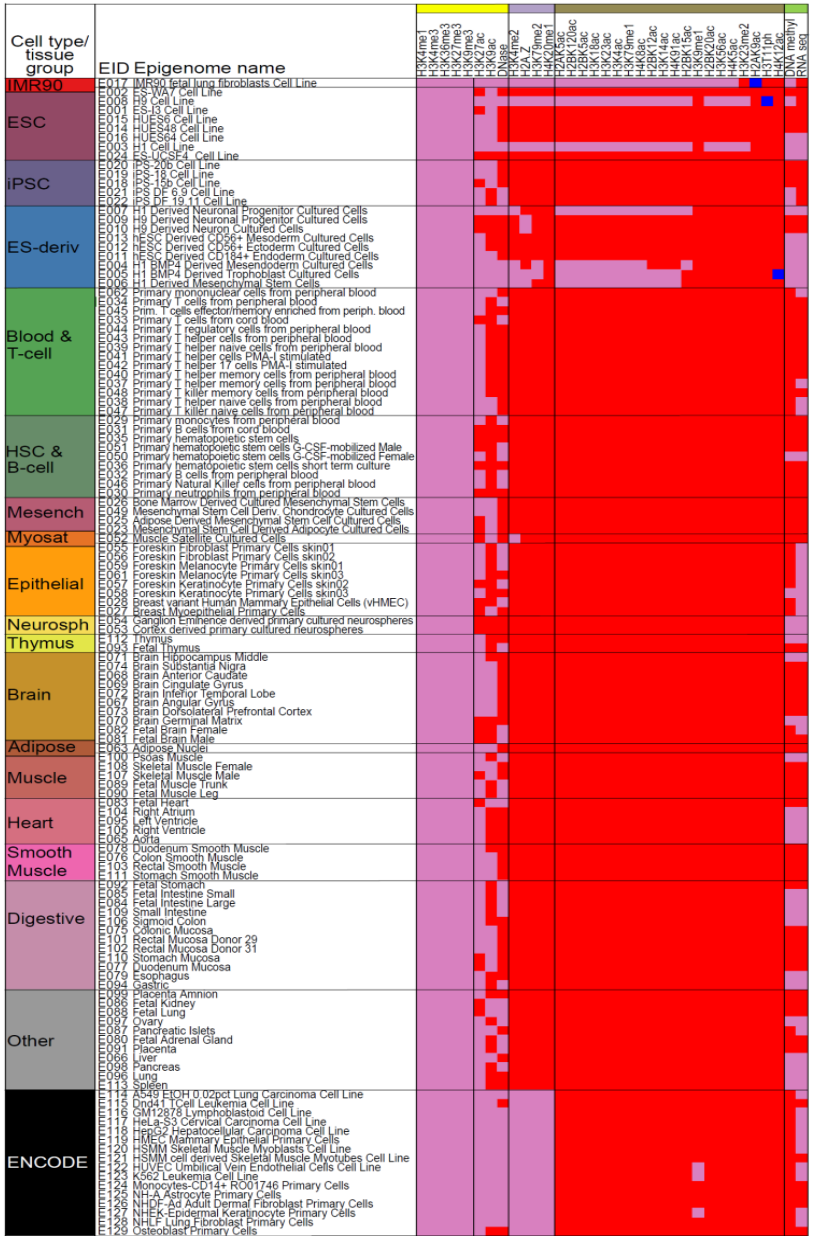
Large-scale Epigenomic Imputation

Jason Ernst

Assistant Professor

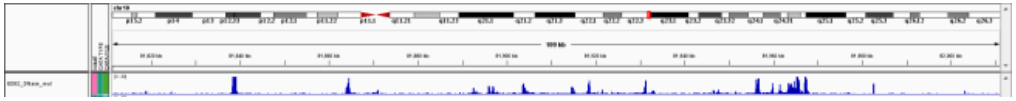
University of California, Los Angeles

Epigenomic Imputation Problem



Problem: Predict mark, cell type data genome-wide assuming no data for the dataset we are trying to predict

- Complete big (mark, tissue) data matrix
- Combines potentially hundreds of datasets to generate more robust and higher quality versions of observed data sets



ROADMAP epigenomics PROJECT

ENCODE

111 Roadmap Epigenomics; 16 ENCODE

Ernst and Kellis, *Nature Biotech* 2015

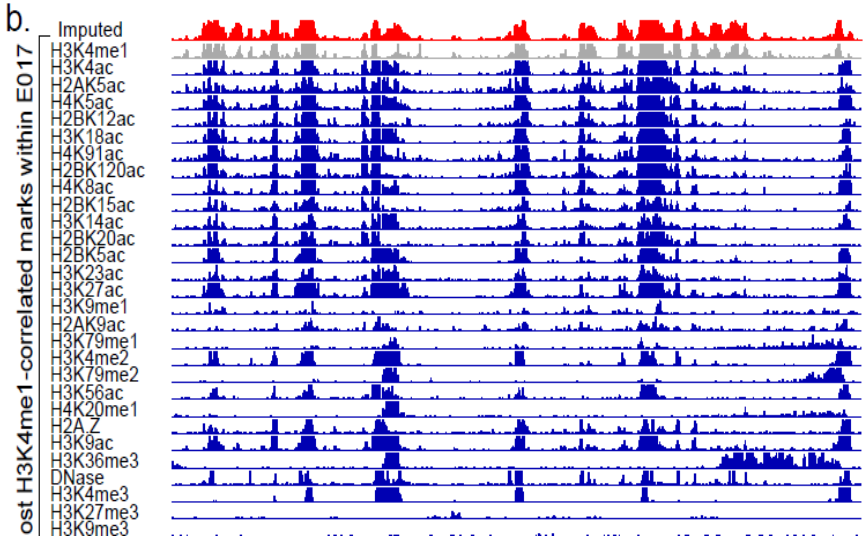
Observed + Imputed

Imputed Only

Observed Only

ChromImpute: Two classes of features

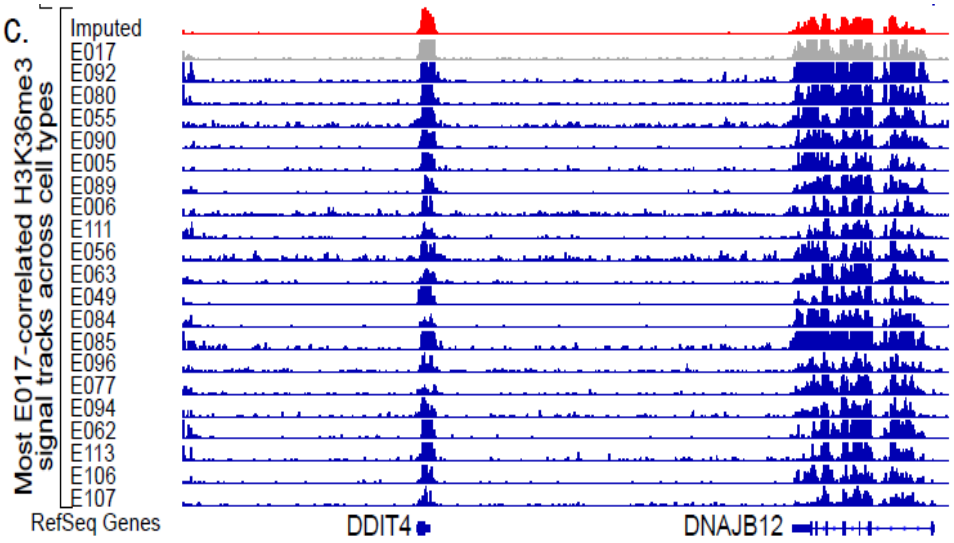
Other marks in same tissue



Features for a mark

- At target position and every 25bp left and right until 500bp.
- At 500bp and every 500bp left and right until 10000 bp.

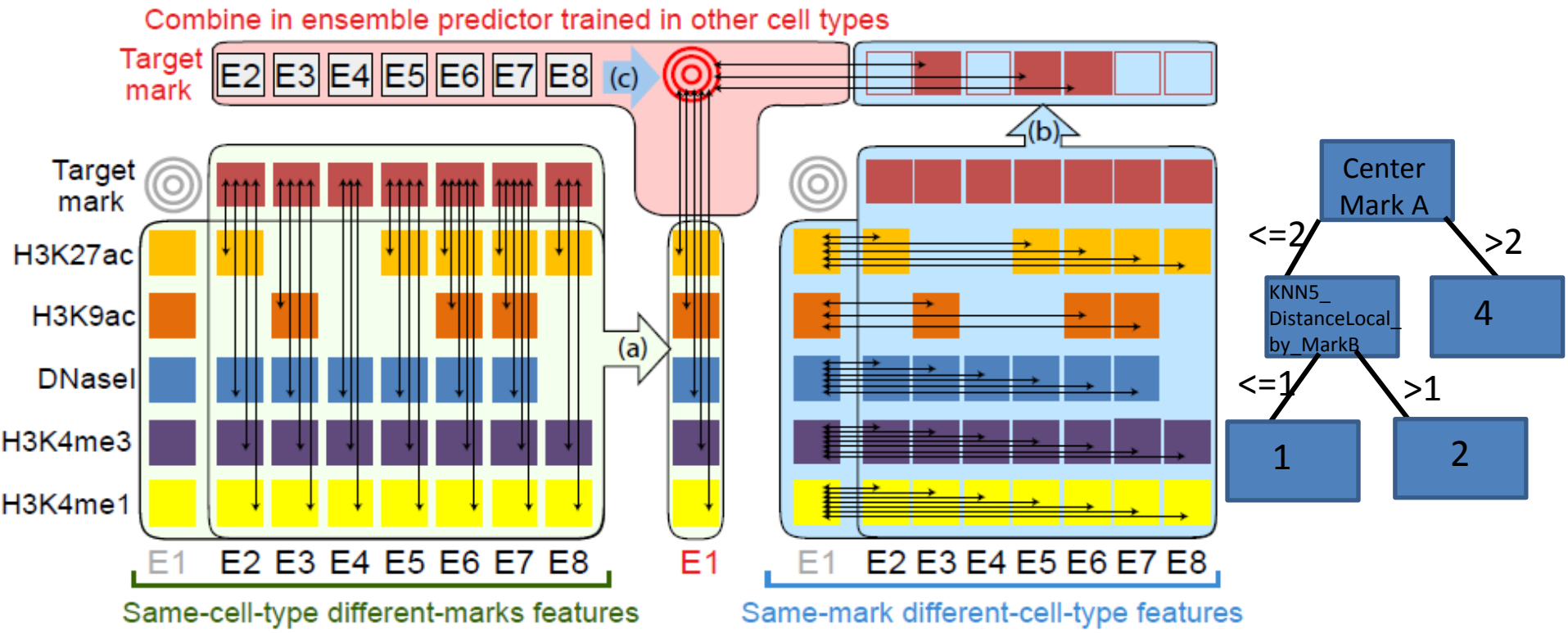
Same mark in other tissues



Features:

- Average target mark signal at target position in K-nearest epigenomes for K=1,…,10
- Separate set of features for distance defined based on each mark in target epigenome and local and global distance

ChromImpute: Training and Prediction strategy

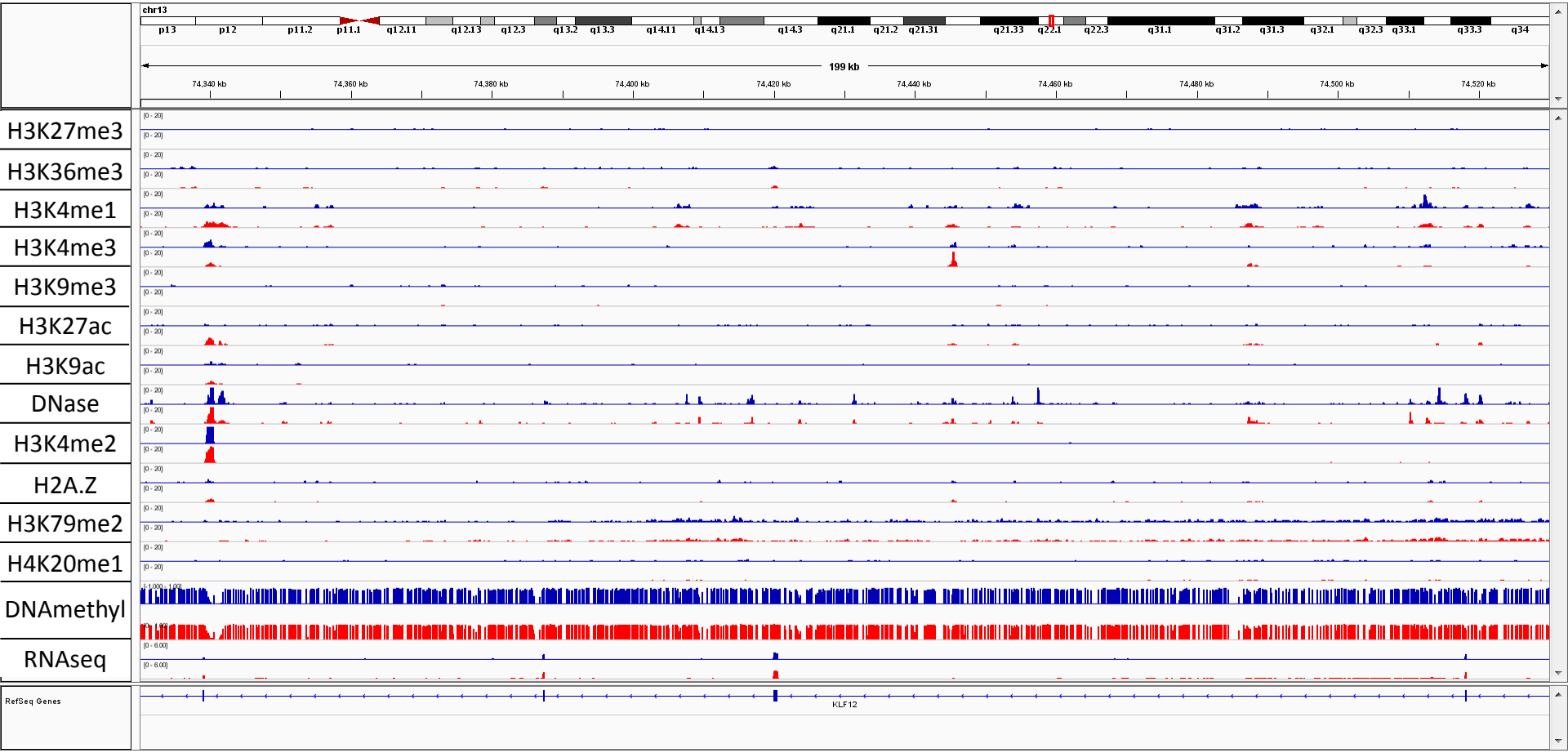


- Assume **no training data** for target mark in target epigenome
- **Separate regression tree(s)** for each epigenome where mark is available
- Restrict features to **common marks** between target and informant tissue
- Apply each regression tree to target epigenome and **average** predictions

Browser Visualizations

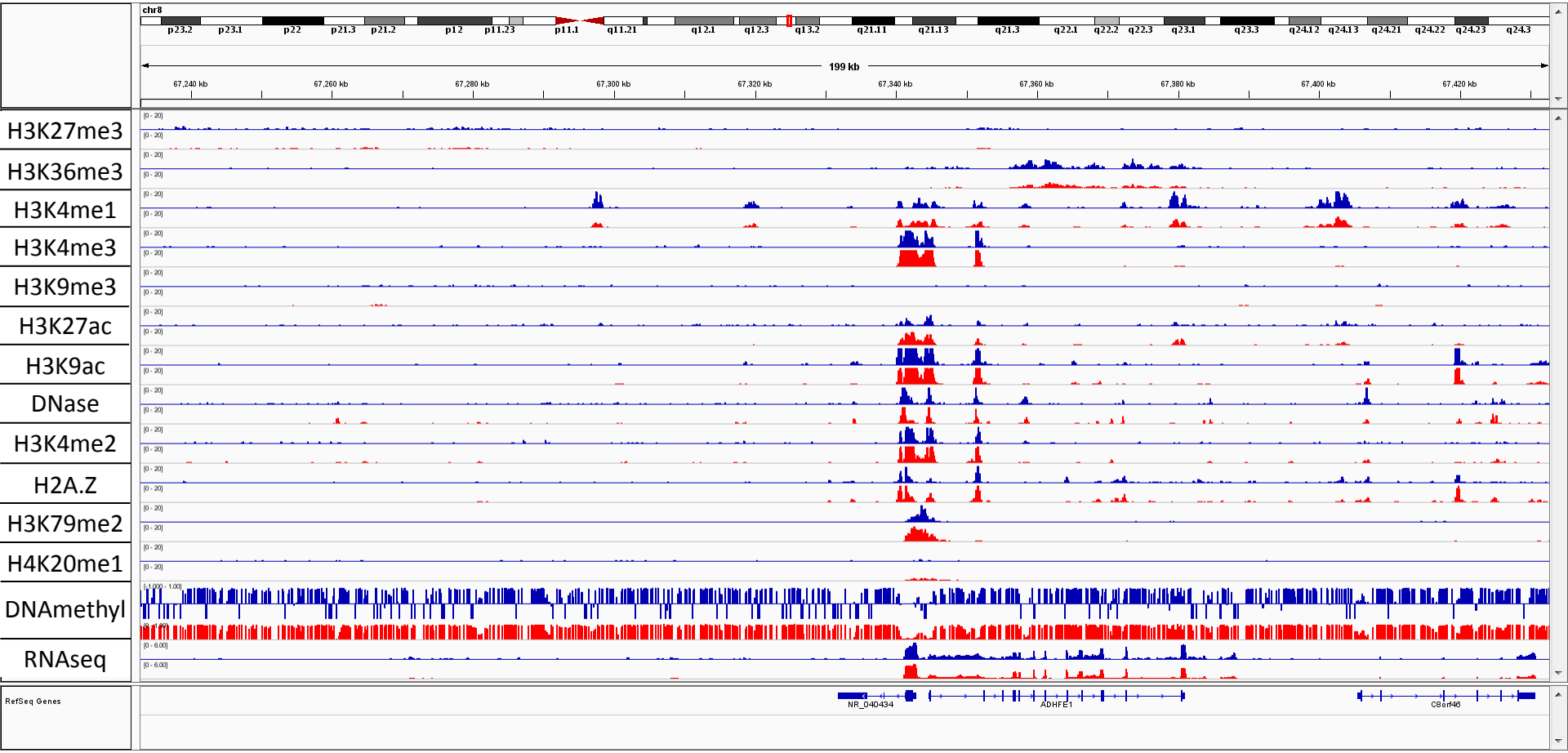
- Randomly selected 9 -200kb regions to visualize and one sample for each mark

