nature biotechnology

OPEN

Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues

Jason Ernst[1–5] & Manolis Kellis[6,7]

With hundreds of epigenomic maps, the opportunity arises to exploit the correlated nature of epigenetic signals, across both marks and samples, for large-scale prediction of additional datasets. Here, we undertake epigenome imputation by leveraging such correlations through an ensemble of regression trees. We impute 4,315 high-resolution signal maps, of which 26% are also experimentally observed. Imputed signal tracks show overall similarity to observed signals and surpass experimental datasets in consistency, recovery of gene annotations and enrichment for disease-associated variants. We use the imputed data to detect low-quality experimental datasets, to find genomic sites with unexpected epigenomic signals, to define high-priority marks for new experiments and to delineate chromatin states in 127 reference epigenomes spanning diverse tissues and cell types. Our imputed datasets provide the most comprehensive human regulatory region annotation to date, and our approach and the ChromImpute software constitute a useful complement to large-scale experimental mapping of epigenomic information.

Genome-wide maps of epigenetic information, including histone modifications, DNA methylation and open chromatin, have emerged as a powerful means to discover tissue and cell type-specific putative functional elements and to gain insights into the genetic and epigenetic basis of disease[1–9]. Given the dynamic nature of epigenomic datasets across cell types and conditions, discovery power increases with broader coverage of diverse samples. However, owing to cost, time or sample material availability, it is not always possible to map every mark in every tissue, cell type and condition of interest. As a result, analyses that require completed sample-mark data matrices sometimes choose to restrict their comparisons to only those marks that have been commonly mapped across different samples, leading to exclusion of marks or samples that did not have full coverage. An additional, often underappreciated issue is that even when a mark is mapped in a sample, it is usually done with few (if any) replicates, which can confound biological comparisons owing to experimental variability. This situation is exacerbated when analyzing large compendiums of datasets whose sheer number increases the likelihood that there will be outlier datasets of lower quality. Lastly, even for high-quality experiments, robustness of the resulting signal level estimates may be reduced because of insufficient sequencing depth, especially for broadly distributed marks that span a large fraction of the genome.

To address these challenges, we developed ChromImpute for large-scale imputation of epigenomic datasets. ChromImpute uses a compendium of epigenomic maps (such as those generated by the NIH Roadmap Epigenomics and ENCODE projects[2,10]) to generate genome-wide predictions of epigenomic signal tracks (such as histone marks, DNA accessibility, DNA methylation, RNA-seq or any coordinate-based signal track). We used ChromImpute to predict signal tracks of histone modifications, DNA accessibility and RNA-seq at 25-base-pair (bp) resolution and whole genome bisulfite DNA methylation data at single-nucleotide resolution (we refer to all of these data types as 'marks' for simplicity). We annotated a total of 127 reference epigenomes, including 111 generated by the Roadmap Epigenomics project[10] and 16 generated by the ENCODE project[2,3]. These span diverse cell types and tissues (we refer to them as 'samples' for simplicity, even though some reference epigenomes were based on multiple independent samples[10]).

We provide a systematic evaluation of the imputed data and demonstrate that the imputed data for a mark in a sample better matches the corresponding observed data than the observed data from any other sample. We also demonstrate how comparison of observed data and imputed data provides a state of the art data quality control metric that complements and surpasses existing methods. Even when a mark has been experimentally profiled in a sample, we show that imputed data are generally more consistent, robust and accurate, as the data leverage information from hundreds of datasets and thus are resilient to noise arising in individual experiments. The 'prior expectation' of a genome-wide signal provided by the imputed data can also be used in conjunction with observed datasets for inference of surprising signal locations in high-quality samples. We also use the imputation quality of subsets of marks to provide recommendations and insights into experiment prioritization. Lastly, we use a compendium of 12 imputed marks in 127 reference epigenomes to predict and annotate a set of 25 chromatin states, providing the most comprehensive annotation of epigenomic state information in the human genome to date.

[1]Department of Biological Chemistry, University of California, Los Angeles, California, USA. [2]Computer Science Department, University of California, Los Angeles, California, USA. [3]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA, Los Angeles, California, USA. [4]Jonsson Comprehensive Cancer Center, University of California, Los Angeles, California, USA. [5]Molecular Biology Institute, University of California, Los Angeles, California, USA. [6]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts, USA. [7]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. Correspondence should be addressed to J.E. (jason.ernst@ucla.edu) or M.K. (manoli@mit.edu).

Received 13 May 2014; accepted 2 February 2015; published online 18 February 2015; doi:10.1038/nbt.3157

