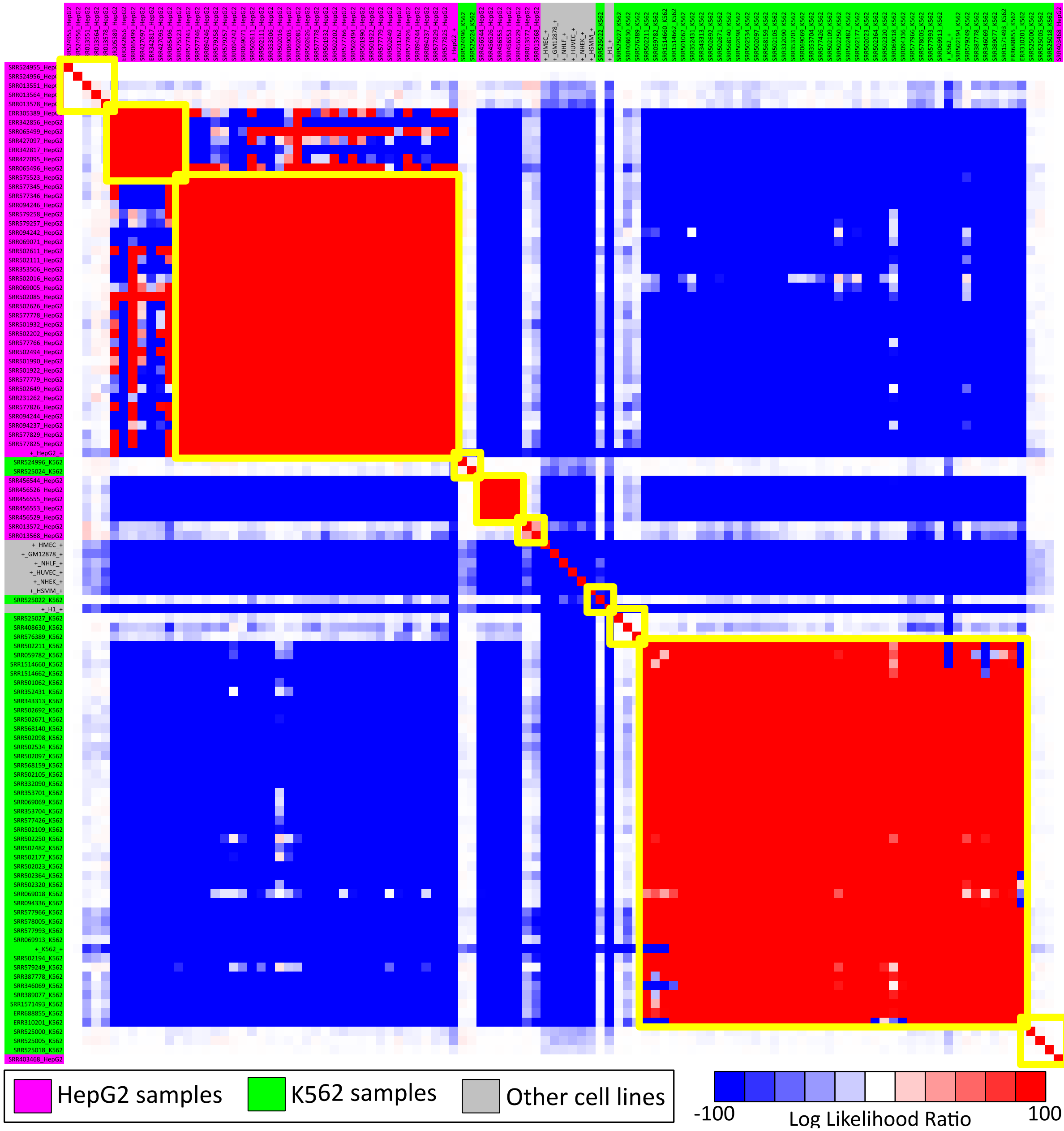
Similarity to other datasets
In the same cell line

Correlation (AUC=0.999967)

EM Matrix (AUC=0.998087)

Applying genetic evidence to SRA

- RNA/ChIP-seq data from SRA
  - 50 each K562/HepG2 + ENCODE
- Most datasets match what is expected
  - Some not consistently. Mixed samples?
- Some match each other, but not most of the same cell line
- A few datasets on their own



Conclusion

- Sample swaps are a common occurrence with large datasets
  - Swaps may occur by the vendor providing cells
- Being able to identify them automatically would be very useful
- Genetics provides very strong evidence of swaps between samples originating from different individuals

Future directions

1. Improved composite score
   – ML approach to finding discriminating features
2. Improvement to genetic score to deal with sparse datasets
   – Normalize individual datasets error rate?
3. Run analysis on ENCODE3 datasets
4. Distribute tools for performing analysis