Browser Screenshots of Random Loci



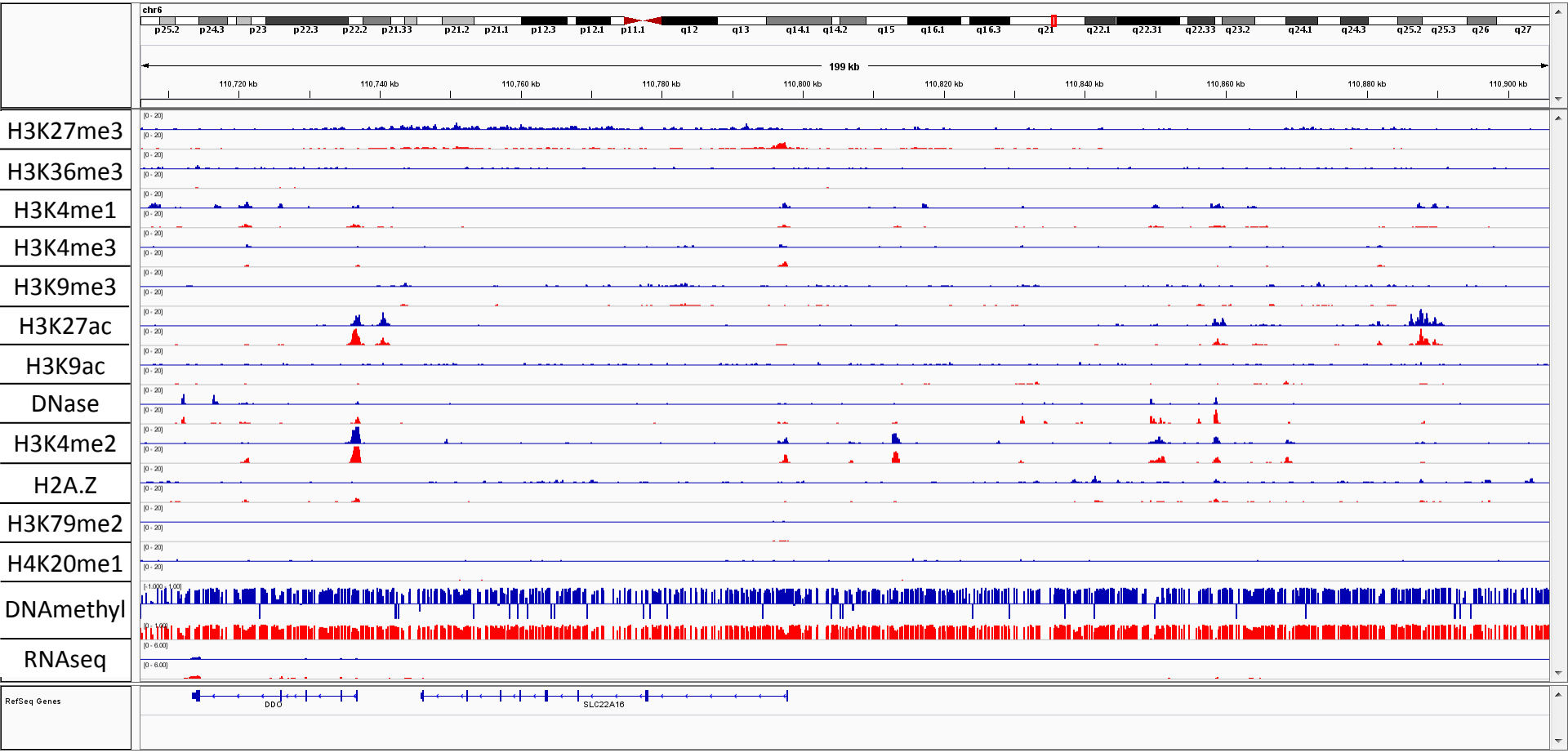
Blue observed; Red Imputed

Browser Screenshots of Random Loci



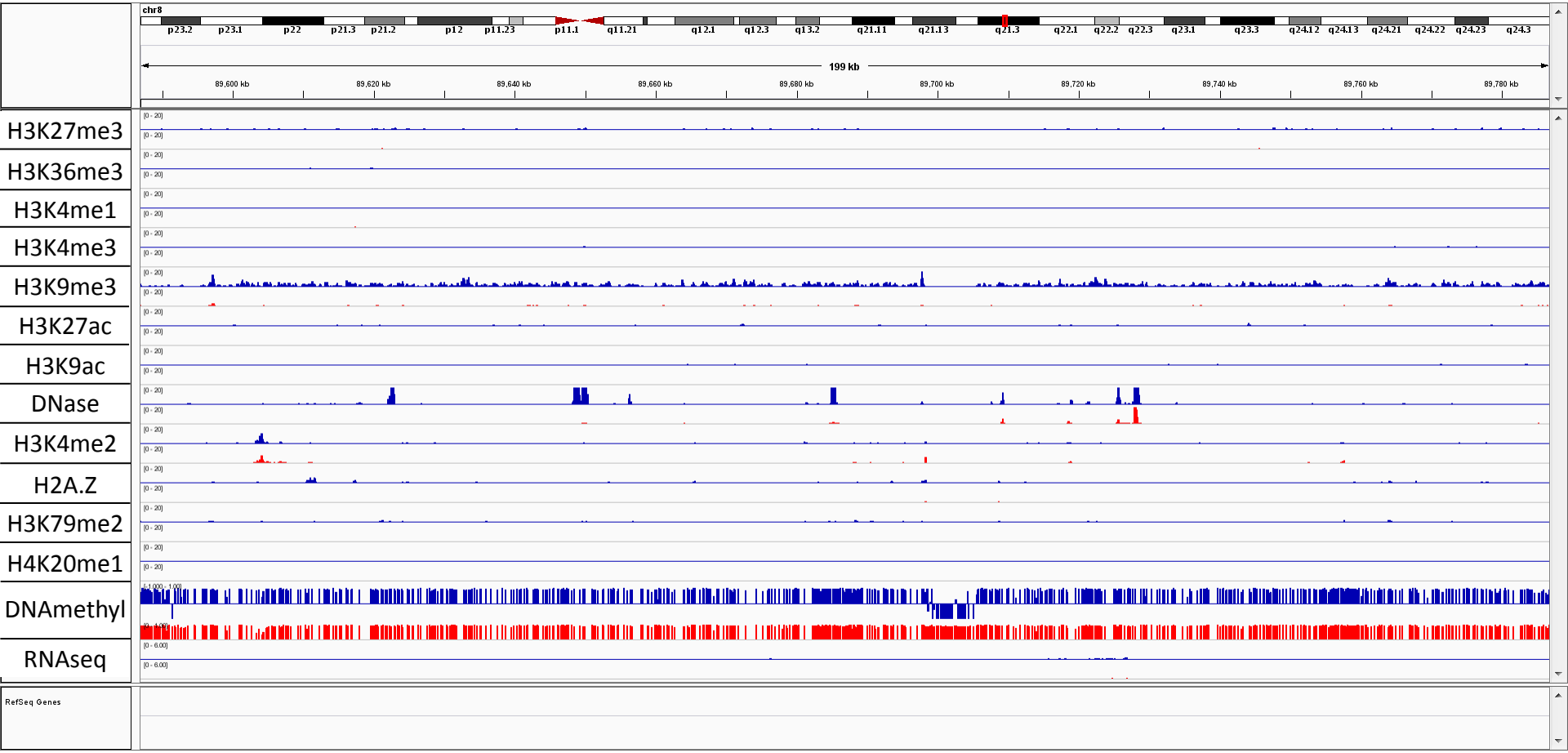
Blue observed; Red Imputed

Browser Screenshots of Random Loci



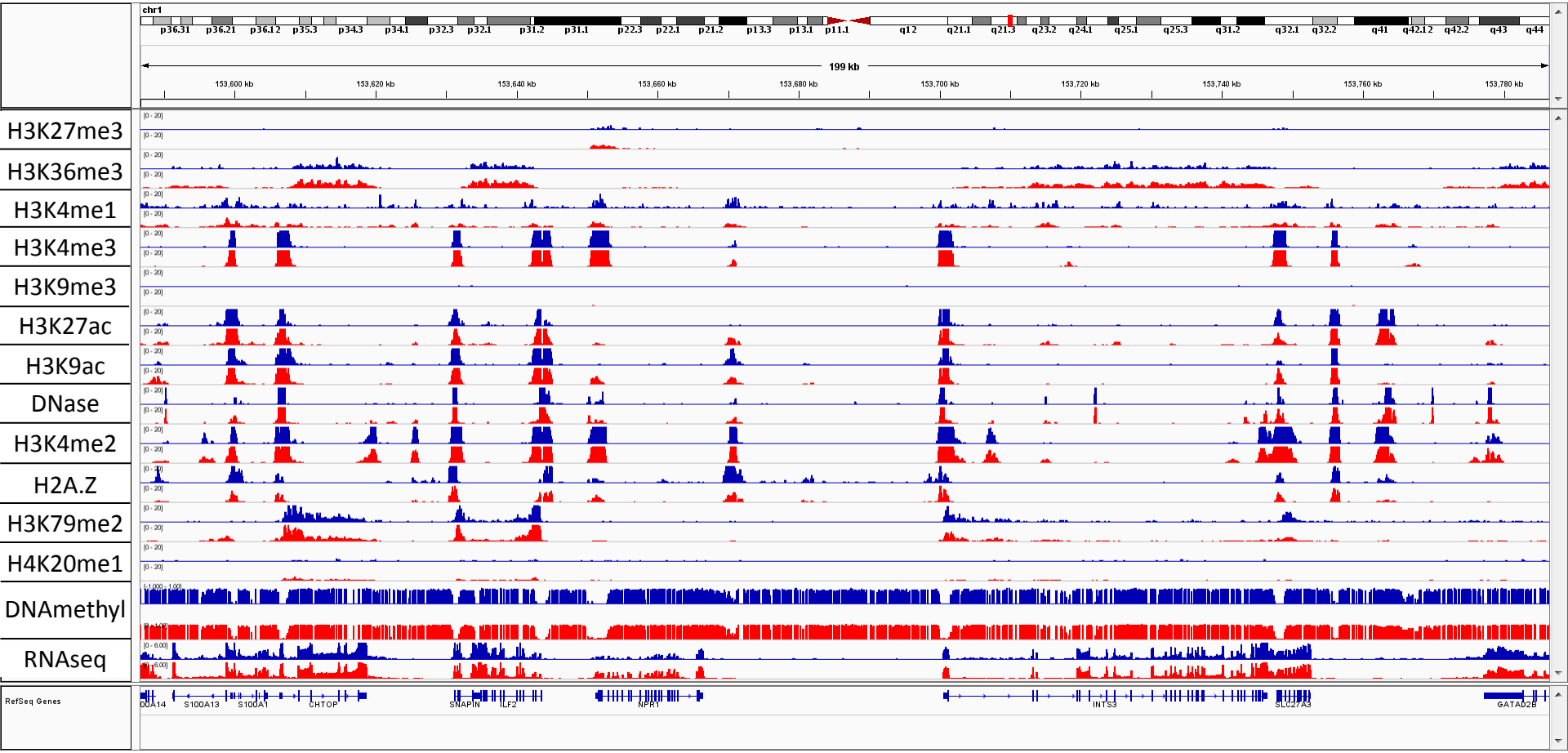
Blue observed; Red Imputed

Browser Screenshots of Random Loci



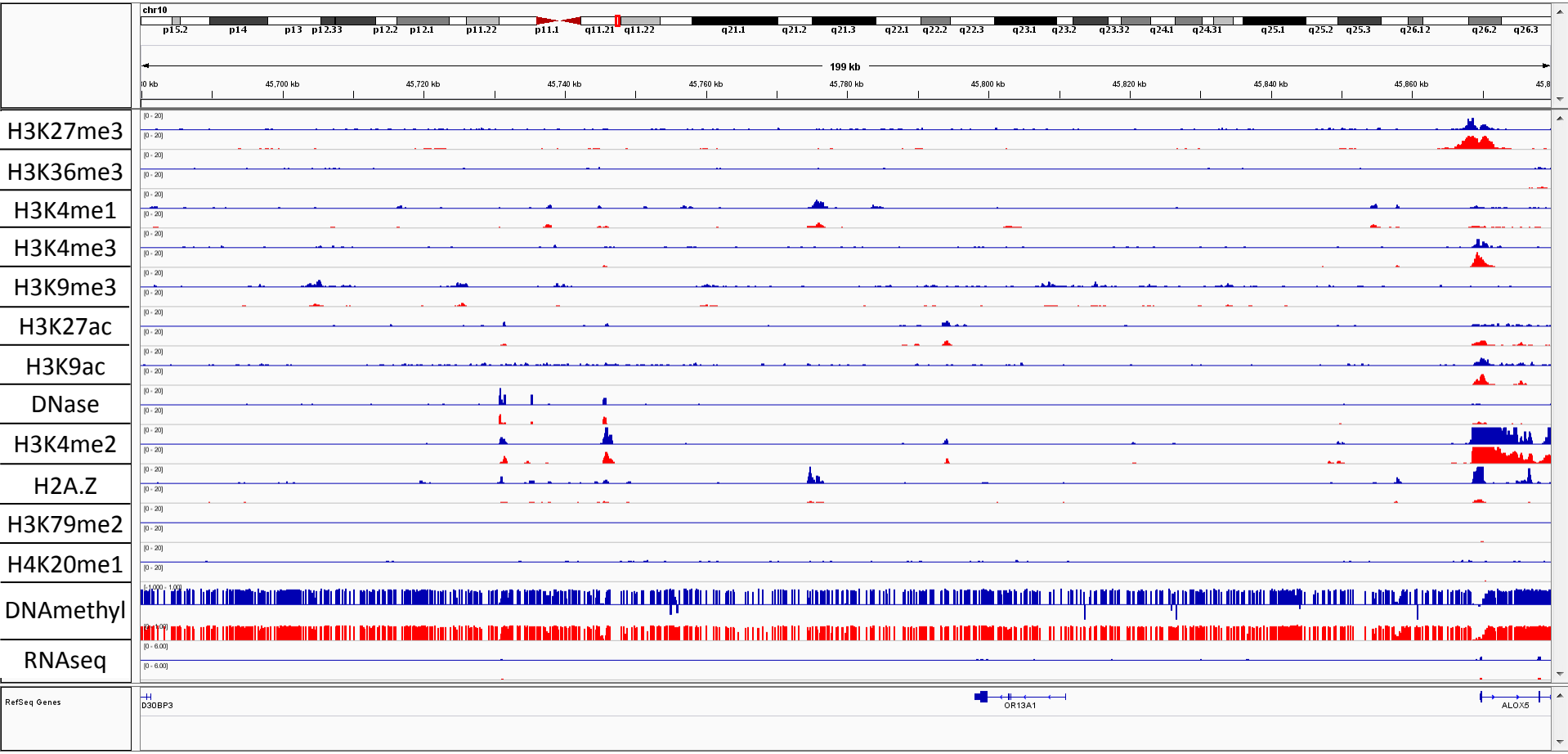
Blue observed; Red Imputed

Browser Screenshots of Random Loci



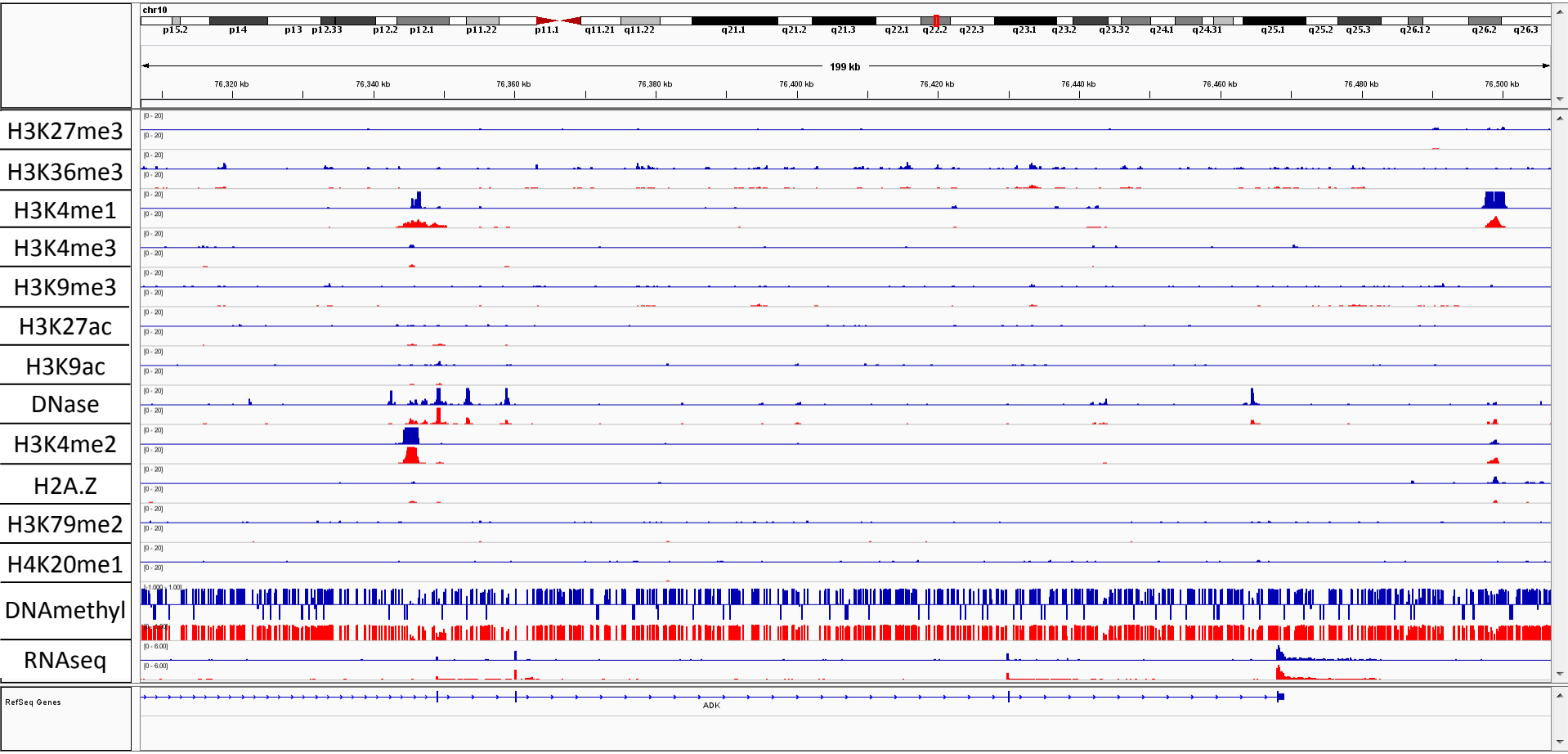
Blue observed; Red Imputed

Browser Screenshots of Random Loci