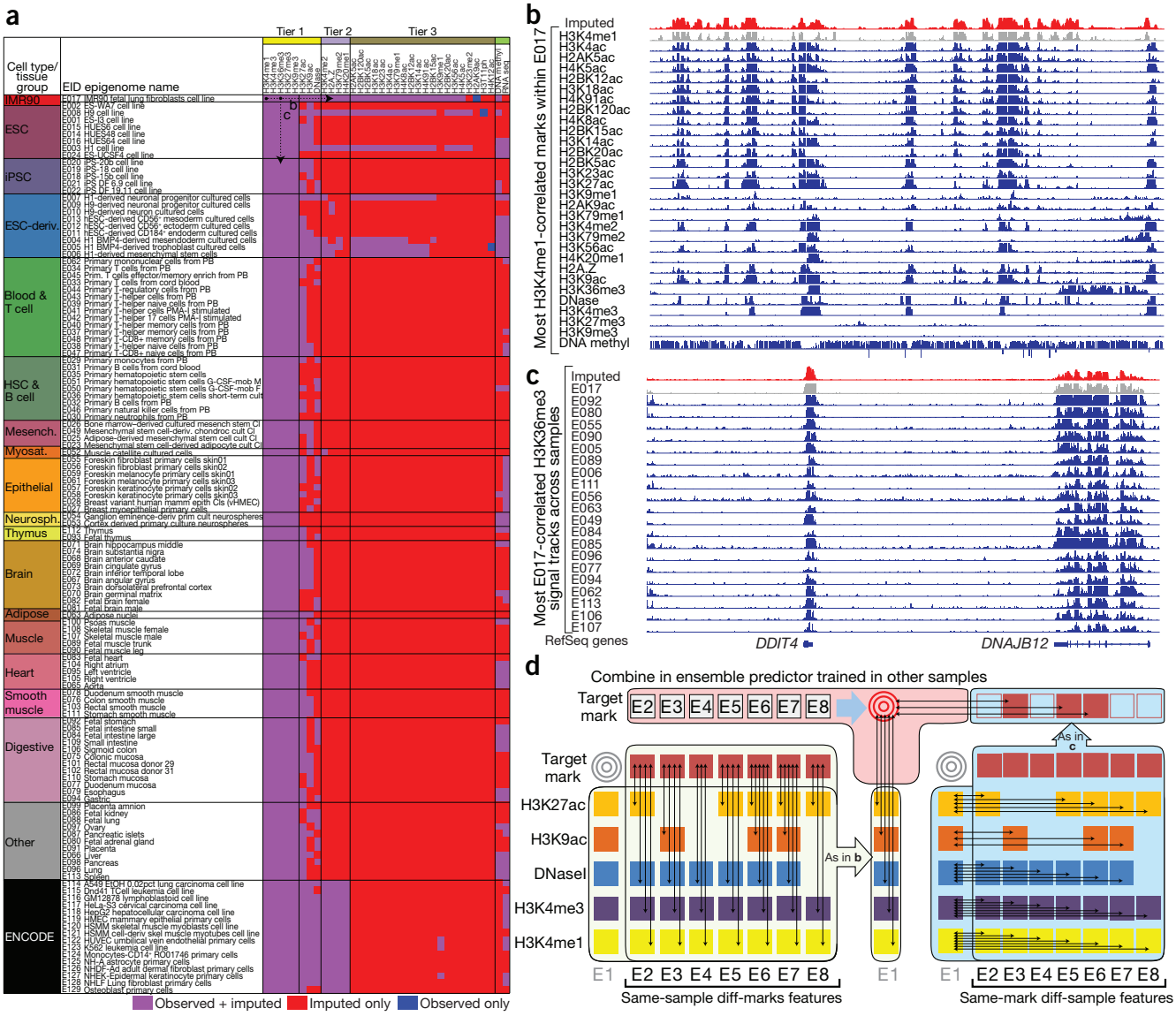

Figure 1 Application and method overview. (**a**) Matrix of observed and imputed datasets across 127 reference epigenomes ('samples'), including 111 from the Roadmap Epigenomics project (rows 1–111) grouped and colored by cell/tissue type, and an additional 16 from ENCODE (rows 112–127), with reference epigenome identifier (EID) and short sample/tissue description. Epigenomic marks (top) are grouped by tiers 1–3 plus RNA-seq and DNA methylation (DNA methyl), based on experimental coverage and imputation strategy. Black dotted arrows on the top denote E017 datasets shown in **b** (horizontal arrow), and H3K36me3 datasets shown in **c** (vertical arrow), illustrating the two dimensions of correlations used in ChromImpute and shown in **d**. PB, peripheral blood; Mesench., Mesenchymal; cult cl, cultured cells. (**b**) Correlation between epigenomic marks in the same sample, one of the two classes of features used for epigenome imputation. Datasets from sample E017 are shown, illustrating their highly correlated nature, comparing the observed signal for H3K4me1 from E017 (gray), the imputed data (red), which was predicted without using the observed data, and the observed tracks for other marks (blue), ordered based on their correlation with the H3K4me1. Imputation of H3K4me1 in E017 (red) does not use the observed data (gray), and instead uses the other samples to learn relationships between H3K4me1 and other marks. DNA methylation values below the horizontal line represent missing data. For the primary imputation of H3K4me1, not all marks shown were used, as only tier 1 marks are used to impute tier 1 marks. (**c**) Multiple signal tracks for H3K36me3 across samples illustrate the highly correlated nature of a given mark across samples, exploited in the second class of features used for epigenome imputation. This example uses the same region as used in **b** to compare the observed signal for H3K36me3 in E017 (gray), H3K36me3 in several other samples (blue), which constitute the basis for highly informative features for H3K36me3 imputation in E017 (red). Observed tracks (blue) are ordered by their global correlation to the observed H3K36me3 signal in E017, though ChromImpute did not have this information when imputing H3K36me3 in E017, and instead determined sample similarity based on other marks, both globally and locally at each position, and then used the H3K36me3 signal in up to ten most-proximal samples for each definition of similarity to compute individual features for each predictor of the ensemble (**d**, right). (**d**) Ensemble strategy for signal track imputation using features that exploit correlations between marks in the same sample (left) and correlations between samples for a given mark (right). We assume that no information is available for the target mark in the target sample (gray targets). Thus, we learn relationships between marks (left side) in other samples (column of E1 sample is not used) and learn relationships between samples (right side) using other marks from which we then compute same-mark features. The ensemble predictor that combines features across marks (**b**) and across samples (**c**) is learned only in other samples (top), and the marks in the target sample are used only during the actual application of the trained ensemble predictors to compute the imputed signals.

RESULTS

ChromImpute method and previous work on imputation

Imputation has been previously explored in a number of bioinformatics settings. For microarray experiments, missing gene expression values have been predicted for specific genes in specific experiments[11]. For genome-wide association studies (GWAS), missing genotype values are routinely predicted for single-nucleotide polymorphisms (SNPs) not directly assayed, by exploiting common haplotype structure[12]. For epigenomic datasets, prediction of both DNA methylation and histone modification datasets has been undertaken from DNA sequence information[13–15], but the static nature of genome sequence limits the ability to generate cell-type-specific predictions for samples not previously used for training, as the motifs driving a given mark frequently differ across samples. Specifically for DNA methylation, imputation has been undertaken using sequence-based features and histone modification data from one sample[16,17], lower resolution assays in conjunction with sequence information and other annotations for predicting high-resolution DNA methylation[18], or assumed phylogenetic relationships between cell types[19]. For histone modifications and other chromatin marks, methods have been developed by us and others, to infer chromatin states based on multiple marks, even in cases with missing data[20–22], but these do not try to infer the actual signal for the missing marks. Several other methods have been developed to model correlations of histone marks with expression or with other marks in a single sample[23–26], which have sometimes been leveraged for imputation on a limited scale, but have not considered across-sample information. In practice, studies interested in a given cell type sometimes use data from a related cell type, which can be viewed as one simple approach to imputation.

Here, we take an ensemble regression-based approach to epigenomic imputation. We impute each target mark in each target sample separately, by combining information from large numbers of datasets that were experimentally determined, but without using any data for the target mark in the target cell type (**Fig. 1a** and **Supplementary Fig. 1**). We leverage two classes of features (**Fig. 1d**).

1. Same-sample (different-mark) information (**Fig. 1b**): The first class of features uses information from the signal of other marks mapped in the target sample, both at the target position and at neighboring sites.
2. Same-mark (different-sample) information (**Fig. 1c**): The second class of features uses information from the signal of the specific mark of interest at the target position in the most similar samples. Similar samples are defined based on similarity with the signal of marks that have been mapped in the target sample both locally and globally. The features in this class are effectively predictions that could be made by a K-nearest neighbor method for various values of K and distance functions.

As no training data are available for the target mark in the target sample, we learn the relationships between the features and the target mark using other samples that contain the target mark. We use regression trees[27], as they can handle nonlinearities (including the constraint that signal values are non-negative), they support combinatorial interactions among features, and they are relatively fast to train. The prediction for each target mark in each target sample is based on an ensemble predictor that averages the values resulting from regression trees trained on each sample in which the target mark is available, thus reducing the impact of biases from any one individual predictor.

Imputation of 4,315 datasets in 127 reference epigenomes

We applied ChromImpute to a compendium of 127 reference epigenomes, including 111 profiled by the NIH Roadmap Epigenomics project[10] and 16 profiled by the ENCODE project[2,3] (**Fig. 1a**). These span diverse tissues and cell types, including embryonic stem cells (ESCs), induced pluripotent stem cells (iPSC), ESC-derived cells, blood and immune cells, skin, brain, adipose, muscle, heart, smooth muscle, digestive, liver, lung and others.

Only five 'core' histone modification marks were experimentally profiled in all 127 reference epigenomes. These are promoter-associated H3K4me3, enhancer-associated H3K4me1, Polycomb repression-associated H3K27me3, transcription-associated H3K36me3 and heterochromatin-associated H3K9me3. Varying subsets of 34 marks were profiled in different epigenomes, including 30 histone modifications (11 histone methylation marks, 18 histone acetylation marks and H3T11ph), histone variant H2A.Z, DNA accessibility (profiled by DNase I hypersensitivity), DNA methylation data (profiled by Whole-Genome Bisulfite Sequencing, WGBS) and RNA-seq data.

Based on these experimentally profiled ('observed') datasets, we imputed the 31 marks observed in at least two epigenomes in all 127 epigenomes, and the three marks mapped in only one epigenome in the remaining 126 epigenomes. In total we generated 4,315 datasets based on imputation, of which only 1,122 (26%) were also experimentally mapped and 3,193 (74%) were available only as imputed data. Signal tracks for all marks were imputed at 25-bp resolution (121 million predictions per track) except for DNA methylation, which was imputed at single-nucleotide resolution for each of 28 million CpGs. Across all marks, samples and positions, we generated a total of 526 billion predicted signal values.

We categorized the 34 epigenomic marks into four classes according to the number of samples in which they were experimentally profiled and our imputation strategy (**Supplementary Fig. 2**).

1. Tier 1 marks were mapped broadly across samples, were used to impute all other datasets and were imputed using only tier 1 marks. They consist of H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3, H3K27ac, H3K9ac and DNA accessibility.
2. Tier 2 marks were mapped broadly only in ENCODE samples, were used to impute tier 2 and tier 3 marks, and were imputed using only tier 1 and tier 2 marks. They consist of H3K4me2, H3K79me2, H4K20me1 and H2A.Z.
3. Tier 3 marks had limited coverage, were used only to impute tier 3 marks and were imputed using all three tiers. They consist of the remaining 20 histone modification marks.
4. DNA methylation and RNA-seq datasets were treated separately as a design choice due to their very distinct natures. RNA-seq datasets were imputed using only tier 1 marks and other RNA-seq datasets and, similarly, DNA methylation datasets, only using tier 1 marks and other DNA methylation datasets.

This tiered approach for histone marks and DNA accessibility datasets enabled us to limit potential biases resulting from the lower number of samples for tier 2 and tier 3 marks (reducing only minimally the information available for making predictions).

Imputed datasets capture missing marks effectively

As an initial control, we assessed by visual inspection the level of similarity between pairs of matching imputed and observed datasets, using nine randomly selected 200-kb regions and 2,000 randomly selected 25-bp regions. For each of the nine broad regions, we randomly selected one sample in which the mark was also experimentally profiled and visualized imputed and observed tracks in detail (**Fig. 2a** and **Supplementary Fig. 3**). For the 2,000 samples, we generated a dense heatmap showing the observed and imputed mark signal across

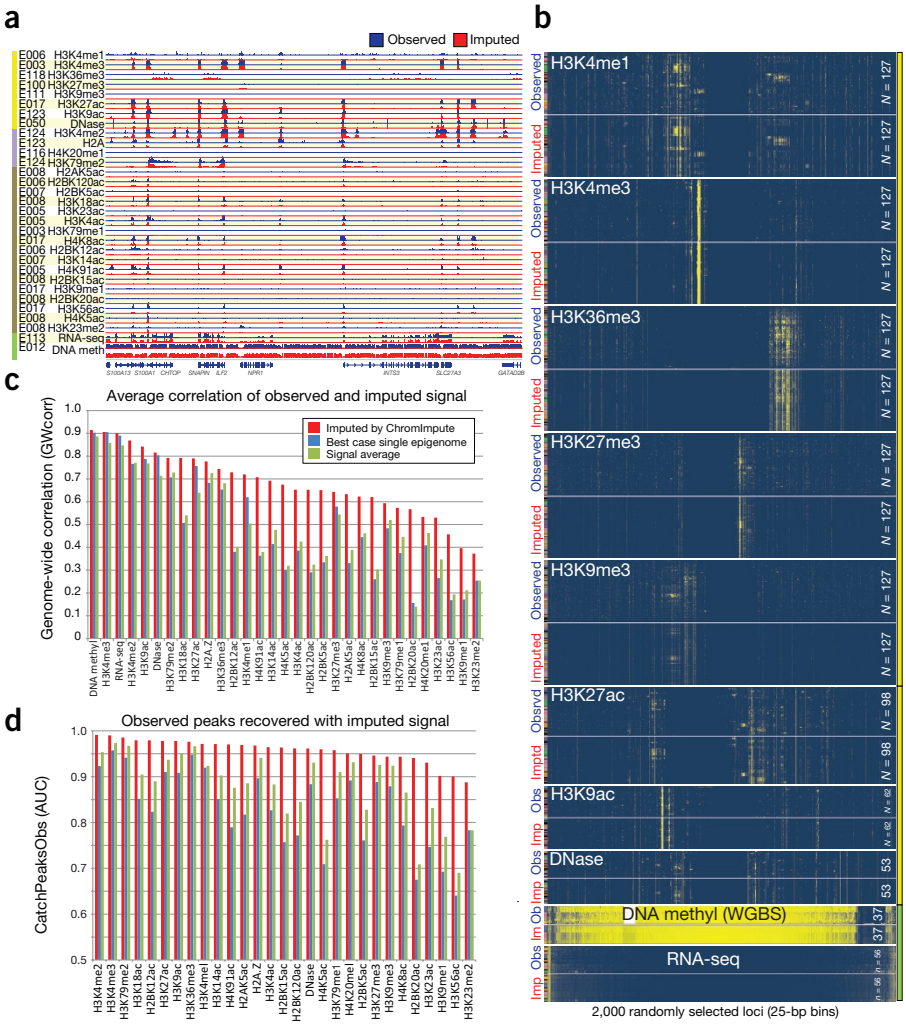

Figure 2 Imputed data are a close match to observed datasets. (**a**) Visualization of one of the randomly selected 200-kb regions illustrates high-resolution concordance between observed (blue) and imputed (red) signal tracks. Imputed tracks are generated at 1-bp resolution for DNA methylation and 25-bp resolution for all other marks and trained without using the observed track. For each mark (row), we show a randomly selected sample (EID from **Fig. 1a**), which also contains observed data for comparison (light purple entries in **Fig. 1a**). This region was chosen among nine randomly selected 200-kb regions (**Supplementary Fig. 3**) as the one with the most signal across all marks. Larger 1.5 Mb context, and example 5-kb close-up are shown in **Supplementary Figure 3c**, illustrating concordance at multiple resolutions. (**b**) Visualization of 2,000 randomly selected 25-bp regions (columns), and their signal (yellow, high; blue, low) across up to 127 samples (rows, colored as in **Fig. 1a**), for tier 1 marks (yellow sidebar) and RNA-seq and DNA methylation (green sidebar) (tier 2 and tier 3 marks are shown in **Supplementary Fig. 4**). Rows and columns are clustered for each mark independently to highlight structure based on observed data (top), and imputed data (generated without using the corresponding observed dataset) are shown below, in the same order, showing clear similarity. WGBS, whole genome bisulfite sequencing. (**c**) Quantitative comparison of observed signal correlation for ChromImpute (red), averaging the mark signal from all other samples (green), and the best-case for selecting a single sample (blue), which is not a realistic method when the target mark signal is not known, as it would be needed to determine the single-best sample. Average correlation is computed based on all samples for which both observed and imputed signals are available. ChromImpute shows consistently higher correlation of observed signals than the two alternate methods (including the unrealistic best case) for all marks. For additional comparisons see **Supplementary Figures 5**–**7**. (**d**) Average AUC for recovering bases covered by a narrow peak call on observed data[10] when ranking based on predicted signal.

every sample in which both were available (**Fig. 2b** and **Supplementary Fig. 4**). Both visual comparisons showed strong agreement between observed and imputed signal, successfully recovering epigenomic features at high resolution, across broad regions (**Fig. 2a** and **Supplementary Fig. 3c**) and in a tissue-specific way (**Fig. 2b**). Beyond the visualizations provided in this paper, imputed and observed tracks are provided for the entire genome through public track hubs on the WashU Epigenome Browser (http://epigenomegateway.wustl.edu/browser/)[28] and the UCSC Genome Browser[29].

We also assessed the ability of ChromImpute to predict missing marks using seven quantitative metrics: (i) the genome-wide correlation between observed and imputed data ("GWcorr," **Fig. 2c**); (ii) the overlap between imputed and observed datasets in the top 1% of the 25-bp bins with the highest signal ("Match1"); (iii) the percentage of the top 1% observed in the top 5% imputed 25-bp bins ("Catch1obs"); (iv) the percentage of the top 1% imputed in the top 5% observed 25-bp bins ("Catch1imp") (**Supplementary Figs. 5**–**7**); (v) the recovery of the top 1% observed and (vi) 1% imputed 25-bp bins based on the full range of signal of the other using the area under the curve (AUC) of a receiver operating characteristic (ROC) curve ("AucObs1" and "AucImp1," **Supplementary Figs. 5**–**7**); (vii) and the AUC recovery of bases covered by observed peak calls based on the full range of signal of the imputed data ("CatchPeakObs," **Fig. 2d** and **Supplementary Figs. 6**–**7**). These 1% and 5% percentages captured the diversity of chromatin states for each mark (**Supplementary Fig. 8**) and captured the majority of high-signal locations (**Fig. 2b** and **Supplementary Fig. 4**; see also genome-wide signal distributions discussed below). For DNA methylation, we used GWcorr and "Methyl25," a previously suggested concordance measure that considered two DNA methylation values to be in agreement if they were within 0.25 of each other[30], as focusing on the top few percent of signal is less meaningful (as the vast majority of CpG dinucleotides in the human genome are highly methylated).

To provide perspective on the performance of ChromImpute in each metric, we compared it to two stringent baselines. The first baseline, 'BestSingle', predicts a missing mark based on the signal of the most similar experimental dataset for the target mark, according to the specific metric measured across any other sample. This baseline is unrealistic as an imputation method because the most similar experiment is not known in advance, and is not available to ChromImpute or to any prediction method. The second baseline, 'SignalAvg', predicts the average signal of the target mark across all other samples and can be thought of as an alternative imputation approach.

ChromImpute showed strong recovery of observed datasets, both in its overall performance, and relative to both stringent baselines. For the GWcorr metric, ChromImpute showed 0.68 correlation on average per mark (vs. 0.50 for both BestSingle and SignalAvg, **Fig. 2c**), outperform-

ing BestSingle for 91% of datasets and SignalAvg for 99% of datasets per mark on average. ChromImpute showed AUC = 0.95 recovery for AucObs1 (vs. 0.84 and 0.88, **Supplementary Fig. 5**) on average per mark, and AUC = 0.96 for CatchPeakObs (vs. 0.83 and 0.88) (**Fig. 2d**). For the Methyl25 metric, ChromImpute outperformed SignalAvg 97% of time, and BestSingle, 76% of the time.

We also compared ChromImpute to several additional imputation approaches. First, we implemented ChromImpute-LR, using the same ensemble training strategy but linear regression instead of regression trees to combine features. ChromImpute had overall similar or better performance than ChromImpute-LR for the tier 1 and 2 marks and much better performance for DNA methylation, although ChromImpute-LR showed somewhat better performance for some tier 3 marks, which had fewer training datasets available (**Supplementary Fig. 9**). Second, for tier 1 histone marks in ESCs and iPSCs, we compared ChromImpute to a predictor based on averaging of increasingly larger number of these near-replicate datasets (**Supplementary Fig. 10**). Predictive power increased by averaging more replicates, but ChromImpute showed better predictive power than ten near-replicates for some marks, and three near-replicates for all marks (**Supplementary Fig. 10**). Third, ChromImpute also outperformed nearest-neighbor predictors of a mark based on local and global distance, a predictor trained on only one sample instead of the full ensemble (**Supplementary Fig. 9**) and a predictor based on averaging active marks in the same sample to predict other active marks and likewise for repressive marks (**Supplementary Fig. 11**), in each case supporting our imputation strategy.

Increased robustness and annotated feature recovery

Although the previous analyses demonstrated that the imputed datasets provided a reasonable approximation to observed datasets, and thus can be beneficial when observed data are not available, we next investigated whether imputed datasets also have distinct advantages that make them valuable even if observed datasets are available. Two potential reasons may lead to advantages for imputed datasets: (i) imputed datasets are based on combining information from many experiments and thus have the potential to be more robust to experimental noise and other confounders than the observed data; (ii) by combining relevant information from many related experiments, imputed data can achieve a higher 'effective' sequencing depth, and thus potentially a higher signal-to-noise ratio.

We used the property that promoter-associated H3K4me3 frequently localizes near transcription start sites (TSS) and that transcription-associated H3K36me3 frequently localizes in gene bodies. We defined two metrics that quantify the extent to which the strongest H3K4me3 signal (at 25-bp resolution) localizes within 2 kb of annotated TSS ("PromRecov," **Fig. 3a**) and the strongest H3K36me3 signal localizes in gene bodies ("GeneRecov" **Fig. 3b**), using AUC for the portion of the ROC curve that has a 5% false-positive rate or less (we primarily focused on this metric instead of the full AUC as we expected many annotated locations not to be marked by the observed or imputed data in any one sample, but saw similar results based on the full AUC (**Supplementary Fig. 12a**,**b**)).