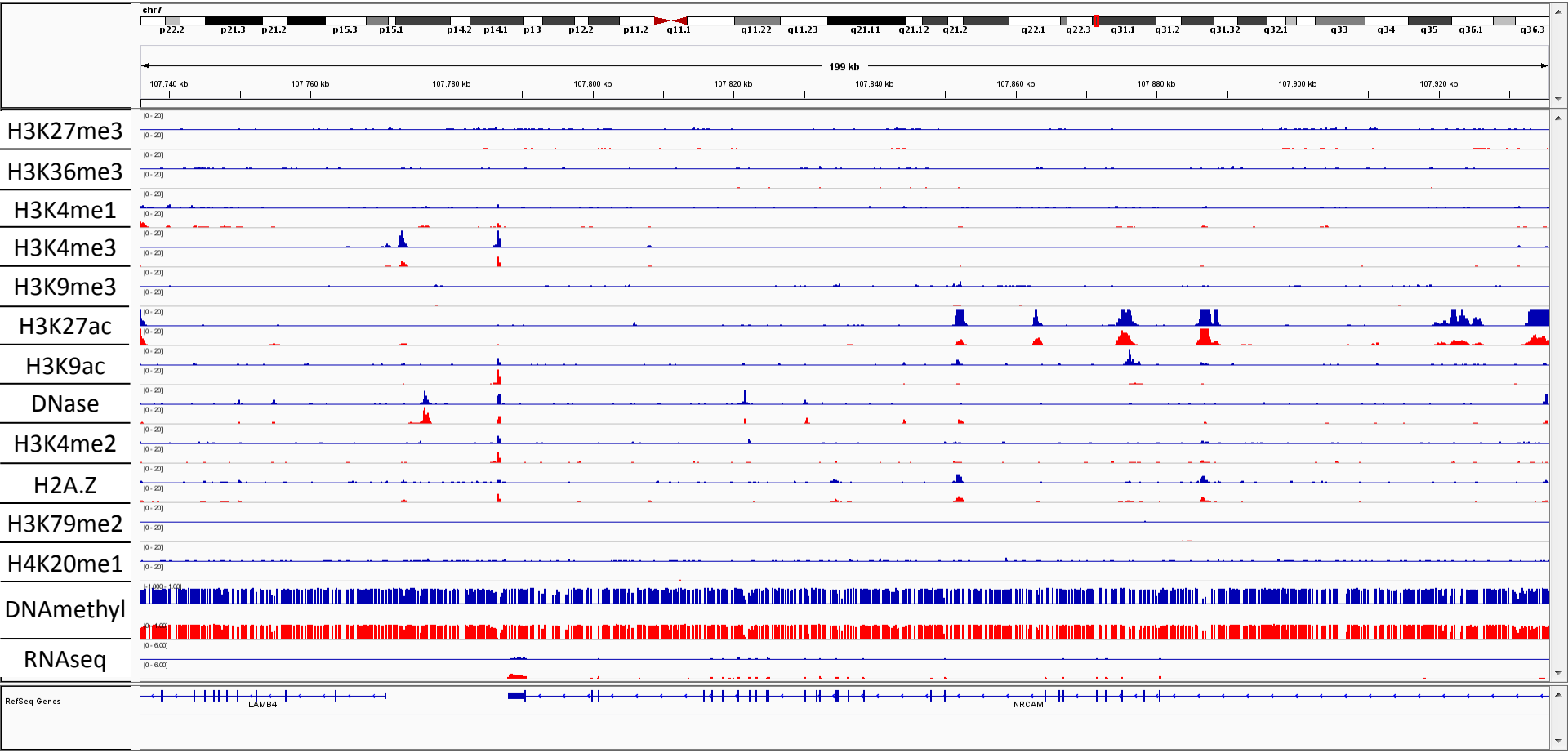
Blue observed; Red Imputed

Browser Screenshots of Random Loci



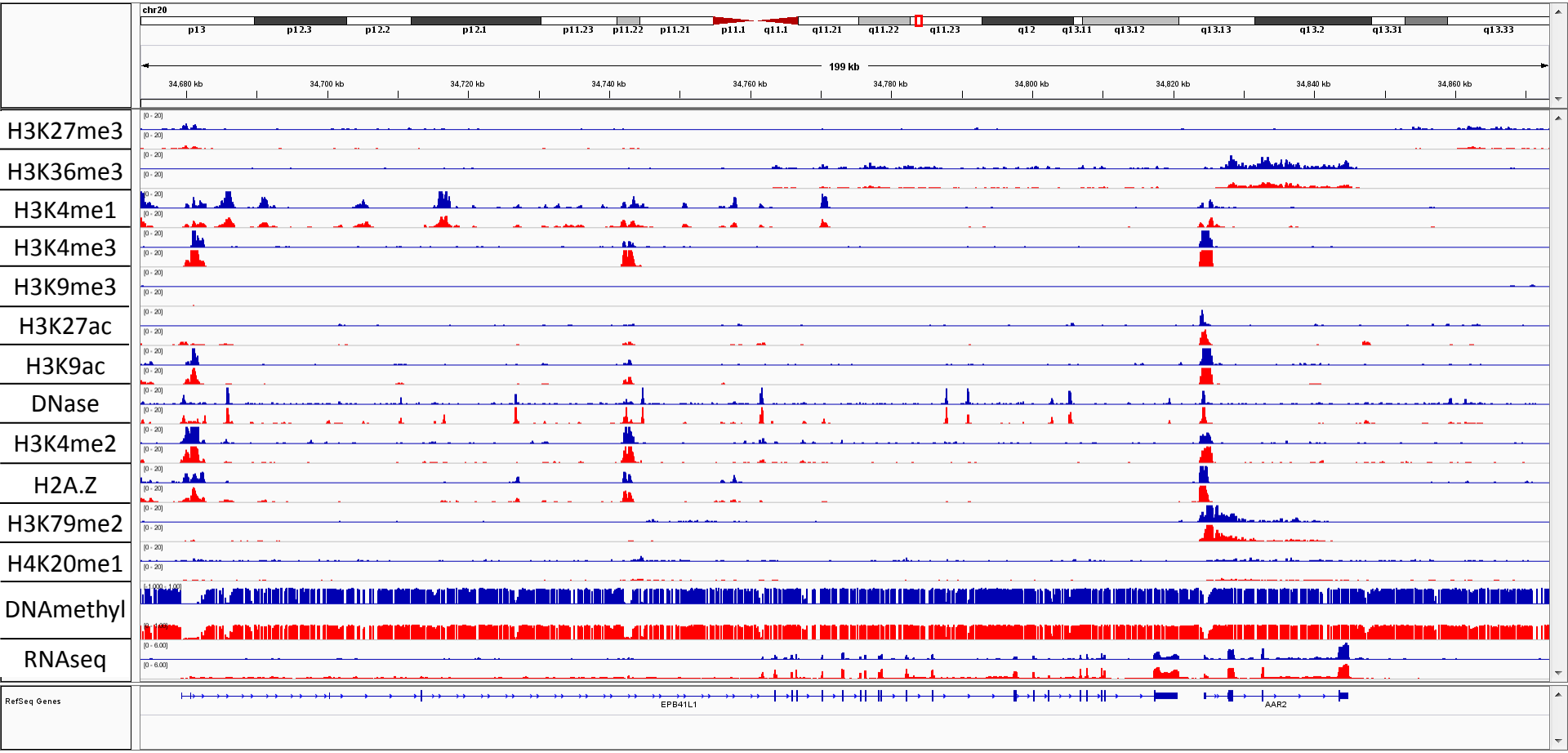
Blue observed; Red Imputed

Browser Screenshots of Random Loci



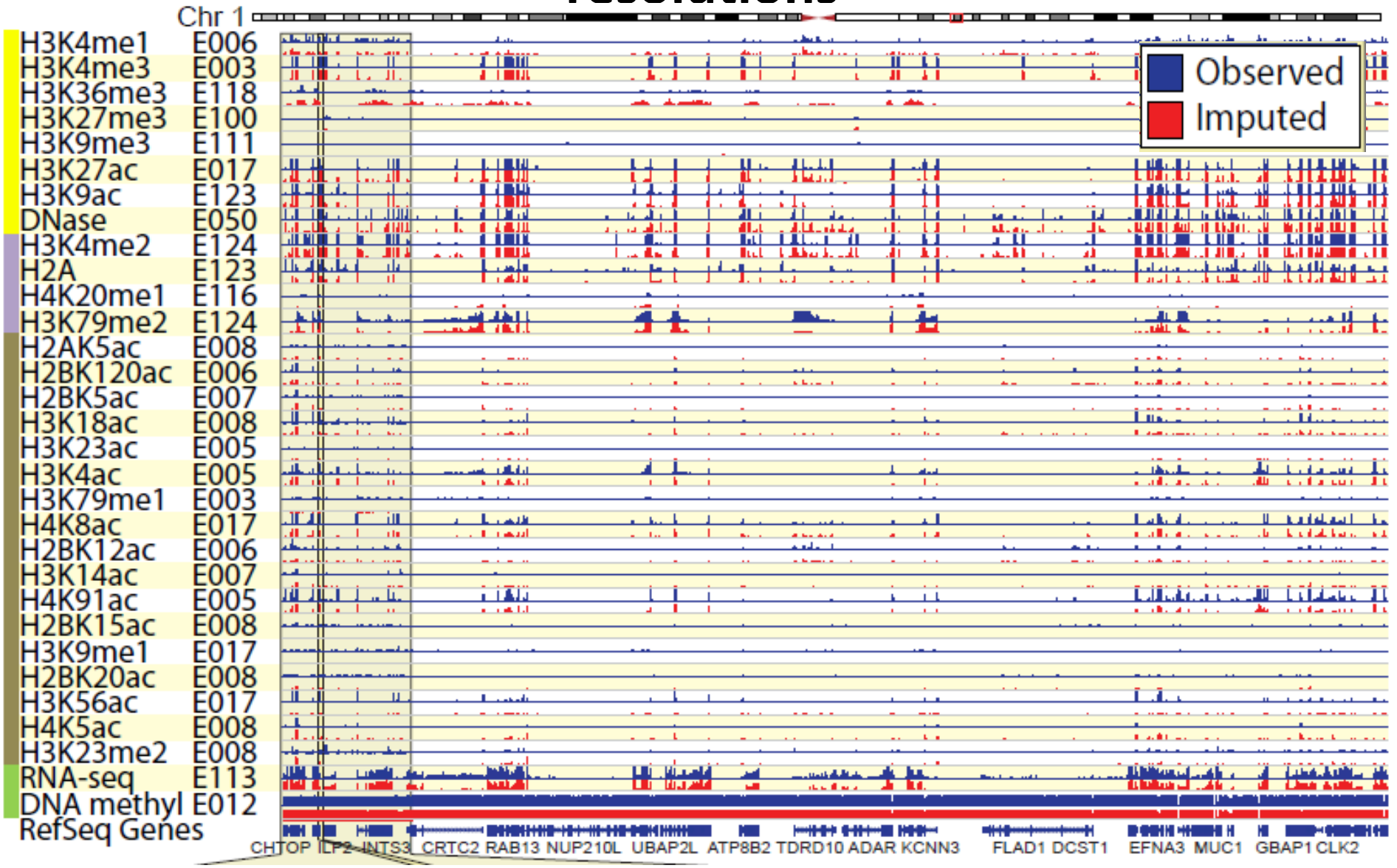
Blue observed; Red Imputed

Browser Screenshots of Random Loci



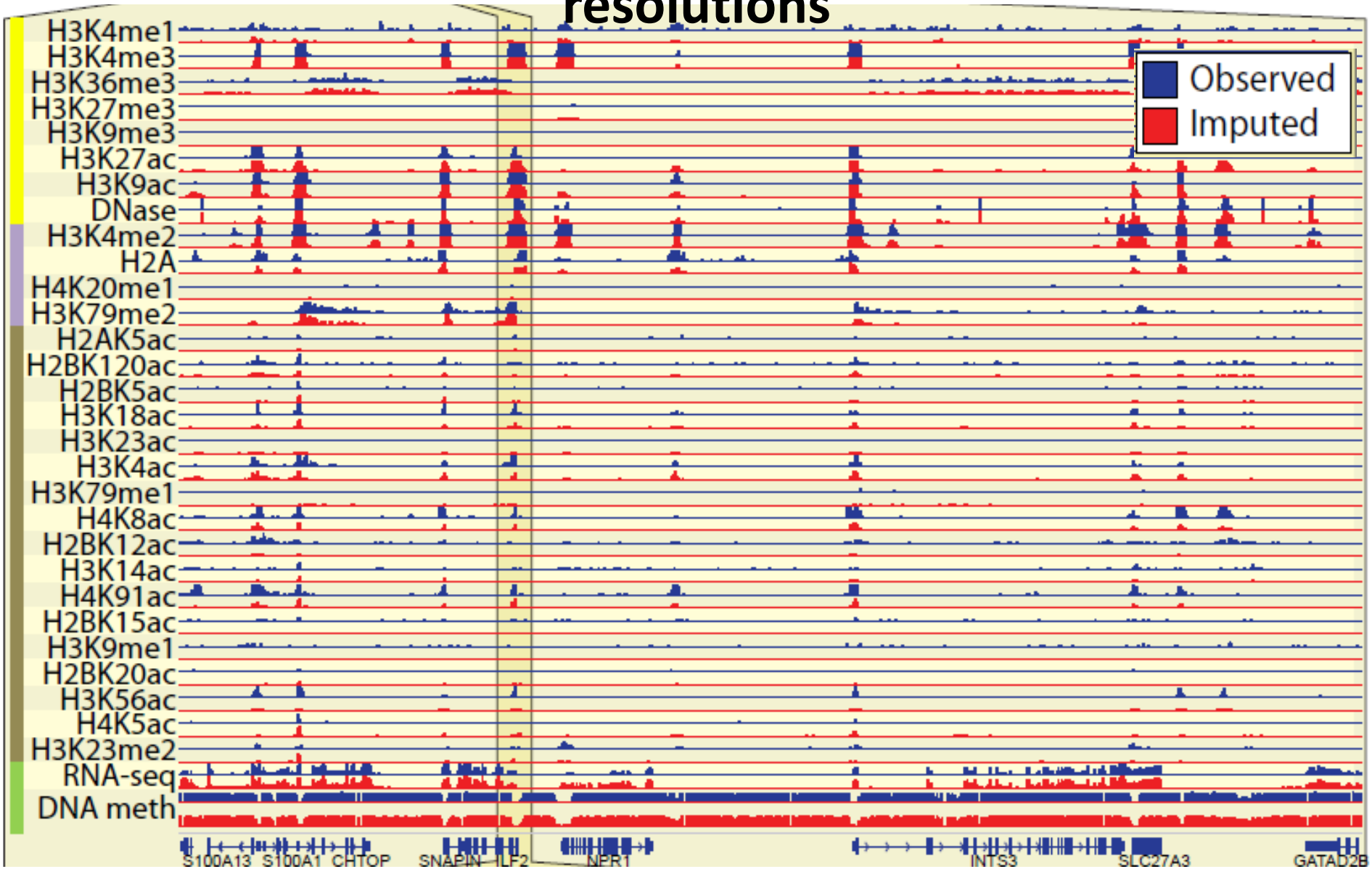
Blue observed; Red Imputed

Imputed data is a close match to observed at multiple resolutions



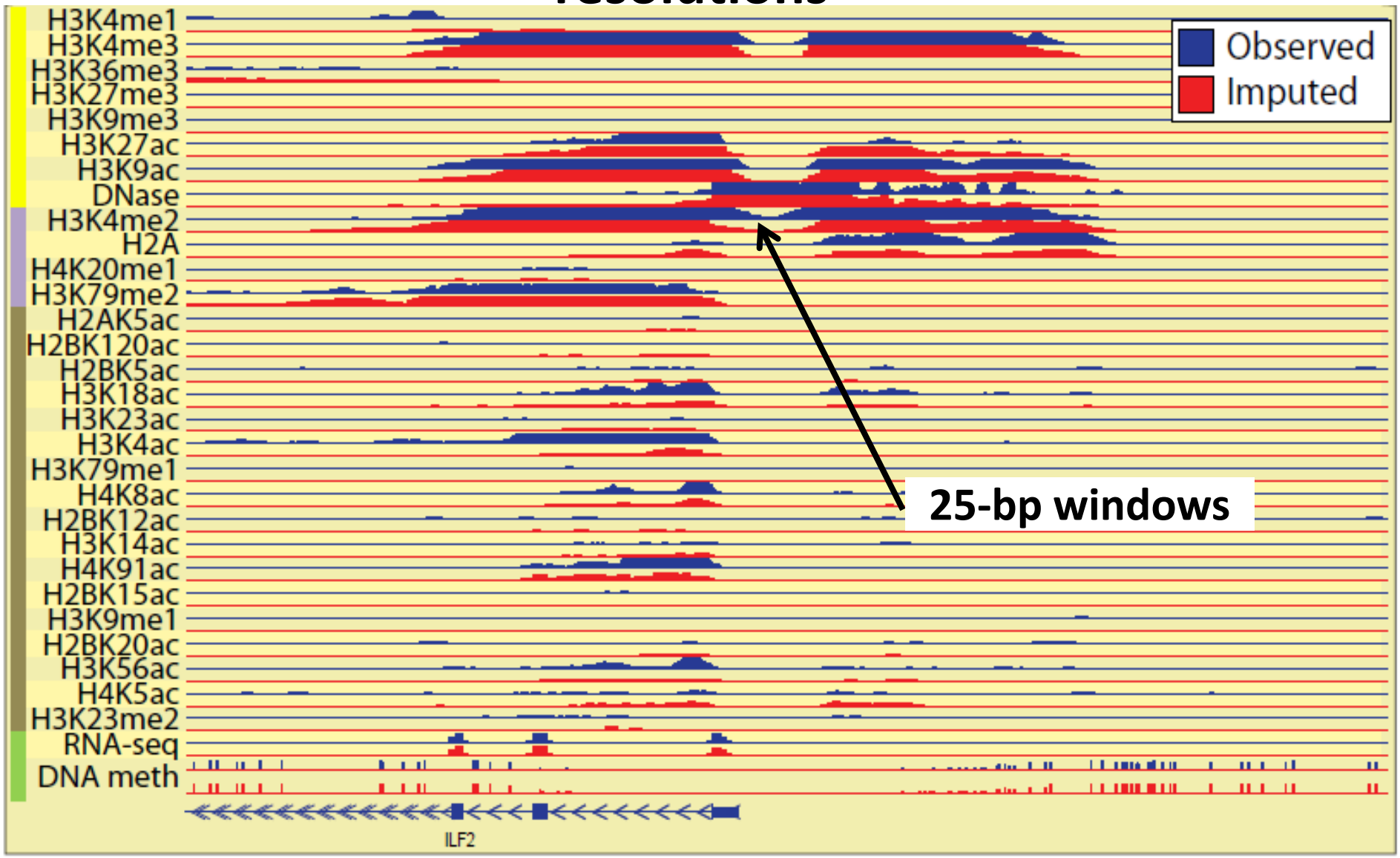
- 2Mb region, 1 tissue per mark