We found that imputed data showed better annotation agreement than observed data for every dataset, often by a large margin (**Supplementary Fig. 13**). In fact, the worst-performing imputed H3K4me3 dataset performed better than 96% of observed H3K4me3 datasets, and the worst-performing imputed H3K36me3 dataset performed better than 91% of observed datasets in the evaluations (**Fig. 3a**,**b**). Recovery of gene bodies for a few of the H3K36me3 observed datasets was only marginally above random, whereas for imputed data, recovery was consistently high. As these results are based only on the rank ordering of signal values, any normalization strategy that preserves the rank ordering (e.g., quantile normalization[31]) would not change these results. We also observed better overall agreement with annotated features when considering peak calls instead of signal level (**Supplementary Fig. 14**).

Additionally, imputed data showed a more robust and consistent signal profile than observed data. Observed H3K4me3 signal proximal to all TSSs showed up to a 95-fold variation between samples (**Fig. 3c**), and observed H3K36me3 showed up to a sevenfold variation in gene bodies (**Fig. 3d**). Suggesting that experimental variability, rather than biological differences, indeed underlies some of these differences, two fetal brain samples (E081 and E082) showed large heterogeneity in their aggregate profiles for H3K4me3 and H3K36me3. E081 showed very flat distributions (**Fig. 3c**,**d**), whereas E082 and the imputed data for E081 and E082 all showed much more recognizable distributions (**Fig. 3c**,**d**). Consistent with experimental confounders, these E081 datasets showed relatively poor scores in both the PromRecov and GeneRecov metrics (**Fig. 3a**,**b**).

Imputed marks also showed higher consistency than observed marks in their genome-wide signal distribution (**Supplementary Fig. 15**). For example, for the observed datasets of H3K36me3 in the two fetal brain samples (E081 and E082), there was an 11.6-fold difference between the amount of the genome that had signal values ≥ 3, whereas imputed data showed only a 1.4-fold difference.

We also used the 28 histone and DNA accessibility marks that were mapped in two different ESC lines (H1 and H9) to compare near replicates for observed and for imputed datasets. We expected that for high-quality datasets, each mark mapped in H1 should show a higher correlation with the corresponding mark in H9 than with other marks in H9 (and conversely for H9 marks). Indeed, this property held more frequently for imputed data versus observed data (**Supplementary Fig. 16**), once more supporting the higher quality of imputed datasets.

Imputed data captured dynamics and sample relationships

To study whether imputed data can capture dynamic epigenomic information across cell types, we evaluated our PromRecov and GeneRecov metrics for tissue-restricted annotations, by focusing specifically on a set of genes that were expressed in the corresponding samples (**Supplementary Figs. 12c**,**d** and **13c**,**d**). Imputed data continued to strongly outperform observed data for the set of expressed genes, with all but one imputed dataset for H3K4me3 showing higher PromRecov, and all but one imputed datasets for H3K36me3 showing higher GeneRecov.

We also compared the ability of imputed and observed data to recover expressed genes as a function of the number of samples in which they were expressed (**Supplementary Fig. 17**). Recovery of both TSS-proximal regions and gene bodies increased greatly with the number of samples in which a given gene is expressed for imputed marks (as expected given the multiple informant samples for each mark) and for observed marks (suggesting that genes detected as more broadly expressed show greater agreement with histone modification marks even for observed data). Notably, imputed H3K4me3 showed higher PromRecov independent of how restricted the expression was to certain samples, even for TSS regions of genes expressed in a single sample. For H3K36me3, observed marks showed a modestly higher recovery of gene bodies for genes expressed in only six samples or fewer (3% of expressed genes in a sample, on average). However, for the remaining genes expressed in larger numbers of samples, imputed datasets consistently outperformed observed datasets.

For all tier 1–3 marks, we directly compared the correlation between observed gene expression levels and the signal data for both

observed and imputed marks (**Supplementary Fig. 18**). For nearly all positively correlated marks, imputed signal showed a greater positive correlation with gene expression than observed signal, both in TSS-proximal regions (**Supplementary Fig. 18a**) and in gene bodies (**Supplementary Fig. 18b**). For negatively correlated marks, observed data showed greater negative correlation with expression than imputed data, but this higher negative correlation was associated with lower-quality observed datasets, and the difference was reduced when focusing only on higher-quality observed data, both in TSS-proximal regions and in gene bodies (**Supplementary Fig. 18c**,**d**).

We also evaluated the ability of both imputed and observed datasets to capture the relationships between tissues and cell types based on genome-wide correlation analysis between pairs of datasets (**Fig. 3e**,**f** and **Supplementary Fig. 19**). Specifically, we compared the imputed and observed data for their ability to group samples in accordance with their tissue group (defined in ref. 10 and shown in **Fig. 1a** of this paper) based on the correlation of individual marks (**Fig. 3e** and **Supplementary Fig. 19**). We found the imputed data showed a correlation matrix with a strongly pronounced block structure, corresponding to the biological groupings of cell types and tissues. This was substantially weaker in observed datasets, suggesting imputed data better captured sample relationships.

To quantify this difference, we evaluated the ability of each tier 1 mark, DNA methylation and RNA-seq to distinguish same-group versus different-group sample pairs (excluding the heterogeneous 'ENCODE' and 'Other' groups), based on the relative genome-wide pairwise correlation, evaluated as the AUC for both observed and imputed signal (**Fig. 3f**). Imputed data consistently outperformed observed data, showing an average AUC of 0.92 versus 0.79 for observed data. The increase in classification power was most pronounced for H3K4me3, H3K36me3, H3K27me3 and H3K9me3, which are generally considered less cell-type specific (AUC = 0.93 vs. 0.70).

These results also held for sample group classification based on histone mark peak call similarity (**Supplementary Fig. 20**), when trying to distinguish pairs of samples having the same anatomy annotation from those that have a different one[10] (with all marks except DNA methylation showing increased accuracy for imputed data compared to observed data, **Supplementary Table 1** and **Supplementary Fig. 20**), and for higher-resolution distinctions beyond the tissue group level, as ChromImpute predictions showed higher correlation with corresponding observed data than predictions obtained by averaging all other same-group experiments (**Supplementary Fig. 21**). We reasoned that perhaps a weighted average of observed and imputed data may further

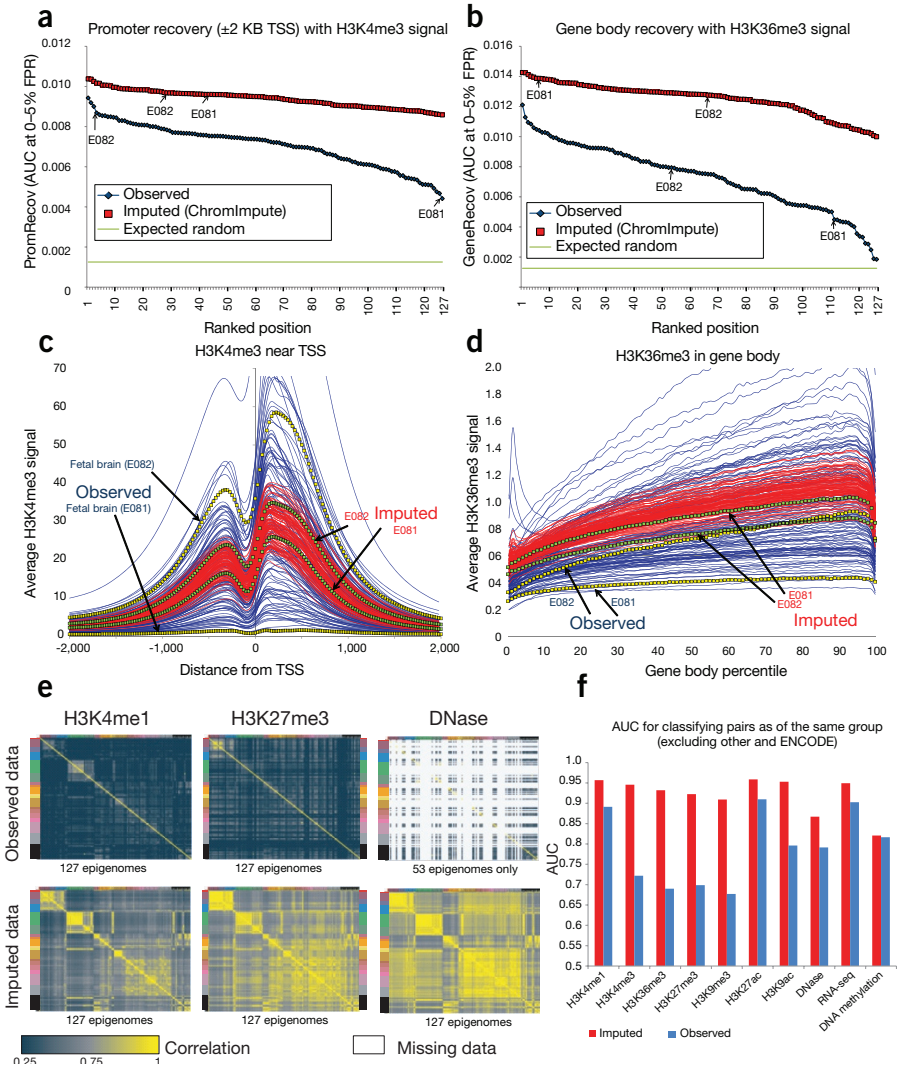

Figure 3 Imputed data shows higher promoter/gene recovery, robustness and biological group recovery. (**a**,**b**) Quantitative comparison of observed (blue) and imputed (red) data in their recovery of annotated promoters (**a**) and gene bodies (**b**), based on the area under the ROC curve up to a 5% false-positive rate (*y* axis) for H3K4me3 signal recovery of locations within 2 kb of TSS (**a**) and H3K36me3 signal recovery of gene bodies (**b**). Arrows indicate two fetal brain samples (E081 and E082) with very different values in the observed data, which show much higher (and more consistent) recovery for imputed data. FPR, false-positive rate. (**c**,**d**) Comparison of aggregate signal for imputed (red) and observed (blue) datasets based on $-\log_{10}$ *P* value of H3K4me3 surrounding the TSS (**c**) and H3K36me3 in gene bodies (**d**). Imputed data show a substantially more consistent profile across all datasets, and in particular for the two fetal brain samples (E081, E082), which show substantial differences in the observed data. (**e**) Pairwise comparison of genome-wide signal correlation for all samples using observed (top) and imputed (bottom) data for H3K4me1, H3K27me3 and DNase (additional marks shown in **Supplementary Fig. 19**), with samples ordered and colored as in **Figure 1a** (left sidebar). Imputed datasets better capture biological relationships between samples than observed datasets, with their correlation structure clearly delineating pluripotent cells, immune cells, adult brain and multiple tissue groups (**Fig. 1a**), whereas observed datasets are much less correlated even for highly similar samples. (**f**) Area under the ROC curve for classifying whether two different pairs of experiments belong to the same group when ranking the pairs based on their correlation. A value of 0.5 could be achieved by random guessing and a value of 1.0 is the maximum possible score. The 'Other' and 'ENCODE' groups were excluded from this analysis as were imputed pairs that were not present in the observed data. This shows quantitatively that the relative similarity of imputed data sets is more consistent with the biological groupings of the samples.

improve classification power, but we did not see substantial improvement in a combination approach relative to just using the imputed data, except for DNA methylation where a balanced combination showed the highest classification power (**Supplementary Fig. 22**).

Imputed data improved GWAS enrichments

As epigenomic maps have recently emerged as an unbiased approach for discovering disease-relevant tissues and cell types[3,32], we also evaluated the impact of epigenome imputation on the interpretation of trait-associated variants from GWAS. We quantified the enrichment (positive or negative) of trait-associated variants from the National Human Genome Research Institute (NHGRI) GWAS catalog[33] in both observed and imputed datasets for each tier 1 mark. We evaluated enrichments both in aggregate across all studies, based on area under an ROC curve up to a 5% false-positive rate (AUC5%) for the signal level recovery of trait-associated SNPs, and at the level of individual studies, based on mark signal rank differences between each study's SNPs and all other SNPs in the GWAS catalog. We evaluated both the number of studies for which there was a significant signal rank difference in at least one sample, and the total number of study-sample pairs that were significant, at varying *P* value thresholds. We then compared both the number of significant studies and the number of significant pairs to the numbers obtained for randomized versions of the GWAS catalog, which also enabled us to obtain a false-discovery rate estimate for each *P*-value threshold (**Supplementary Table 2**).

For all tier 1 active marks, imputed data resulted in substantially greater recovery of SNPs in the GWAS catalog than the observed data (**Supplementary Fig. 23**), and more significant enrichments for both the number of studies and the number of study-sample pairs, across all tested significance thresholds (**Fig. 4a** and **Supplementary Figs. 24** and **25**). In addition, the imputed data yielded a stronger enrichment for each enriched study-sample pair in the large majority of cases (**Fig. 4b** and **Supplementary Fig. 26**). We confirmed that the actual GWAS catalog yielded substantially more significant associations than randomized versions, for both the observed and imputed data across a range of *P*-value significance thresholds (**Fig. 4a** and **Supplementary Figs. 24** and **25**). Imputed data performance was substantially higher than that of the average mark signal across all available samples (**Supplementary Fig. 24b**), emphasizing that the higher performance was not simply due to averaging multiple samples. We also confirmed that the samples with the strongest positive enrichments for a given study were generally biologically relevant for active marks. For H3K27ac, for example, we found that liver was the most enriched sample for various cholesterol phenotypes, immune-related cells for various immune-related disorders, colonic mucosa for ulcerative colitis. Many additional biologically meaningful enrichments were found for diverse studies and cell types (**Fig. 4c–f** and **Supplementary Table 2**).

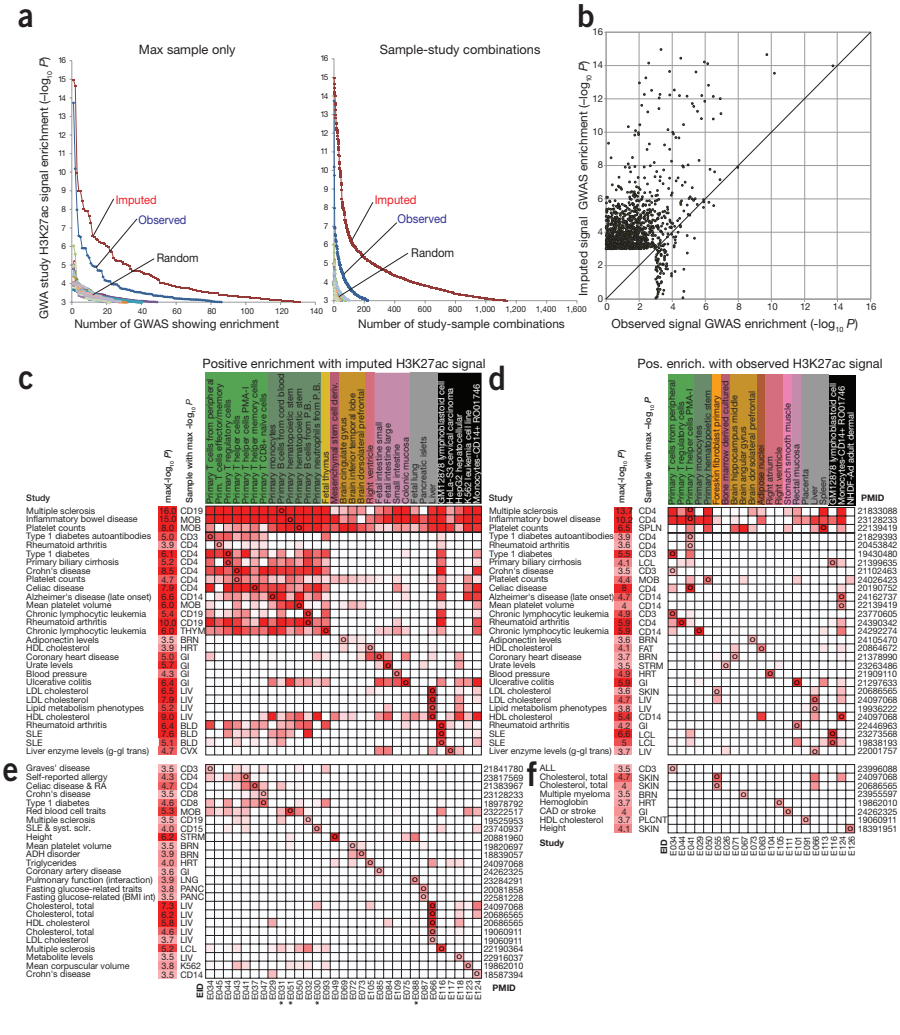

Figure 4 Overlap with trait-associated genetic variants from GWAS. (**a**) (Left) The *x* axis shows the number of GWAS in which there was at least one sample for which the H3K27ac signal was significantly enriched at significance level indicated on the *y* axis using a Mann-Whitney *U* Test. This is shown for the observed data (blue), the imputed data restricted to the 98 samples with observed data (red), and the observed and imputed data based on ten randomizations of the GWAS catalog. (Right) The same as on left, but counting study-sample combinations as opposed to just studies. (**b**) A scatter plot showing the $-\log_{10} P$ value computed for each study-sample combination based on the observed data (*x* axis) and imputed data (*y* axis) for each combination that had a *P* value of 10^{-3} or better based on either the imputed or the observed data for H3K27ac. The diagonal line is the $y = x$ line showing most of the most-significant enrichments based on either the observed or imputed data are for the imputed data. Additional marks can be found in **Supplementary Figures 24–26**. (**c–f**) Enrichment matrices (heatmaps) showing all studies (rows) with uncorrected $-\log_{10} P \geq 3.5$ and positive enrichment for at least one reference epigenome (columns) based on H3K27ac imputed data (**c**,**e**) and observed data (**d**,**f**). For each study (rows) is shown the trait, most-significant *P* value ($-\log_{10} P$), max-sample abbreviation and PubMed identifier (PMID). Only samples that showed the highest-significance positive enrichment for at least one study are shown. Studies in **c**,**d** were significant ($-\log_{10} P \geq 3.5$) for both observed and imputed data. Top three rows show studies with broad enrichment across samples. (**e**,**f**) Same enrichments for studies that were only significantly enriched using imputed (**e**) or observed (**f**) H3K27ac signal. Asterisks denote H3K27ac signal tracks that exist only as imputed data. Expanded enrichments for all samples, all tier 1 marks and additional GWAS are in **Supplementary Table 2**. SLE, systemic lupus erythematosus; ADH, attention deficit hyperactivity; ALL, acute lymphocytic leukemia; P.B., peripheral blood.