Imputed data is a close match to observed at multiple resolutions



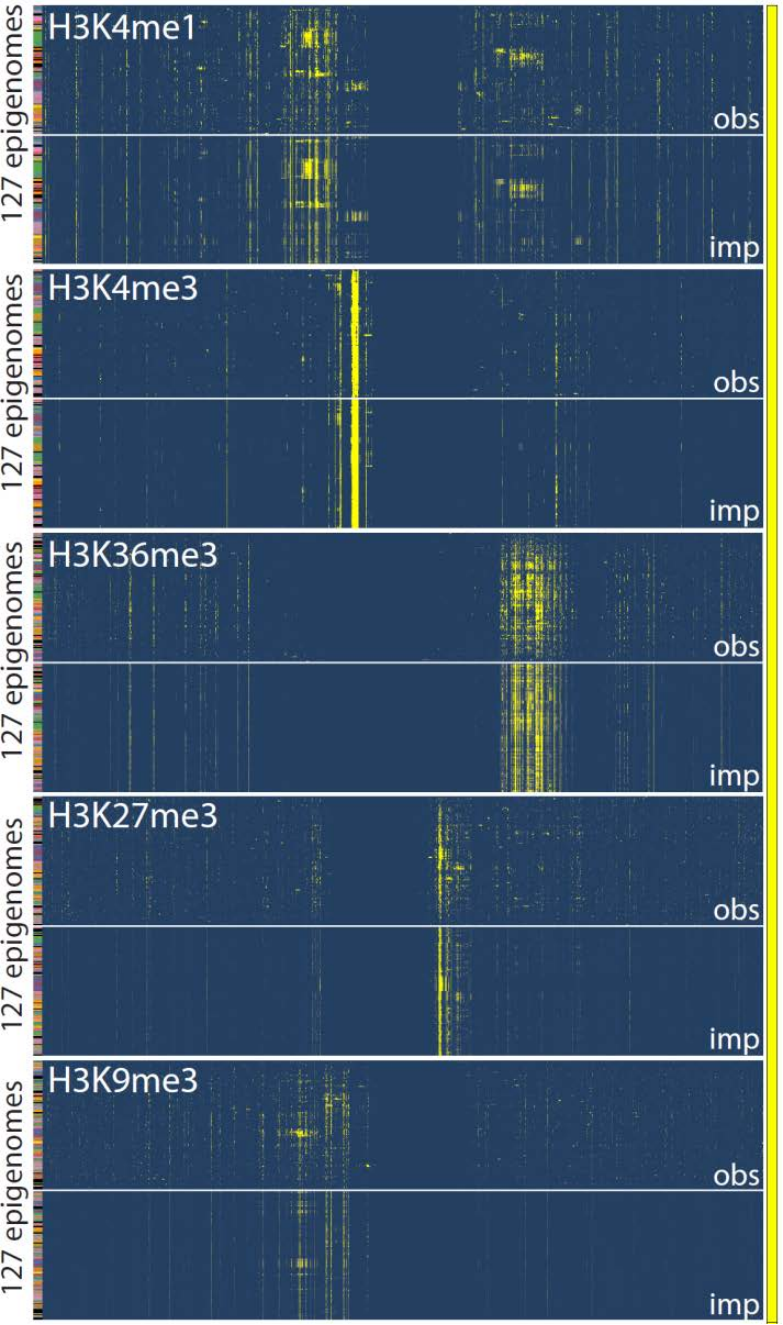
- 200kb region, 1 tissue per mark

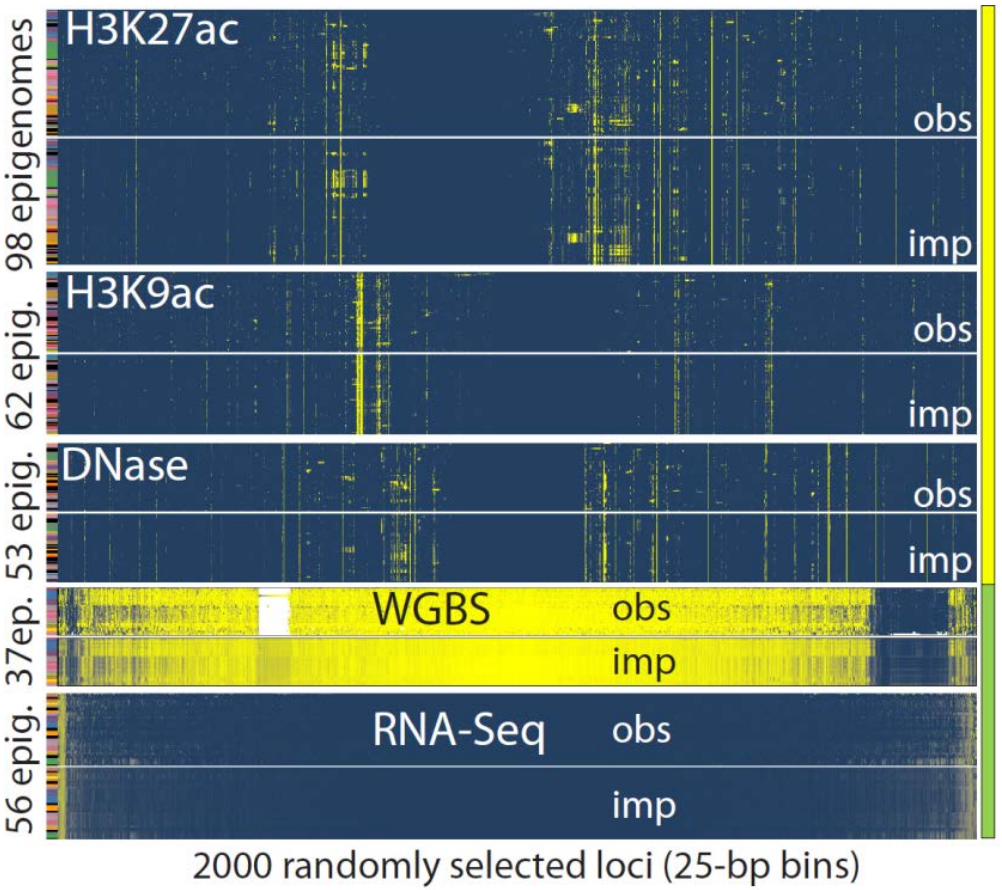
Imputed data is a close match to observed at multiple resolutions



- 10kb region, at 25bp bins

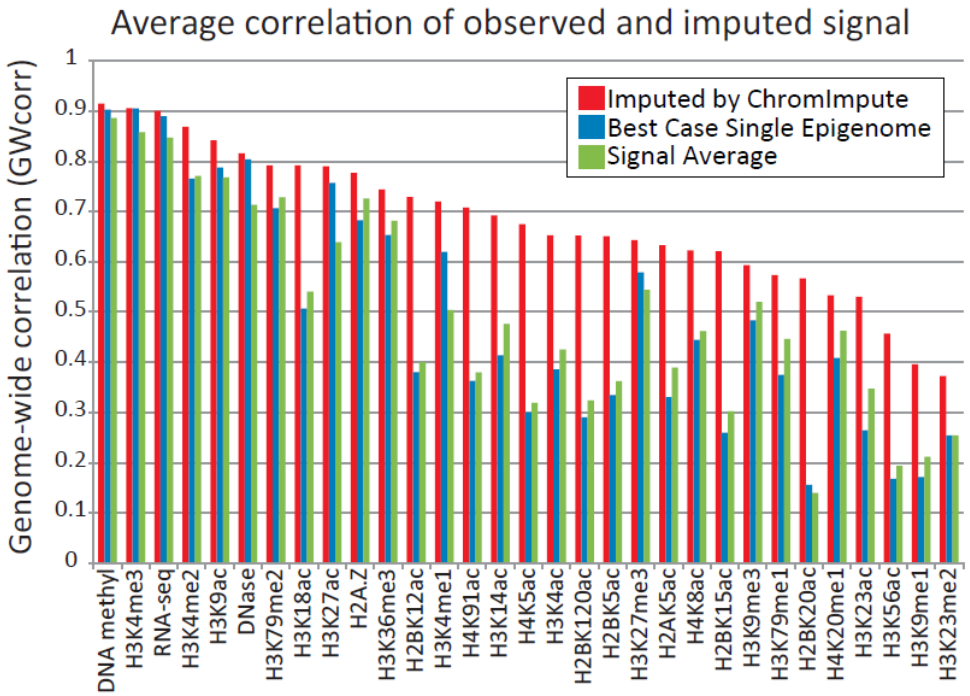
Observed/Imputed Data at 2000 Random Positions

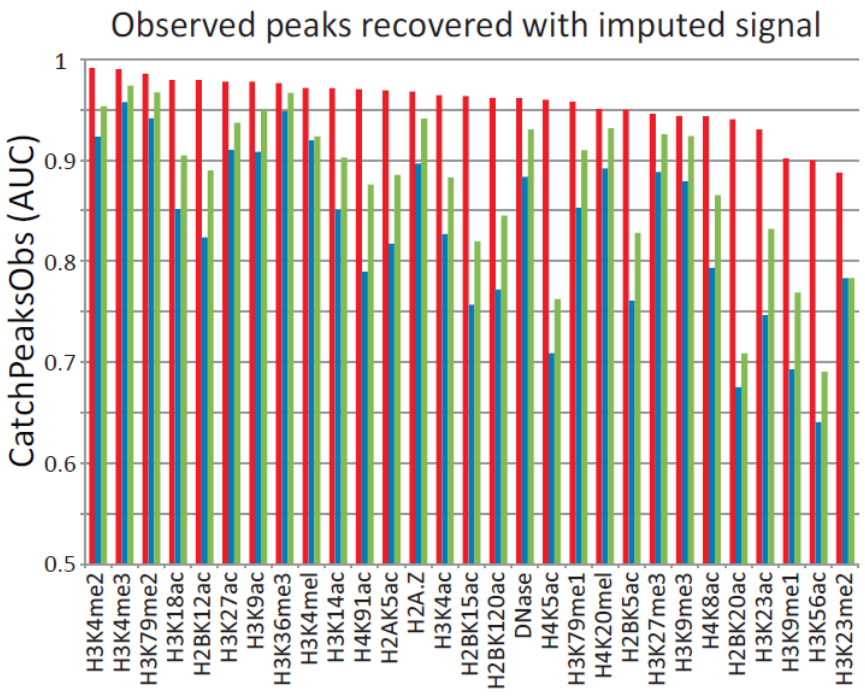




- Captures cell type specificity
- Dynamic changes across marks

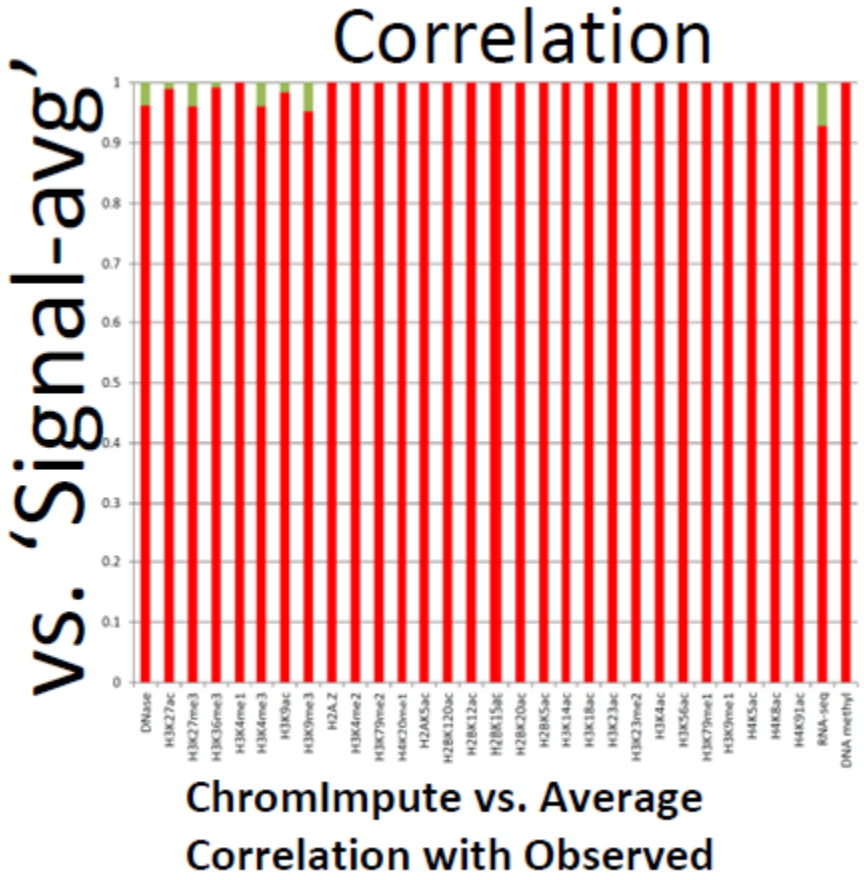
ChromImpute Outperforms Two Stringent Baselines



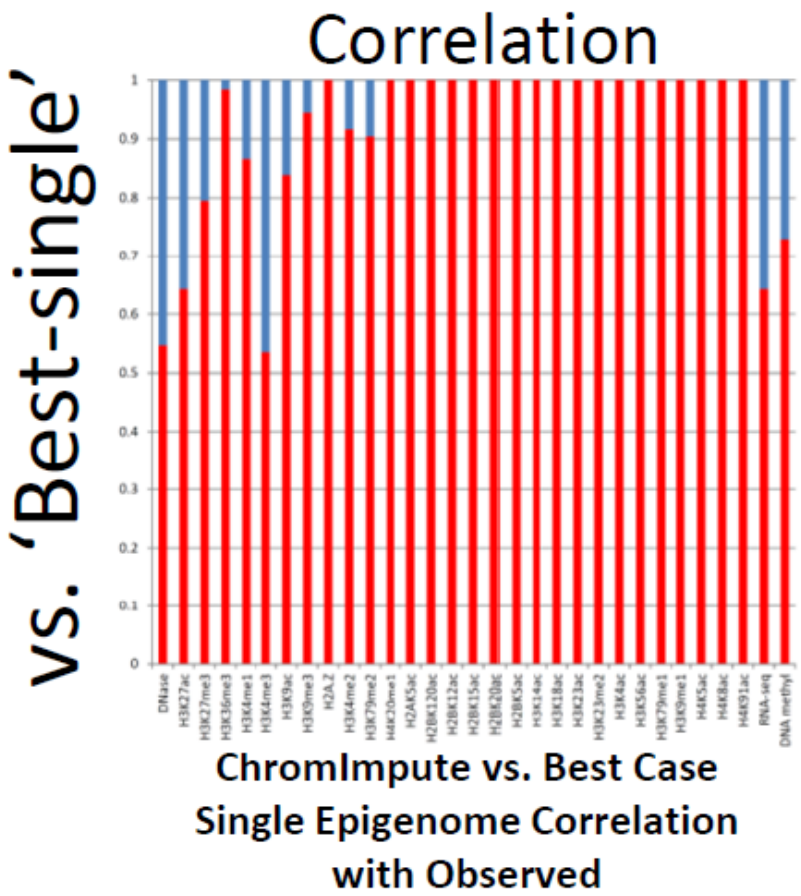


- Signal Average – average of mark across all other epigenomes
- Best Case Single Epigenome – upper bound on performance when selecting one epigenome

ChromImpute Outperforms Baselines on Vast Majority of Individual Data Sets

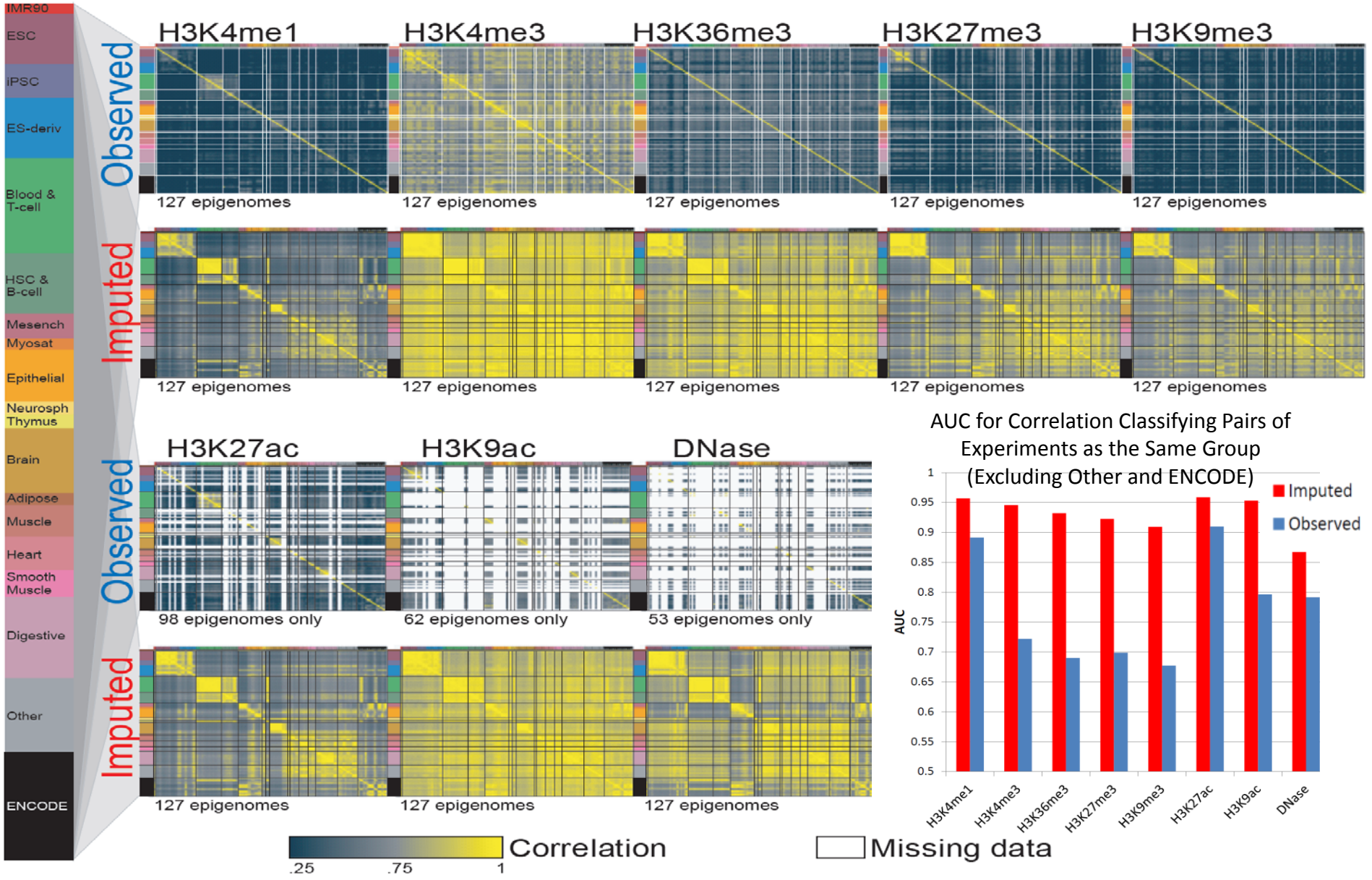


ChromImpute vs. Average Correlation with Observed



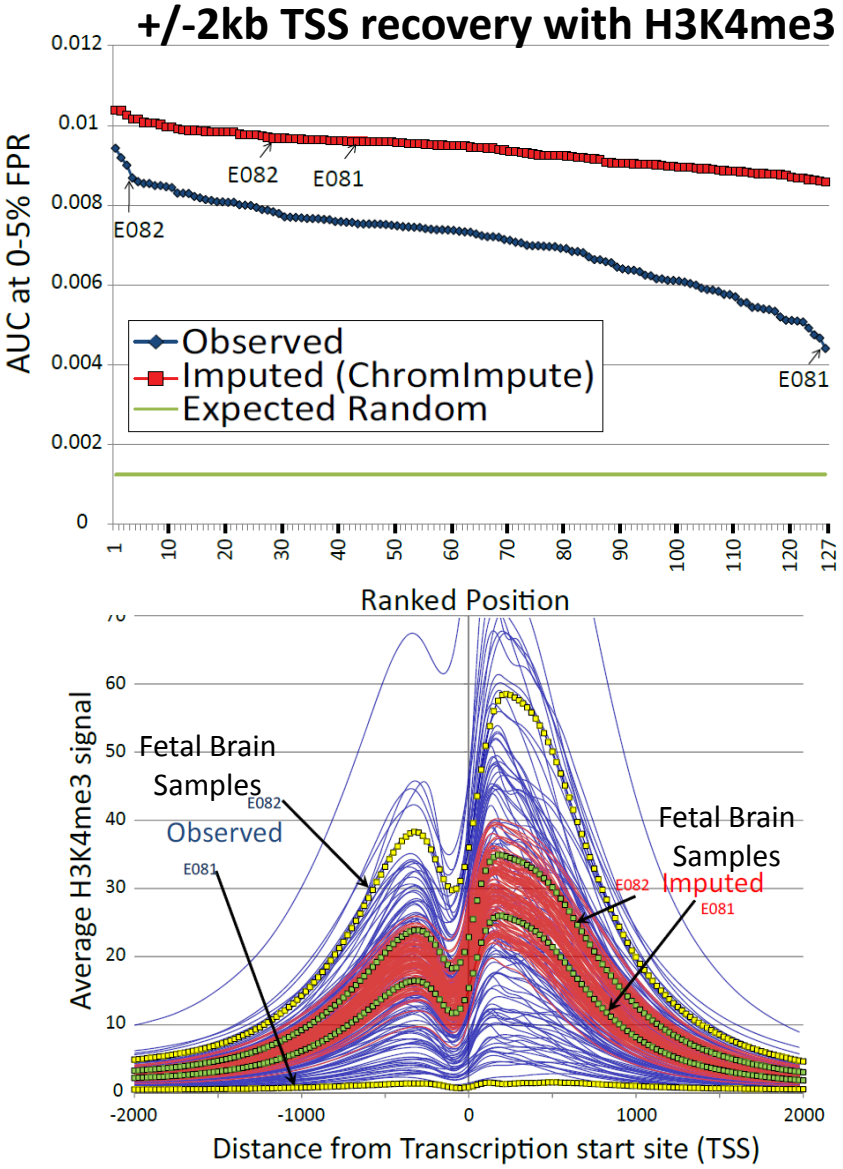
ChromImpute vs. Best Case Single Epigenome Correlation with Observed

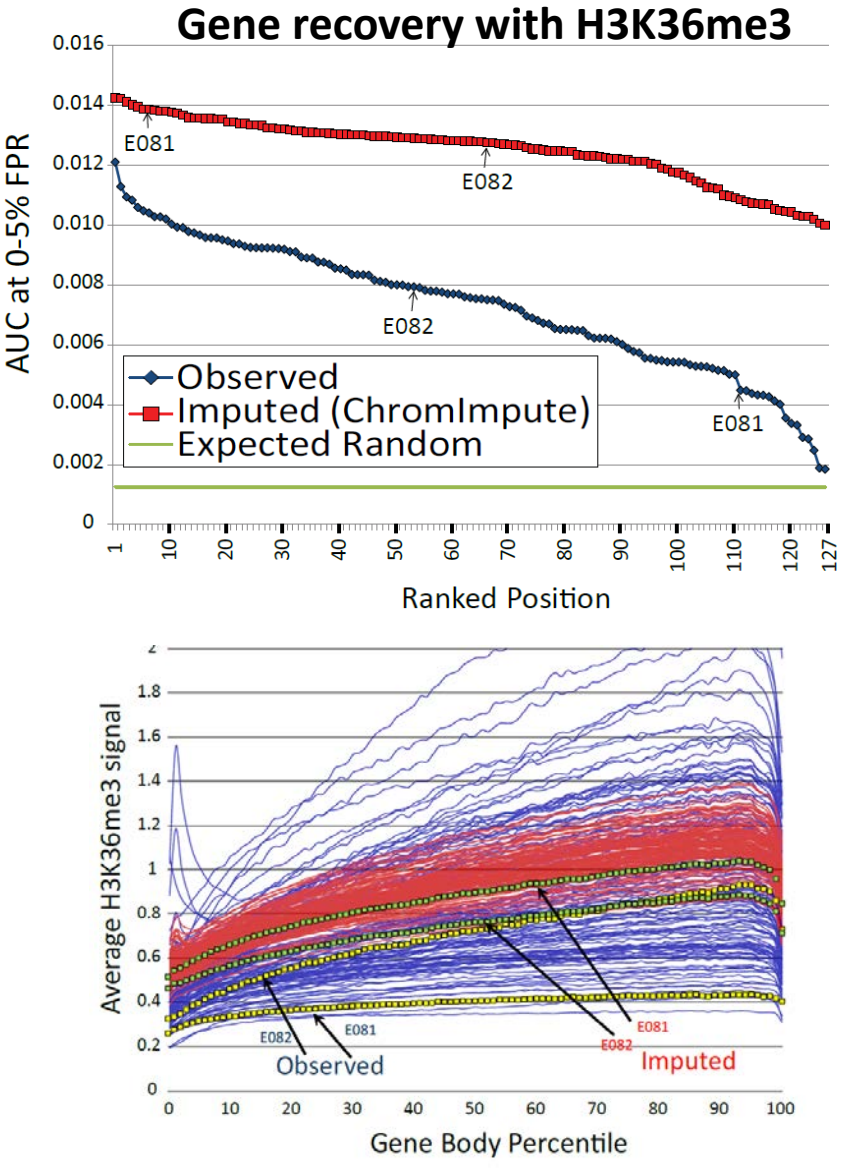
Imputed data capture tissue specificity/relationships



- **Better tissue coherence than observed datasets!**

Imputed: Better agreement with TSS and gene annotations





- **Unbiased comparison of observed/imputed data**

Observed/imputed discrepancy ➔ Flag low-quality data

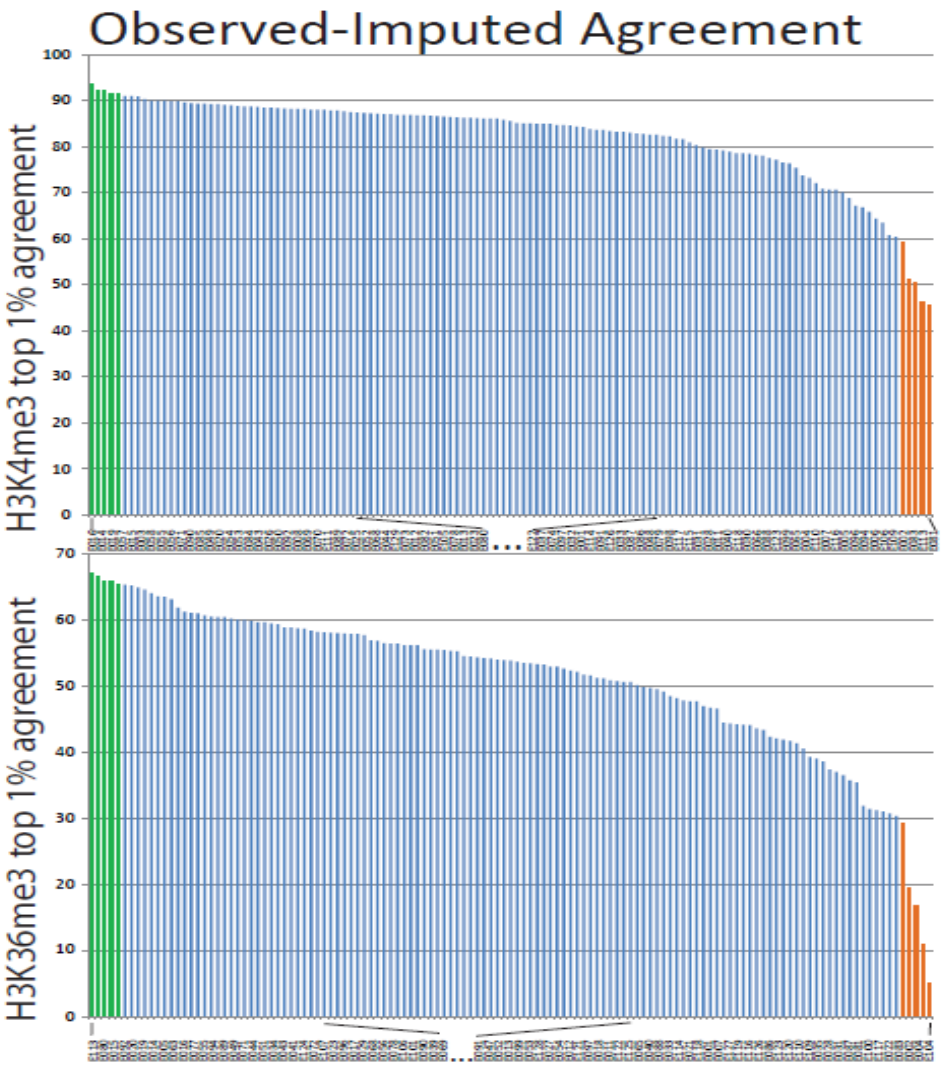
H3K4me3 lowest-AUC 5% datasets

Sample	AUC 5%	Read Depth	Poisson	SPOT	FindPeaks	NSC	RSC	Impute Top 1%	Impute correlation
E081	127	68	125	126	126	126	125	**127**	**127**
E113	126	111	86	121	124	115	102	**126**	121
E083	125	118	102	120	122	124	**127**	125	125
E002	124	**127**	42	59	78	82	1	124	126
E004	123	28	**122**	**122**	120	116	91	114	106
E106	122	115	106	106	114	106	96	**122**	**122**
E065	121	89	**117**	113	108	112	103	113	115
E109	120	122	103	102	115	109	122	**123**	**123**
E096	119	86	99	111	112	95	34	**119**	93
E007	118	28	104	112	110	90	58	**116**	94
Median		100	104	113	115	111	99	**123**	122

H3K36me3 lowest-AUC 5% datasets

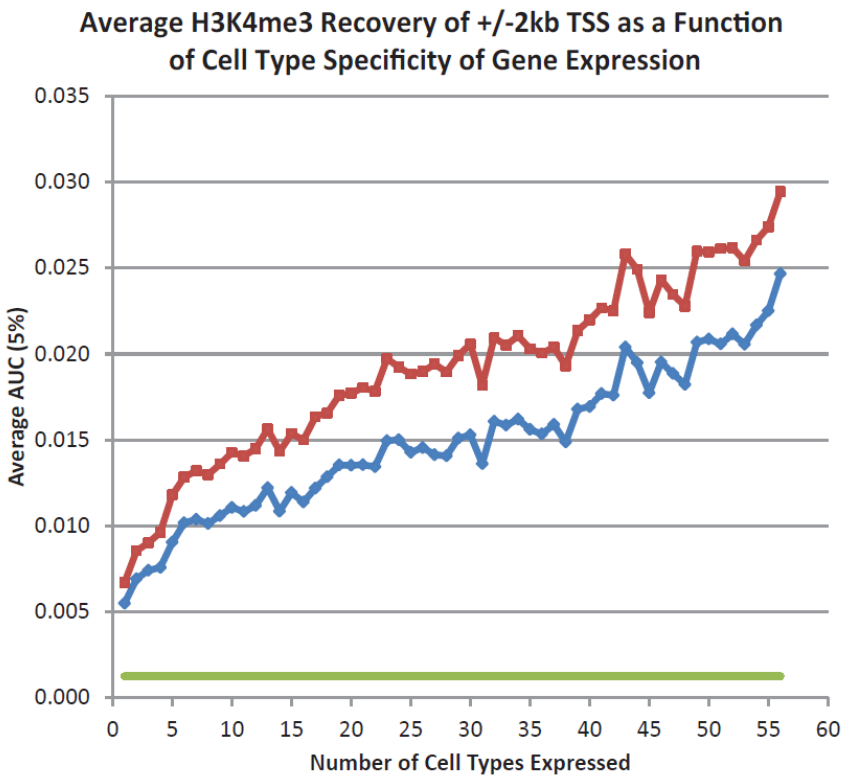
E104	127	**127**	50	9	49	1	70	**127**	**127**
E004	126	37	**127**	**127**	**127**	88	17	126	126
E002	125	124	122	100	109	38	3	**125**	**125**
E022	124	116	78	79	92	22	32	**123**	122
E087	123	**125**	95	49	82	10	61	119	123
E021	122	37	112	110	108	46	42	103	**115**
E083	121	115	120	114	122	85	118	**124**	**124**
E007	120	37	**123**	113	115	47	74	106	114
E109	119	**126**	70	38	68	26	103	115	**120**
E100	118	113	108	94	105	52	14	**121**	**121**
Median		116	110	97	107	42	52	122	**123**