These results help validate the biological relevance of imputed datasets, based on an orthogonal annotation source, and help illustrate imputed datasets as a potentially useful resource for interpreting GWAS results.

Imputed datasets are informative for quality control

We next studied whether discrepancy between imputed and observed datasets is indicative of lower-quality experiments and can be used as a quality control (QC) metric. We ranked all H3K4me3 and H3K36me3 datasets based on PromRecov and GeneRecov scores, respectively, providing an independent benchmark informative of dataset quality (**Fig. 5a**). We then compared several QC metrics previously applied to these datasets[10], based on their ability to flag the worst-ranked datasets. These metrics are based on the proportion of reads falling in enriched regions as determined by various methods (signal proportion of tags (SPOT)[34], pre-binned regions enriched based on a Poisson distribution[10] and FindPeaks[35]) and signal correlations between forward and reverse reads (normalized strand correlation (NSC) and relative strand cross-correlation (RSC))[36].

Traditional QC metrics indeed flagged several worst-ranked H3K4me3 and H3K36me3 datasets, but failed to detect several cases, especially for lower read depths. This was more pronounced for H3K36me3, where two metrics (NSC, RSC) failed to detect the majority of low-GeneRecov datasets, and several datasets (E104, E022, E087, E109) were not detected as problematic by any of the traditional QC metrics. A deeper

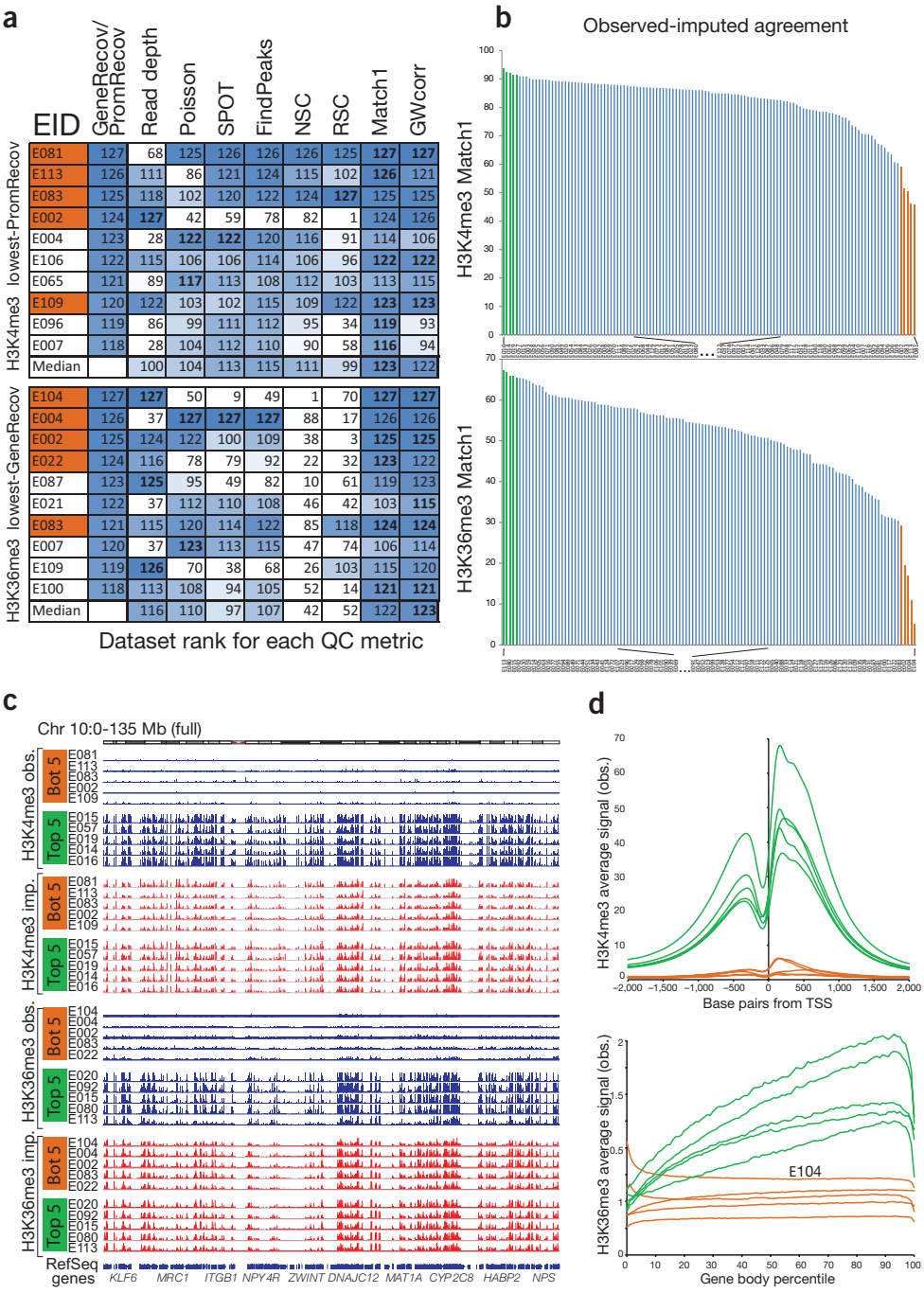

Figure 5 Low similarity between imputed and observed data reveals low-quality datasets. (**a**) Comparison of QC metrics (columns) for the ten datasets (rows) showing lowest agreement with gene and promoter annotations (**Fig. 3a**,**b**), based on H3K4me3 PromRecov (top) and H3K36me3 GeneRecov (bottom). Each entry shows rank (out of 127) for GeneRecov/PromRecov, read depth and each QC metric (Poisson statistic, Signal Proportion of Tags (SPOT), FindPeaks, Normalized and Relative Strand Correlation between forward and reverse strands (NSC and RSC)), and similarity between imputed and observed data (Match1 and GWcorr). Orange-shaded EIDs denote the five worst-agreement datasets from **b**. Data sets with the same read depth (a result of highly sequenced datasets being previously downsampled to the same number of reads[10]) are given the same expected rank if ties were broken randomly. Most-problematic datasets (based on lack of gene or ±2 kb TSS annotation recovery) are sometimes missed by traditional QC measures but consistently show low imputation agreement. (**b**) Distribution of agreement between top 1% observed signal and top 1% imputed signal locations for H3K4me3 (top) and H3K36me3 (bottom), highlighting five worst-similarity (orange) and five highest-similarity (green) datasets. (**c**) Observed (blue) and imputed (red) signal tracks for worst-similarity (orange) and best-similarity (green) datasets for H3K4me3 (top) and H3K36me3 (bottom) for the entire chromosome 10 (0–135 Mb). Datasets with the lowest agreement have a relatively flat signal, suggesting that when observed and imputed datasets disagree most, it is usually the observed datasets that are of lowest quality. (**d**) Aggregation of observed signal for H3K4me3 surrounding the TSS (top) and H3K36me3 in gene bodies (bottom) for the five best-agreement (green) and worst-agreement (orange) datasets, highlighting the unusual profiles of some worst-agreement datasets, suggesting they are of lower quality, even though they were not flagged by traditional QC metrics.

understanding of the sources of lower-quality datasets is beyond the scope of this paper, but the low read depth of several flagged datasets (**Fig. 5a** and **Supplementary Fig. 27**) suggests that deeper sequencing in some cases could improve overall quality.

By contrast, imputation-based QC metrics were consistently able to capture worst-ranked datasets, even when traditional QC metrics failed (**Fig. 5a**). We evaluated two imputation-based QC metrics, the first based on our Match1 score (overlap of the top 1% of imputed signal with observed signal) (**Supplementary Fig. 8**) and the second based on our GWcorr score (genome-wide correlation in signal between imputed and observed signal tracks). Both performed well, showing the best agreement with PromRecov and GeneRecov at detecting the worst datasets (**Fig. 5a**). Notably, the E104 Right Atrium H3K36me3 dataset (which both the GeneRecov and imputation metrics ranked as the worst H3K36me3 dataset and had the lowest sequencing coverage depth) was rated as the single highest-quality H3K36me3 dataset, based on the NSC metric, and was considered among the ten highest-quality H3K36me3 datasets by SPOT. The metagene plot of this sample shows inconsistencies with the typical pattern for H3K36me3 and is suggestive of possible antibody cross-reactivity (**Fig. 5d**), illustrating how QC measures based on agreement with imputed data can be used to identify likely problematic datasets that are missed by other QC measures, which are ineffective in cases of label swaps or antibody cross-reactivity.