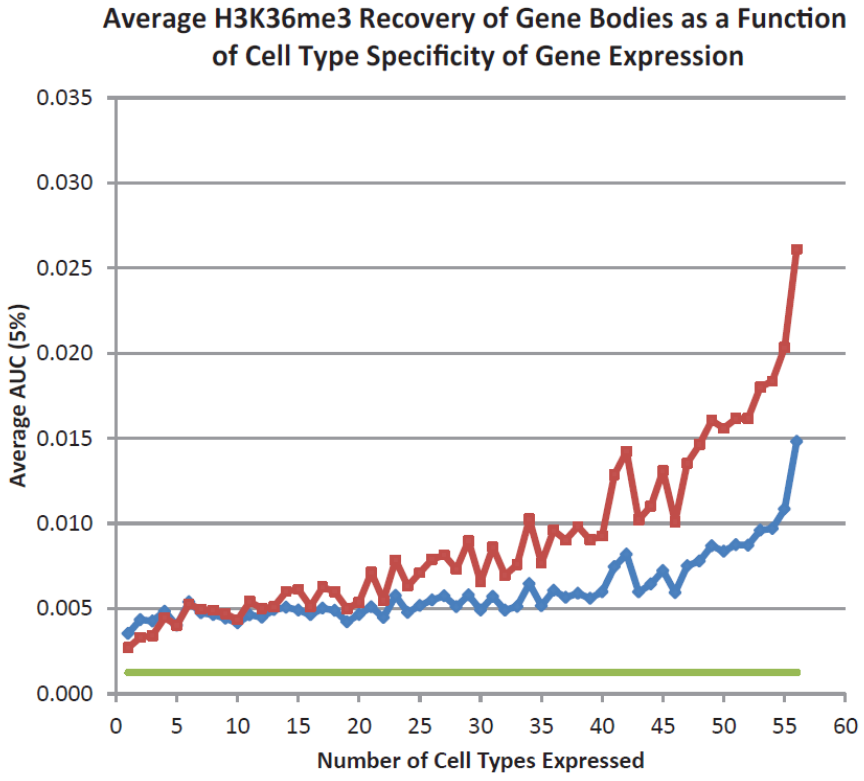
Dataset Rank for each QC metric

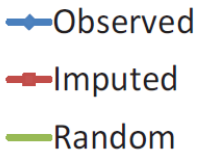


- **Existing QC metrics can fail for wrong Ab, cross-reactivity, label-swap**

Predictive Performances Increases for More Broadly Expressed Genes

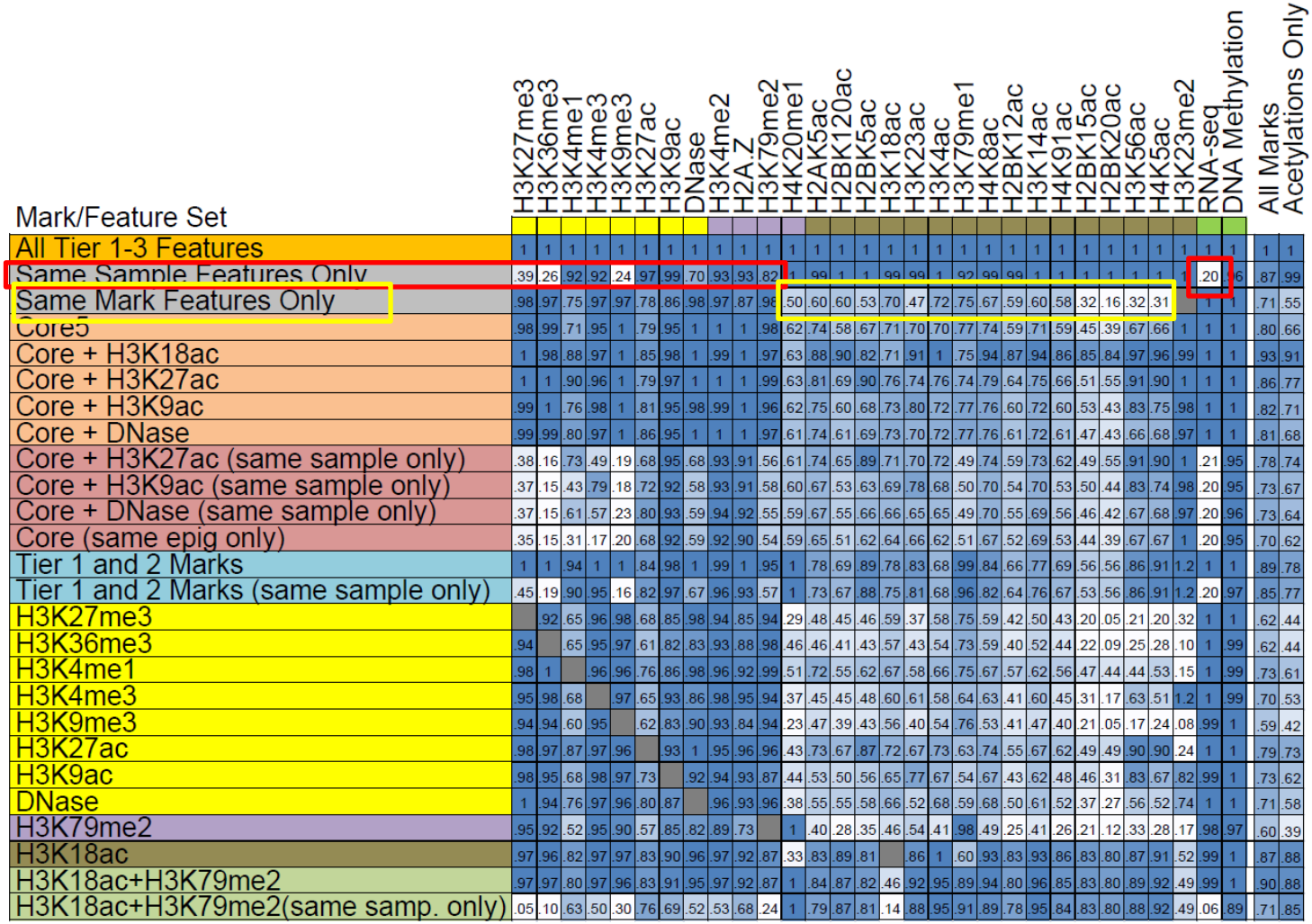






Expressed level RPKM>= 0.5

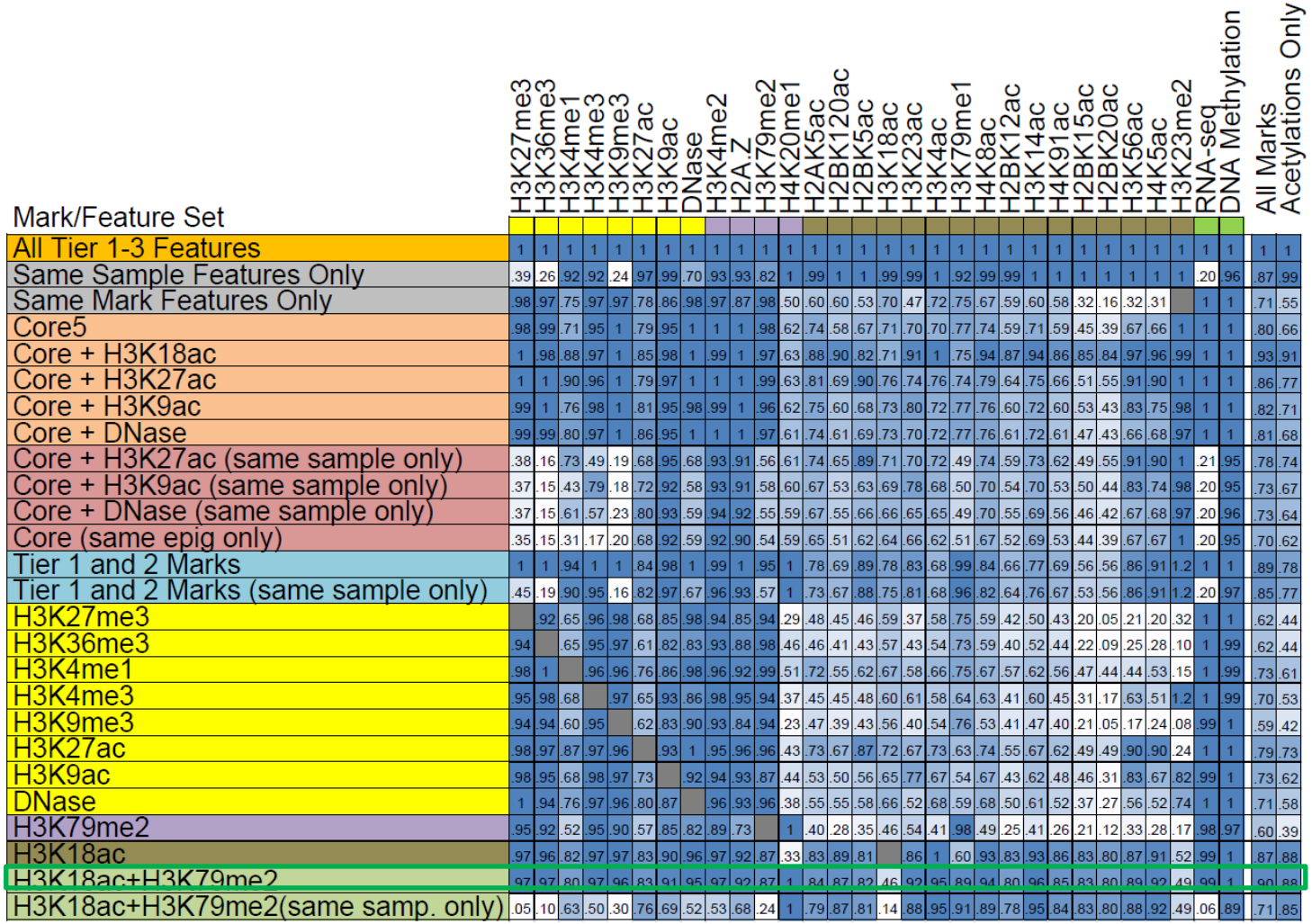
Mark prioritization from imputation performance



Mark/Feature Set	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3	H3K27ac	H3K9ac	DNase	H3K4me2	H2A.Z	H3K79me2	H4K20me1	H2AK5ac	H2BK120ac	H2BK5ac	H3K18ac	H3K23ac	H3K4ac	H3K79me1	H4K8ac	H2BK12ac	H3K14ac	H4K91ac	H2BK15ac	H2BK20ac	H3K56ac	H4K5ac	H3K23me2	RNA-seq	DNA Methylation	All Marks	Acetylations Only
All Tier 1-3 Features	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Same Sample Features Only	.39	.26	.92	.92	.24	.97	.99	.70	.93	.93	.82	1	.99	1	1	.99	.99	1	.92	.99	.99	1	1	1	1	1	1	1	.20	.96	.87	.99
Same Mark Features Only	.98	.97	.75	.97	.97	.78	.86	.98	.97	.87	.98	.50	.60	.60	.53	.70	.47	.72	.75	.67	.59	.60	.58	.32	.16	.32	.31	1	1	1	.71	.55
Core5	.98	.99	.71	.95	1	.79	.95	1	1	1	.98	.62	.74	.58	.67	.71	.70	.70	.77	.74	.59	.71	.59	.45	.39	.67	.66	1	1	1	.80	.66
Core + H3K18ac	1	.98	.88	.97	1	.85	.98	1	.99	1	.97	.63	.88	.90	.82	.71	.91	1	.75	.94	.87	.94	.86	.85	.84	.97	.96	.99	1	1	.93	.91
Core + H3K27ac	1	1	.90	.96	1	.79	.97	1	1	1	.99	.63	.81	.69	.90	.76	.74	.76	.74	.79	.64	.75	.66	.51	.55	.91	.90	1	1	1	.86	.77
Core + H3K9ac	.99	1	.76	.98	1	.81	.95	.98	.99	1	.96	.62	.75	.60	.68	.73	.80	.72	.77	.76	.60	.72	.60	.53	.43	.83	.75	.98	1	1	.82	.71
Core + DNase	.99	.99	.80	.97	1	.86	.95	1	1	1	.97	.61	.74	.61	.69	.73	.70	.72	.77	.76	.61	.72	.61	.47	.43	.66	.68	.97	1	1	.81	.68
Core + H3K27ac (same sample only)	.38	.16	.73	.49	.19	.68	.95	.68	.93	.91	.56	.61	.74	.65	.89	.71	.70	.72	.49	.74	.59	.73	.62	.49	.55	.91	.90	1	.21	.95	.78	.74
Core + H3K9ac (same sample only)	.37	.15	.43	.79	.18	.72	.92	.58	.93	.91	.58	.60	.67	.53	.63	.69	.78	.68	.50	.70	.54	.70	.53	.50	.44	.83	.74	.98	.20	.95	.73	.67
Core + DNase (same sample only)	.37	.15	.61	.57	.23	.80	.93	.59	.94	.92	.55	.59	.67	.55	.66	.66	.65	.65	.49	.70	.55	.69	.56	.46	.42	.67	.68	.97	.20	.96	.73	.64
Core (same epig only)	.35	.15	.31	.17	.20	.68	.92	.59	.92	.90	.54	.59	.65	.51	.62	.64	.66	.62	.51	.67	.52	.69	.53	.44	.39	.67	.67	1	.20	.95	.70	.62
Tier 1 and 2 Marks	1	1	.94	1	1	.84	.98	1	.99	1	.95	1	.78	.69	.89	.78	.83	.68	.99	.84	.66	.77	.69	.56	.56	.86	.91	1.2	1	1	.89	.78
Tier 1 and 2 Marks (same sample only)	.45	.19	.90	.95	.16	.82	.97	.67	.96	.93	.57	1	.73	.67	.88	.75	.81	.68	.96	.82	.64	.76	.67	.53	.56	.86	.91	1.2	.20	.97	.85	.77
H3K27me3		.92	.65	.96	.98	.68	.85	.98	.94	.85	.94	.29	.48	.45	.46	.59	.37	.58	.75	.59	.42	.50	.43	.20	.05	.21	.20	.32	1	1	.62	.44
H3K36me3	.94		.65	.95	.97	.61	.82	.83	.93	.88	.98	.46	.46	.41	.43	.57	.43	.54	.73	.59	.40	.52	.44	.22	.09	.25	.28	.10	1	.99	.62	.44
H3K4me1	.98	1		.96	.96	.76	.86	.98	.96	.92	.99	.51	.72	.55	.62	.67	.58	.66	.75	.67	.57	.62	.56	.47	.44	.44	.53	.15	1	.99	.73	.61
H3K4me3	.95	.98	.68		.97	.65	.93	.86	.98	.95	.94	.37	.45	.45	.48	.60	.61	.58	.64	.63	.41	.60	.45	.31	.17	.63	.51	1.2	1	.99	.70	.53
H3K9me3	.94	.94	.60	.95		.62	.83	.90	.93	.84	.94	.23	.47	.39	.43	.56	.40	.54	.76	.53	.41	.47	.40	.21	.05	.17	.24	.08	.99	1	.59	.42
H3K27ac	.98	.97	.87	.97	.96		.93	1	.95	.96	.96	.43	.73	.67	.87	.72	.67	.73	.63	.74	.55	.67	.62	.49	.49	.90	.90	.24	1	1	.79	.73
H3K9ac	.98	.95	.68	.98	.97	.73		.92	.94	.93	.87	.44	.53	.50	.56	.65	.77	.67	.54	.67	.43	.62	.48	.46	.31	.83	.67	.82	.99	1	.73	.62
DNase	1	.94	.76	.97	.96	.80	.87		.96	.93	.96	.38	.55	.55	.58	.66	.52	.68	.59	.68	.50	.61	.52	.37	.27	.56	.52	.74	1	1	.71	.58
H3K79me2	.95	.92	.52	.95	.90	.57	.85	.82	.89	.73		1	.40	.28	.35	.46	.54	.41	.98	.49	.25	.41	.26	.21	.12	.33	.28	.17	.98	.97	.60	.39
H3K18ac	.97	.96	.82	.97	.97	.83	.90	.96	.97	.92	.87	.33	.83	.89	.81		.86	1	.60	.93	.83	.93	.86	.83	.80	.87	.91	.52	.99	1	.87	.88
H3K18ac+H3K79me2	.97	.97	.80	.97	.96	.83	.91	.95	.97	.92	.87	1	.84	.87	.82	.46	.92	.95	.89	.94	.80	.96	.85	.83	.80	.89	.92	.49	.99	1	.90	.88
H3K18ac+H3K79me2(same samp. only)	.05	.10	.63	.50	.30	.76	.69	.52	.53	.68	.24	1	.79	.87	.81	.14	.88	.95	.91	.89	.78	.95	.84	.83	.80	.88	.92	.49	.06	.89	.71	.85

Evaluation of performance for subset of marks/features relative to prediction with all features on deep epigenomes

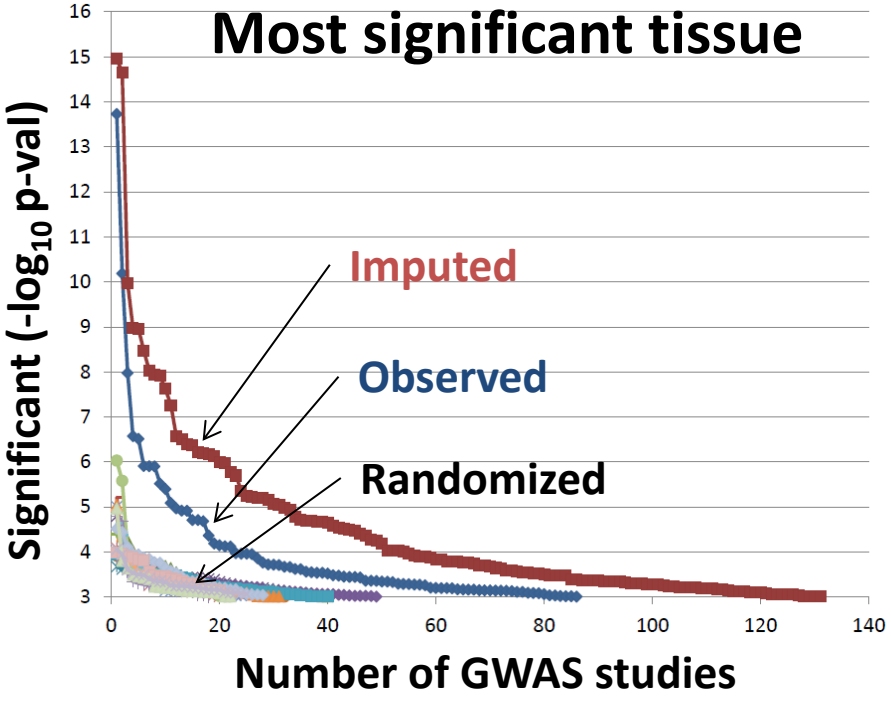
Mark prioritization from imputation performance

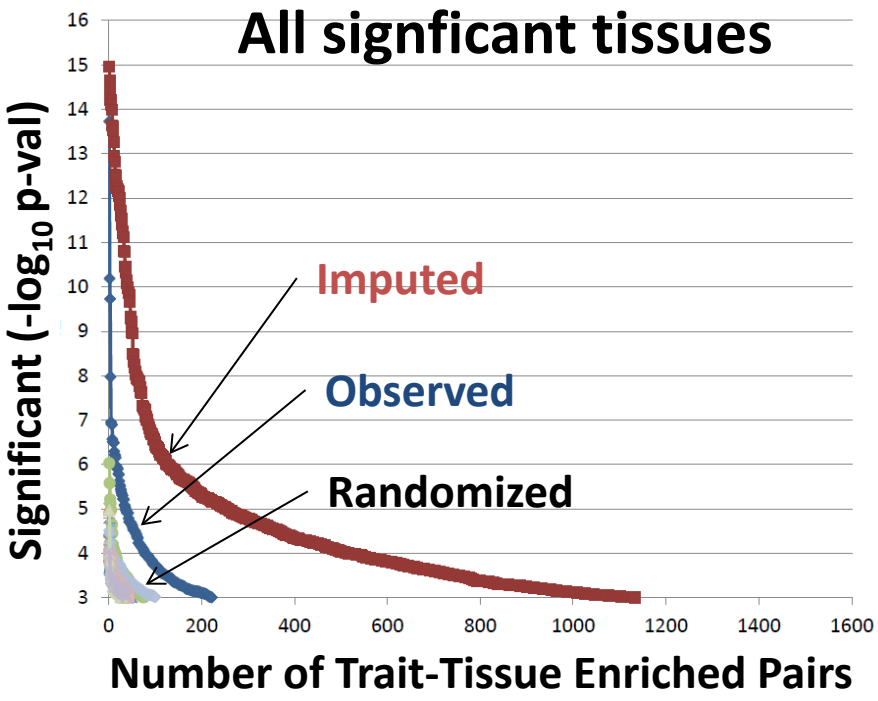


Mark/Feature Set	H3K27me3	H3K36me3	H3K4me1	H3K4me3	H3K9me3	H3K27ac	H3K9ac	DNase	H3K4me2	H2A.Z	H3K79me2	H4K20me1	H2AK5ac	H2BK120ac	H2BK5ac	H3K18ac	H3K23ac	H3K4ac	H3K79me1	H4K8ac	H2BK12ac	H3K14ac	H4K91ac	H2BK15ac	H2BK20ac	H3K56ac	H4K5ac	H3K23me2	RNA-seq	DNA Methylation	All Marks	Acetylations Only
All Tier 1-3 Features	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Same Sample Features Only	.39	.26	.92	.92	.24	.97	.99	.70	.93	.93	.82	1	.99	1	1	.99	.99	1	.92	.99	.99	1	1	1	1	1	1	1	.20	.96	.87	.99
Same Mark Features Only	.98	.97	.75	.97	.97	.78	.86	.98	.97	.87	.98	.50	.60	.60	.53	.70	.47	.72	.75	.67	.59	.60	.58	.32	.16	.32	.31		1	1	.71	.55
Core5	.98	.99	.71	.95	1	.79	.95	1	1	1	.98	.62	.74	.58	.67	.71	.70	.70	.77	.74	.59	.71	.59	.45	.39	.67	.66	1	1	1	.80	.66
Core + H3K18ac	1	.98	.88	.97	1	.85	.98	1	.99	1	.97	.63	.88	.90	.82	.71	.91	1	.75	.94	.87	.94	.86	.85	.84	.97	.96	.99	1	1	.93	.91
Core + H3K27ac	1	1	.90	.96	1	.79	.97	1	1	1	.99	.63	.81	.69	.90	.76	.74	.76	.74	.79	.64	.75	.66	.51	.55	.91	.90	1	1	1	.86	.77
Core + H3K9ac	.99	1	.76	.98	1	.81	.95	.98	.99	1	.96	.62	.75	.60	.68	.73	.80	.72	.77	.76	.60	.72	.60	.53	.43	.83	.75	.98	1	1	.82	.71
Core + DNase	.99	.99	.80	.97	1	.86	.95	1	1	1	.97	.61	.74	.61	.69	.73	.70	.72	.77	.76	.61	.72	.61	.47	.43	.66	.68	.97	1	1	.81	.68
Core + H3K27ac (same sample only)	.38	.16	.73	.49	.19	.68	.95	.68	.93	.91	.56	.61	.74	.65	.89	.71	.70	.72	.49	.74	.59	.73	.62	.49	.55	.91	.90	1	.21	.95	.78	.74
Core + H3K9ac (same sample only)	.37	.15	.43	.79	.18	.72	.92	.58	.93	.91	.58	.60	.67	.53	.63	.69	.78	.68	.50	.70	.54	.70	.53	.50	.44	.83	.74	.98	.20	.95	.73	.67
Core + DNase (same sample only)	.37	.15	.61	.57	.23	.80	.93	.59	.94	.92	.55	.59	.67	.55	.66	.66	.65	.65	.49	.70	.55	.69	.56	.46	.42	.67	.68	.97	.20	.96	.73	.64
Core (same epig only)	.35	.15	.31	.17	.20	.68	.92	.59	.92	.90	.54	.59	.65	.51	.62	.64	.66	.62	.51	.67	.52	.69	.53	.44	.39	.67	.67	1	.20	.95	.70	.62
Tier 1 and 2 Marks	1	1	.94	1	1	.84	.98	1	.99	1	.95	1	.78	.69	.89	.78	.83	.68	.99	.84	.66	.77	.69	.56	.56	.86	.91	1.2	1	1	.89	.78
Tier 1 and 2 Marks (same sample only)	.45	.19	.90	.95	.16	.82	.97	.67	.96	.93	.57	1	.73	.67	.88	.75	.81	.68	.96	.82	.64	.76	.67	.53	.56	.86	.91	1.2	.20	.97	.85	.77
H3K27me3		.92	.65	.96	.98	.68	.85	.98	.94	.85	.94	.29	.48	.45	.46	.59	.37	.58	.75	.59	.42	.50	.43	.20	.05	.21	.20	.32	1	1	.62	.44
H3K36me3	.94		.65	.95	.97	.61	.82	.83	.93	.88	.98	.46	.46	.41	.43	.57	.43	.54	.73	.59	.40	.52	.44	.22	.09	.25	.28	.10	1	.99	.62	.44
H3K4me1	.98	1		.96	.96	.76	.86	.98	.96	.92	.99	.51	.72	.55	.62	.67	.58	.66	.75	.67	.57	.62	.56	.47	.44	.44	.53	.15	1	.99	.73	.61
H3K4me3	.95	.98	.68		.97	.65	.93	.86	.98	.95	.94	.37	.45	.45	.48	.60	.61	.58	.64	.63	.41	.60	.45	.31	.17	.63	.51	1.2	1	.99	.70	.53
H3K9me3	.94	.94	.60	.95		.62	.83	.90	.93	.84	.94	.23	.47	.39	.43	.56	.40	.54	.76	.53	.41	.47	.40	.21	.05	.17	.24	.08	.99	1	.59	.42
H3K27ac	.98	.97	.87	.97	.96		.93	1	.95	.96	.96	.43	.73	.67	.87	.72	.67	.73	.63	.74	.55	.67	.62	.49	.49	.90	.90	.24	1	1	.79	.73
H3K9ac	.98	.95	.68	.98	.97	.73		.92	.94	.93	.87	.44	.53	.50	.56	.65	.77	.67	.54	.67	.43	.62	.48	.46	.31	.83	.67	.82	.99	1	.73	.62
DNase	1	.94	.76	.97	.96	.80	.87		.96	.93	.96	.38	.55	.55	.58	.66	.52	.68	.59	.68	.50	.61	.52	.37	.27	.56	.52	.74	1	1	.71	.58
H3K79me2	.95	.92	.52	.95	.90	.57	.85	.82	.89	.73		1	.40	.28	.35	.46	.54	.41	.98	.49	.25	.41	.26	.21	.12	.33	.28	.17	.98	.97	.60	.39
H3K18ac	.97	.96	.82	.97	.97	.83	.90	.96	.97	.92	.87	.33	.83	.89	.81		.86	1	.60	.93	.83	.93	.86	.83	.80	.87	.91	.52	.99	1	.87	.88
H3K18ac+H3K79me2	.97	.97	.80	.97	.96	.83	.91	.95	.97	.92	.87	1	.84	.87	.82	.46	.92	.95	.89	.94	.80	.96	.85	.83	.80	.89	.92	.49	.99	1	.90	.88
H3K18ac+H3K79me2(same samp. only)	.05	.10	.63	.50	.30	.76	.69	.52	.53	.68	.24	1	.79	.87	.81	.14	.88	.95	.91	.89	.78	.95	.84	.83	.80	.88	.92	.49	.06	.89	.71	.85

H3K18ac + H3K79me2 more informative for most mark imputations than core set in a new cell type given an existing roughly uniform coverage compendium

Imputed signal data shows stronger H3K27ac-GWAS associations





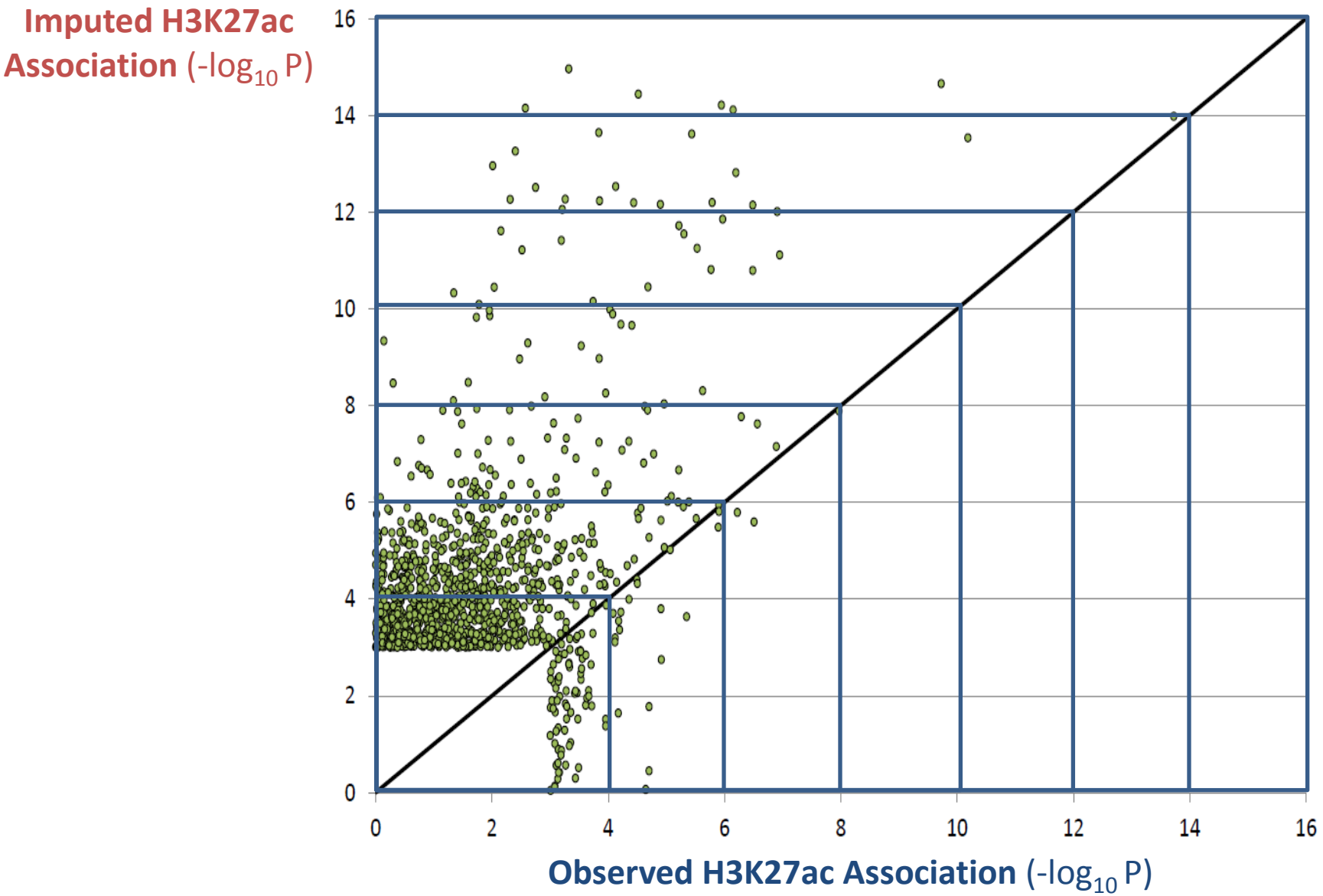
Method:

- H3K27ac association for GWAS catalog (Hindorff et al, 2009)
- GWAS-Tissue association vs. all GWAS SNPs (Mann-Whitney test)
- Restrict to 98 common samples (1MB pruned)

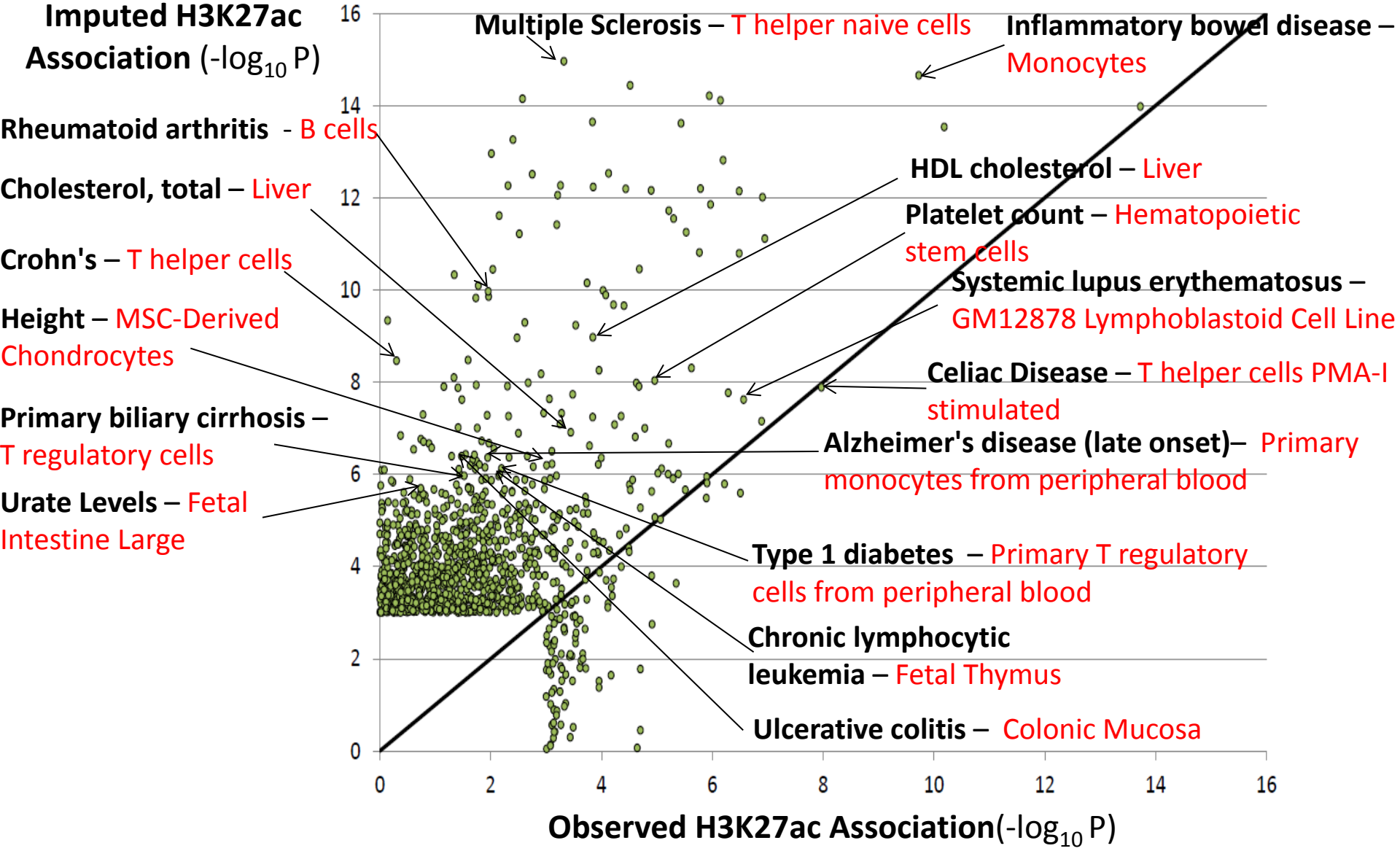
Results: Imputed H3K27ac shows higher association than observed