Observed datasets varied substantially in their agreement with their corresponding imputed datasets (**Fig. 5b** and **Supplementary Table 3** and **Supplementary Fig. 28**). Moreover, the observed signal tracks for the worst-scoring samples (Match1 metric) showed striking visual differences from the best samples, whereas the corresponding imputed signal tracks had a consistently strong signal (**Fig. 5c**,**d**). When correlating QC metrics and read depth across all samples (**Supplementary Fig. 27**), the GWcorr and Match1 metrics showed among the highest correlations with both PromRecov and GeneRecov and were better correlated with sequencing depth for all histone marks, while being distinct from other QC metrics for all marks, highlighting that imputation-based QC measures capture important information, which is complementary, from existing QC metrics.

Imputed data identified unexpected signal regions

Although many high-quality experiments will globally agree with the imputed data, there could be specific locations for which the imputed data do not match the observed data. Because the imputed data constitute a form of prior expectation on the observed data, genomic locations where the two disagree can pinpoint biologically interesting locations and in some cases tissue-specific regulatory drivers.

To investigate this application of imputed datasets, we analyzed genomic locations showing strong DNA accessibility in observed data, but weak or no DNA accessibility in imputed data. Sequence motif analysis of these locations revealed an enrichment of biologically relevant regulatory motifs with known cell type–specific roles (**Supplementary Fig. 29**). For example NFKB motifs were found using primary monocyte DNA accessibility (E029) consistent with immune regulation, and PAX2 motifs in fetal kidney DNA accessibility (E086) consistent with roles in kidney development[37].

Thus, even for high-quality datasets, building a prior expectation of signal across the entire genome can also be informative for identifying locally dissimilar locations, which may be associated with cell type–specific and tissue-specific regulatory processes. However, if a mark that is highly correlated with the mark of interest is already present, then the imputation may already provide a close enough approximation to the true signal so that dissimilar locations may be due to biological or experimental noise, rather than cell type–specific regulation.

Imputation feature usage varies across marks

We next sought to gain information about the utilization of different marks and features for imputing datasets. We first studied the frequency with which each feature was utilized in our regression trees, at the root (**Supplementary Fig. 30a**) or at any position (**Supplementary Figs. 30b** and **31**) when it was available. We did this both for the primary imputation analyzed above, treating tier 1, tier 2 and tier 3 marks separately, given their differences in coverage, and another imputation restricted to the seven samples with deep coverage of many marks[9,10], treating all tier 1–3 marks uniformly, given their similar coverage.

For nearly all acetylation marks, the most frequent feature at the root was another acetylation mark at the same genomic position in the same sample, reflecting the highly correlated and dynamic nature of acetylation marks. For histone methylations, DNA accessibility, RNA-seq and DNA methylation, the most informative feature for the root was more often based on the same mark in a set of nearest K samples, consistent with their more stable nature across cell types.

When considering any position in the regression tree, the most frequently used features were from other marks in the same sample and the same position, although all positions surrounding the target genomic location were used quite often (**Supplementary Fig. 31**). DNA accessibility was less frequently used at the exact target position compared to histone mark features (**Supplementary Fig. 31**), reflecting the slight displacement of nucleosomes relative to open-chromatin regions, and thus the offset of histone modification marks relative to DNA accessibility peaks.

Chromatin state annotation using many imputed marks

Given the importance of chromatin mark combinations for distinguishing biologically meaningful features and different classes of regulatory elements, we used ChromHMM[20,21] to discover chromatin states based on imputed marks. Chromatin state analysis based on observed data in the Roadmap Epigenomics project primarily focused on the five marks common to all 127 samples (H3K4me1, H3K4me3, H3K36me3, H3K27me3 and H3K9me3) or only six marks (with H3K27ac) for 98 samples[10], with the number of samples rapidly decreasing as additional marks are considered due to missing datasets. ChromHMM explicitly handles missing data, but absence of a particular mark can result in dramatic reduction in the genomic coverage of corresponding chromatin states in the samples that are missing a defining mark (e.g., a DNA accessibility-dominated chromatin state shows 60-fold reduction for samples that lack DNA accessibility, **Supplementary Fig. 32**). Epigenomic mark imputation circumvents these limitations and provides a practical alternative to the missing-data strategy of ChromHMM, enabling learning of chromatin states jointly on uniform signal tracks for large numbers of epigenomic features across large numbers of samples.

We first trained a 25-state model jointly[3] across all 127 samples (**Fig. 6b**,**c**) using all tier 1 and 2 marks. This captured multiple types of promoter, enhancer, open chromatin, transcribed and repressed states and shows specific gene annotation, conservation, DNA methylation, and RNA-seq enrichments (**Fig. 6b**,**c** and **Supplementary Fig. 33**). Compared to the 15-state chromatin state model based on observed data in the 127 samples[10] (**Supplementary Fig. 33**), the 12-mark model better distinguished active versus poised enhancer states (using H3K27ac and H3K9ac) and captured novel states (e.g., state 19_DNase showing DNA accessibility but lacking enhancer/promoter marks and state 5_Tx5′ associated with 5′ ends of transcripts and based on H3K79me2). Because of the increased stability and robustness of imputed data, imputation-based chromatin states showed more consistent genome coverage across samples (**Supplementary Fig. 34**),

better agreement with annotated gene bodies and TSS, both for all genes (**Supplementary Fig. 35a**,**b**) and for a set of genes expressed in a given tissue (**Supplementary Fig. 35c**,**d**), and better discrimination of evolutionarily conserved elements (**Supplementary Fig. 36**)[38]. Additionally, we saw better recovery of a sample that was not included in any of our training data (an osteoblast DNA accessibility dataset[39],