- More significant P-value for **most-significant tissue** in each trait
- Higher total number of significant tissues across **all tissues and traits**

Imputation improves trait-relevant tissue association

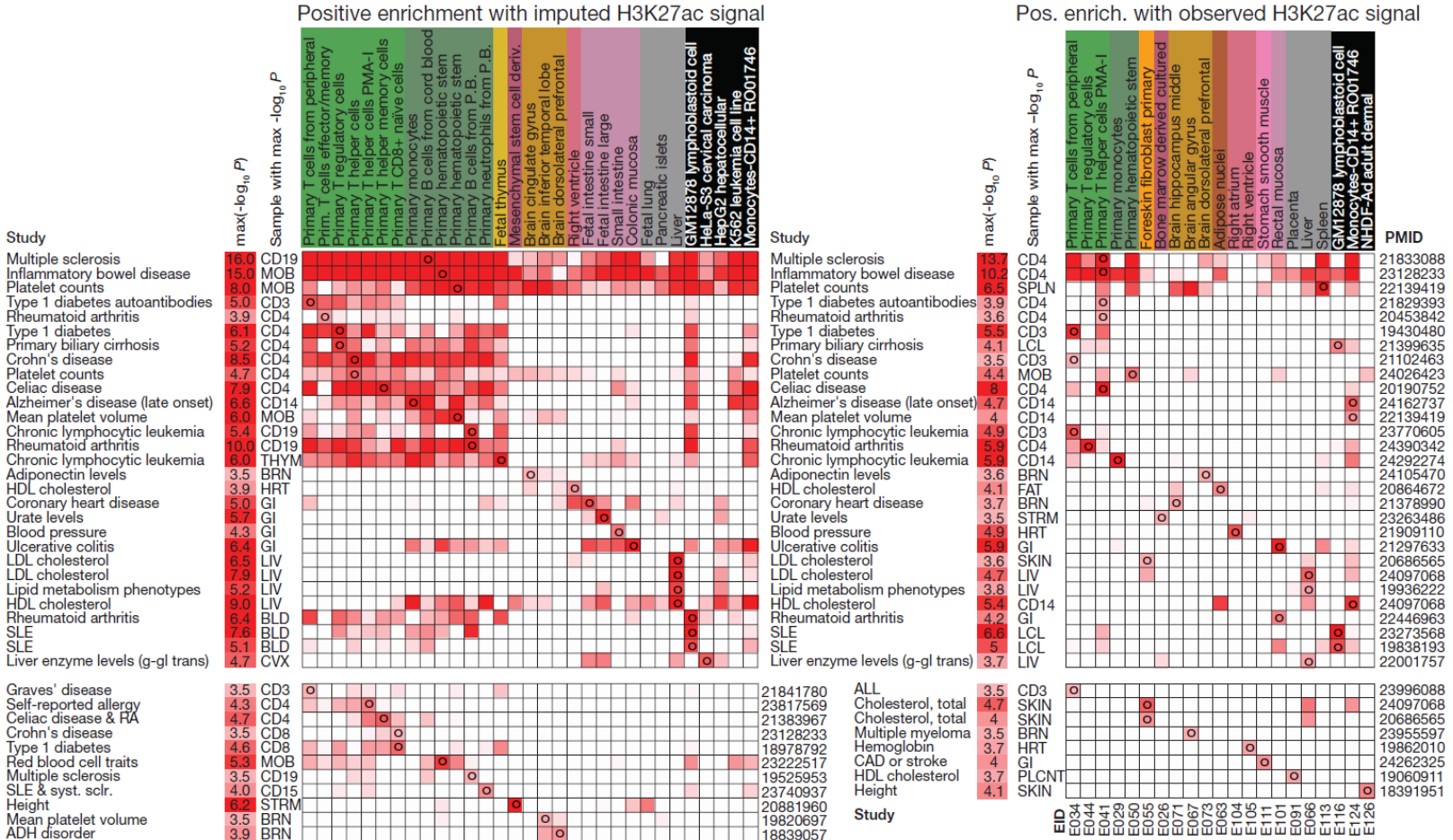


Imputation improves trait-relevant tissue association

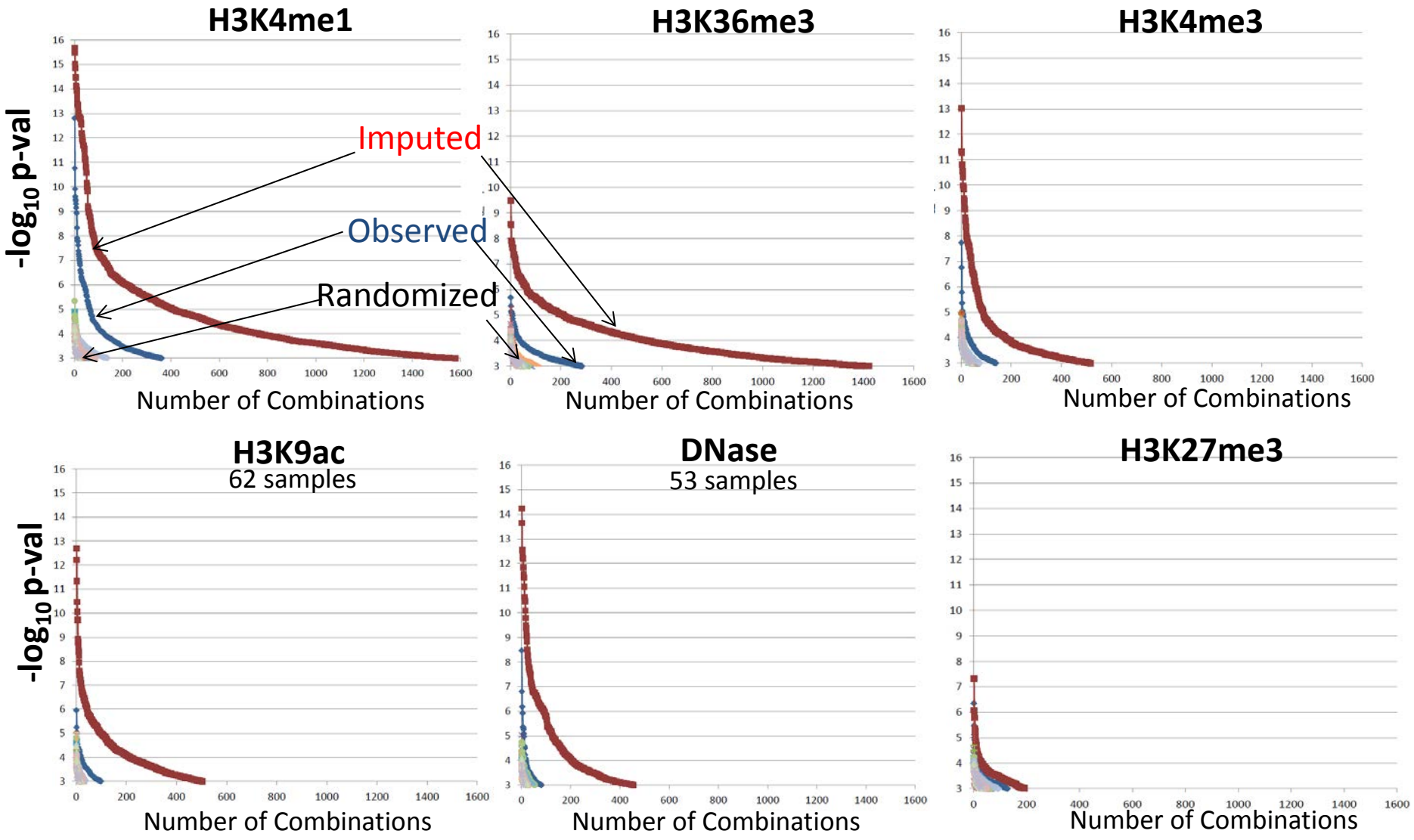


Most significant enrichment shown for observed or imputed data

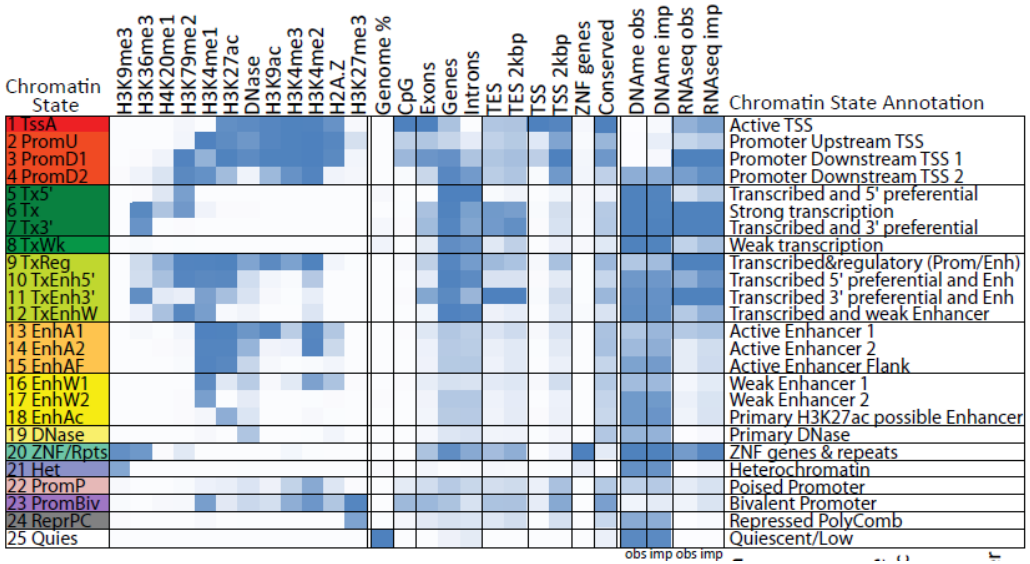
Imputation improves trait-relevant tissue association

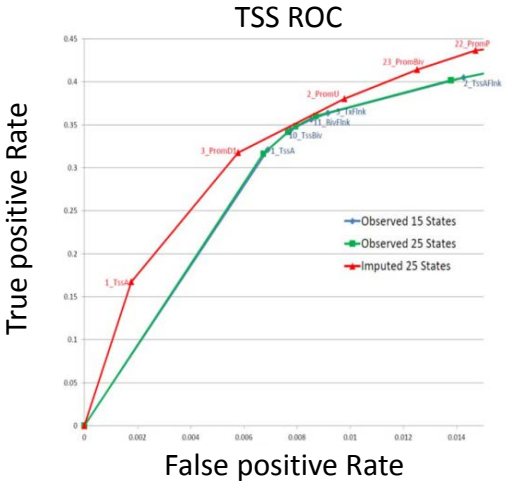


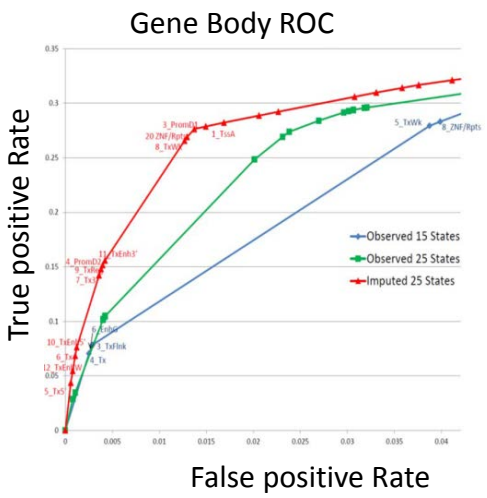
Significant Sample-Study Combinations Additional Marks

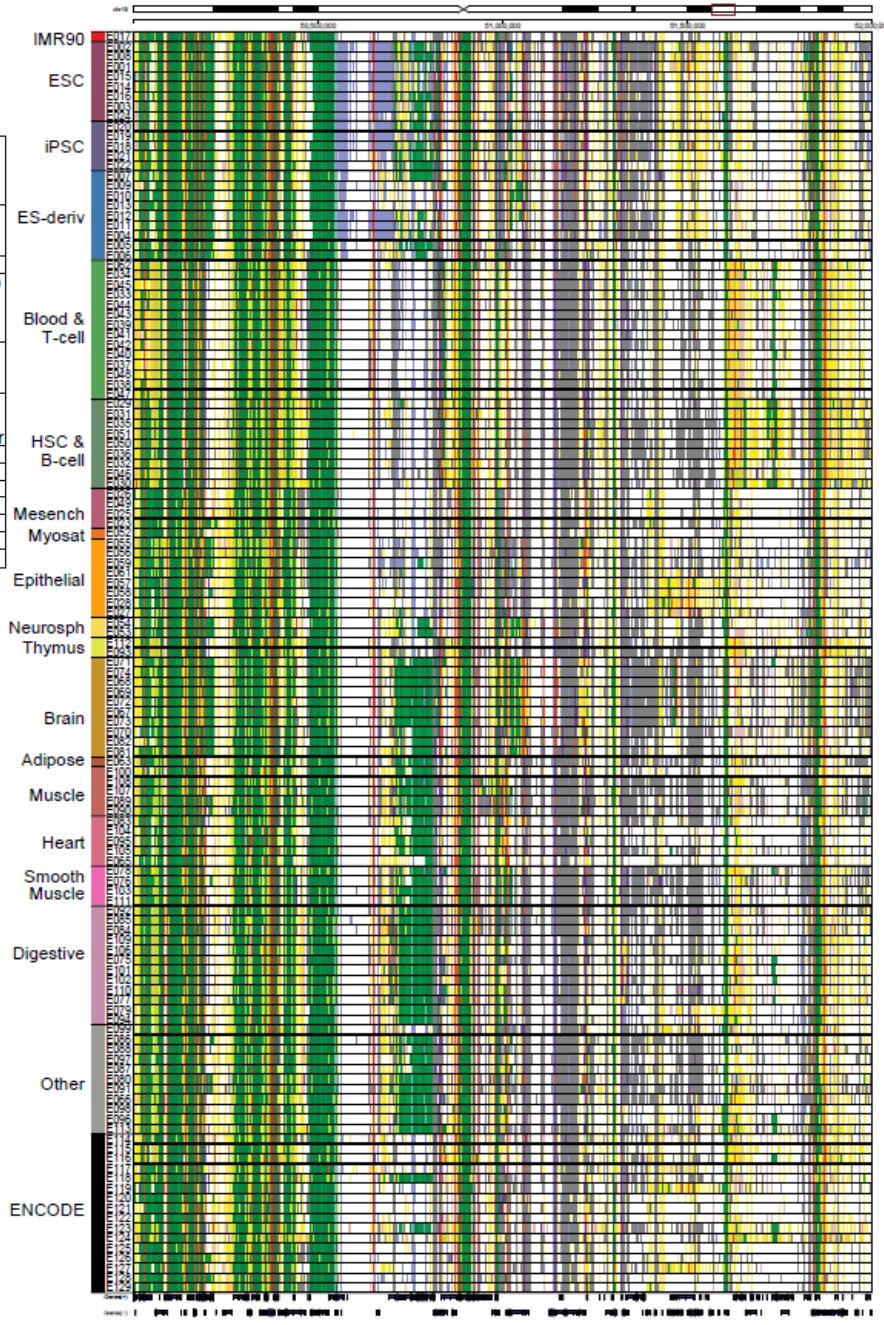


25 chromatin states from 12 marks imputed in 127 cells









Chromatin states based on ChromHMM (Ernst and Kellis, 2012)
Observed model based on 5-core marks

Summary

- ChromImpute method to impute epigenomic data
  - Predict data sets not experimentally mapped
  - Provides a more robust version of experimentally mapped data
- Imputed data and chromatin states a resource to interpret locations identified by GWAS

Acknowledgements

- Manolis Kellis
- Roadmap Epigenomics Consortium
  - Anshul Kundaje
  - Wouter Meuleman
  - Misha Bilenky
- ENCODE Consortium
- Funding: NIH, NSF, Sloan

URLs:
http://www.biolchem.ucla.edu/labs/ernst/ChromImpute/ (software)
http://compbio.mit.edu/roadmap (data links)
http://epigenomegateway.wustl.edu/browser/roadmap/ (browser view)