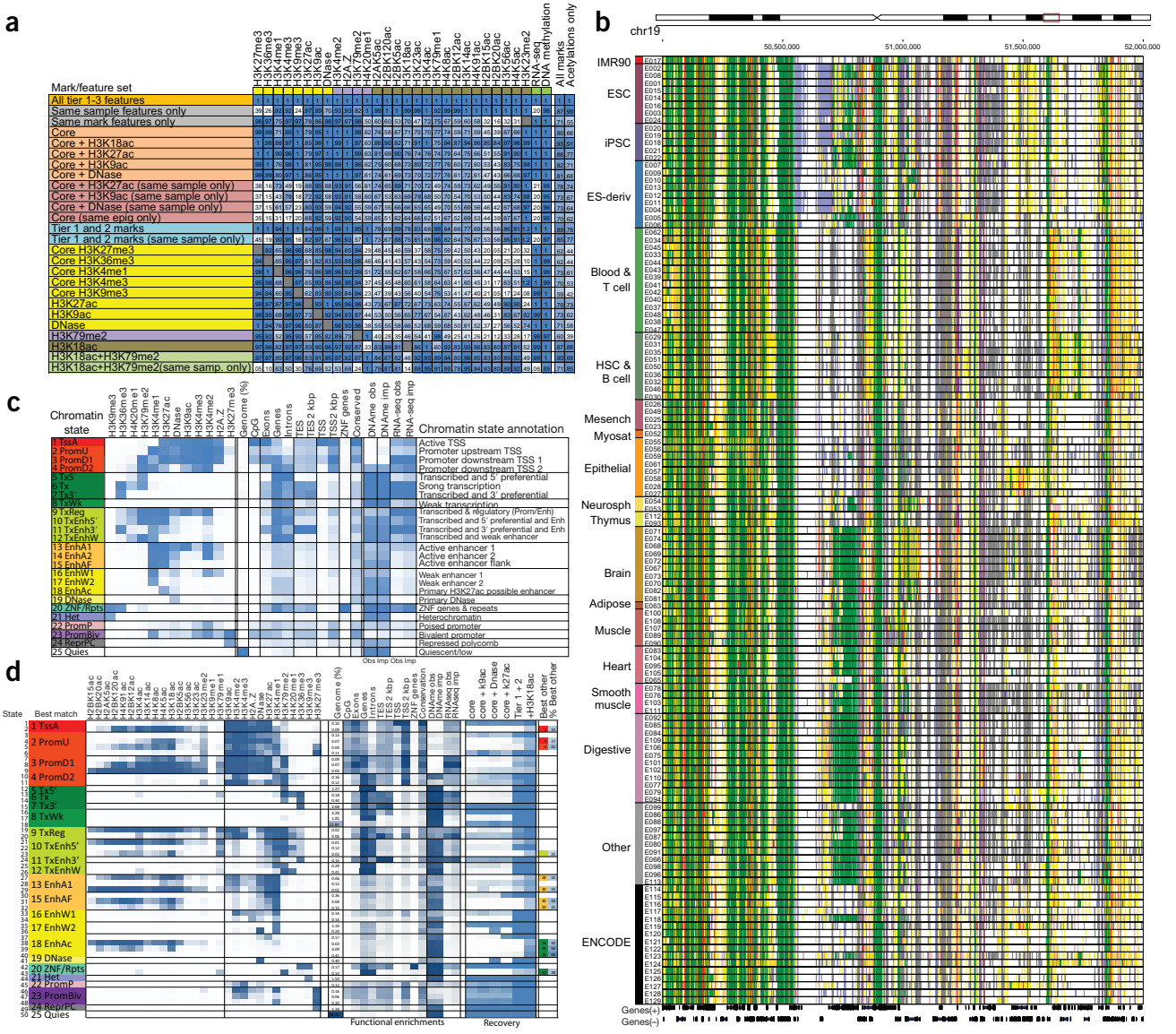

Figure 6 Imputation using mark subsets and chromatin state learning. (**a**) Imputation agreement for each mark (columns) using subsets of features (rows) in top 1% signal bins or 0.25 concordance measure for DNA methylation, for Chr10 relative to agreement achieved when using all features based on the seven samples with deep mark coverage without making distinctions between the tier 1–3 marks. Same-sample features are most important for acetylation marks, and same-mark features are most important for H3K27me3, H3K36me3, H3K9me3 and RNA-seq. Profiling of only H3K18ac and H3K79me2 allows higher relative imputation agreement than all five core marks, assuming a compendium with uniform coverage of marks. Performance for additional subsets is shown in **Supplementary Figure 42**. The last two columns show the average performance of the feature subset over all target marks and specifically for acetylations. Core=H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3. For the purpose of computing these averages for mark subsets, if the target mark was included in the subset then a value of 1 was used for the target mark; the imputation performance restricted to other marks in the subset, when available, is provided in the table. The H3K18ac+H3K79me2 and tier 1 and 2 mark evaluations were limited to the five samples that were deeply profiled across marks and also had experimentally profiled H3K79me2. (**b**) Portion of a chromatin state segmentation using imputed data of 12 marks across 127 samples using the 25-state model and colors shown in **c**. Segmentation is highly consistent for similar samples but is able to capture highly dynamic regulatory elements across different samples. (**c**) Chromatin state model using 12 marks and 25 states, trained jointly using imputed data across all 127 samples. For each state (rows) are shown its emission parameters, genome coverage, relative functional enrichments for diverse annotations and conserved elements, and median observed and imputed DNA methylation and RNA-seq signal (**Supplementary Fig. 33**), followed by a candidate state annotation. (**d**) Expanded chromatin state model trained using 50 states and 29 marks in seven samples with deep mark coverage. States are grouped and labeled by the maximum-enrichment 25-state model match. Additional marks in this model are shown to the left of the vertical line. Emission parameters and functional enrichments (similar to **c**), and percentage of locations recovered for each state using subsets of marks (**Supplementary Figs. 40**, **41** and **43**). '+H3K18ac' denotes the subset of tier 1 and 2 marks extended by H3K18ac. When the same chromatin state was not maximally recovered with tier 1 and 2 marks, the last two columns denote the best other state and its percent assignment.

Supplementary Fig. 37) including for sample-specific sites; in addition we captured major sample type differences in chromatin states (e.g., ESC/iPSC samples showed consistently more abundant bivalent promoter states[40], **Supplementary Fig. 38**), with differences in some cases more pronounced than for chromatin states based on observed data (**Supplementary Fig. 38**).

We also trained a 50-state model using imputed data for 29 marks across the seven deeply covered samples. The model showed distinct state emission parameters, diverse functional enrichments, and relatively consistent correlations in emission parameters and mark frequency across samples for nearly all states (**Fig. 6d** and **Supplementary Figs. 39–41**).

Accurate imputation using a limited number of marks

To help prioritize marks for experimental profiling in new cell types, we studied the subset of marks that provide the highest-accuracy imputation. We considered two settings, the first ('unrelated setting') assuming that new samples are largely dissimilar to any existing in the compendium and can rely only on same-sample features, and the second ('related setting') assuming that new samples are related to an existing compendium of datasets with roughly uniform coverage of each mark that can be used to impute in the new sample.

In both settings, we assessed the predictive power of a subset of features by comparing the agreement achieved between the observed signal and the imputed signal using the subset of features, relative to the agreement achieved using all features. We chose this 'relative agreement' metric to avoid penalizing the prediction of marks that are hard to impute even when using all features due to low-quality signal. We evaluated this relative agreement using the Match1 metric (except for DNA methylation, where we used Methyl25 in place), and using the coefficient of determination (R^2). We restricted these evaluations to the seven deep-coverage samples on chr10 and did not make distinctions between the tier 1–3 marks when performing the imputation (**Supplementary Fig. 8**).

In the 'unrelated' setting (same-sample features only), imputation of H3K36me3, H3K9me3, H3K27me3 and RNA-seq showed the lowest relative Match1 scores (20–39%) (**Fig. 6a** and **Supplementary Fig. 42a**), followed by DNA accessibility (70%), H3K79me2 (82%), and H3K4me1/2/3, H2A.Z and H3K79me1 (92–93%), suggesting a prioritization based on the marks that are hardest to impute using same-sample features, even if all other marks are used. All acetylation marks showed higher relative Match1 scores (97–100%), but H3K27ac had the lowest relative score among them (97%), suggesting it contains the most unique information. Relative Match1 score recovery was 87%, on average, across all marks when using all same-sample features, 70% when using only the five core marks (counting experimentally mapped marks as 100% recovered), 73% using the core marks and either DNA accessibility or H3K9ac, 78% using the core marks and H3K27ac, and 85% using all tier 1 and 2 marks (**Fig. 6a** and **Supplementary Fig. 42a**). R^2 values showed overall similar results and conclusions, but revealed a lower relative agreement for DNA methylation (**Supplementary Fig. 42b**), also highlighting its unique information relative to other marks in the same sample.

In the 'related' setting (both same-sample and same-mark features), the five core marks resulted in 80% Match1 relative recovery on average across all marks, which increased, respectively, to 86%, 82% and 81% with inclusion of H3K27ac, H3K9ac or DNA accessibility, and increased to 89% using all tier 1 and 2 marks (**Fig. 6a**). Recovery of acetylation marks was on average lower (66%) using only the five core marks, but increased to 77%, 71% and 68%, respectively, with inclusion of H3K27ac, H3K9ac or DNA accessibility. Using one or two marks led to sometimes surprisingly high recovery of many other marks. For example, H3K18ac was the single mark giving the highest average recovery of all others marks (87%; 88% for acetylation marks), and greater than 80% recovery for all marks except H4K20me1, H3K79me1 and H3K23me2. Profiling of H3K79me2 was highly complementary, resulting in 98% recovery for H4K20me1 and H3K79me1; and profiling of H3K79me2 in combination with H3K18ac resulted in 90% average recovery of marks in a new cell type, when leveraging the entire existing data compendium, but only 71% average recovery using same-sample features.

We also used chromatin states to evaluate the 'unrelated' setting, based on the ability of subsets of the 29 marks to recover each of the 50 chromatin states learned from imputed data in the seven deeply covered samples when treating the remaining marks as missing[20] (**Fig. 6d** and **Supplementary Fig. 43**). We found that holding out any of DNA accessibility, H3K9me3, H3K36me3, H3K4me1, H3K27me3 or H3K27ac resulted in at least one 'missing' state (<20% recovery; **Supplementary Fig. 43a**). Holding out H2A.Z, H3K79me2, H4K20me1, H3K79me1, H3K4me3 or H3K4me2 resulted in at least one state with less than 70% recovery. No single mark in isolation led to substantial state recovery beyond the states that were primarily defined by that mark (**Supplementary Fig. 43d**). Using only the five core marks and treating all remaining marks as missing data resulted in 31% average recovery of assigned locations for each state (**Fig 6d** and **Supplementary Fig. 43c**). Including H3K27ac, H3K9ac or DNA accessibility increased average recovery to only 35–37%, and the greatest average state recovery of any mark was 43% with the additional inclusion of H3K18ac. Using all tier 1 and 2 marks together increased the average recovery to 65%, with only 12 states showing 30% or less recovery (**Fig. 6d** and **Supplementary Fig. 43b**). Inclusion of H3K18ac with the tier 1 and tier 2 marks increased average state recovery to 77%, with all states showing greater than 30% recovery. These results suggest substantial additional diversity of chromatin states not captured based on the chromatin marks that have received extensive mapping by the Roadmap Epigenomics and ENCODE projects.

DISCUSSION

In this paper we introduced a computational approach for prediction (imputation) of genome-wide epigenomic signals applied at 25-bp resolution. The method imputes both missing and existing datasets by leveraging correlations of epigenomic marks within a given sample and similarities in the epigenomic landscape of related samples, and it is applicable to any type of functional data that can be represented as a signal track. We developed and applied an array of quantitative metrics and tests to evaluate the accuracy of the imputed data. We showed that the imputed data of a mark in a sample is of high resolution and a better match to the observed data than using the average of all other observed datasets of that mark (an important baseline comparison for any such study), and it is also a better match than even the single closest dataset (a benchmark that would require knowledge of the target mark and is thus not possible in practice).

We showed that imputed data outperformed observed data based on a number of analyses: (i) similarity to annotated gene features; (ii) consistency across closely related samples; (iii) capture of biological relationships between tissue and cell types; (iv) correlation with observed gene expression; (v) enrichment of SNPs identified in GWAS; (vi) chromatin state capture of TSS, gene bodies, tissue-restricted activity and conserved elements. The observed data only showed a modest advantage in identifying genes showing the most tissue-specific expression patterns (approximately 3% of genes in each sample). Furthermore, disagreement between observed and imputed data were usually due to lower-quality experimental datasets, and not low-quality imputation.

Our benchmarks show that in practice, observed data are not always an uncontested gold standard, but that both observed and imputed data are of important and complementary value, each with its own merits, and each likely to have both false-negative and false-positive signals. Certainly, when high-quality, deeply sequenced and extensively replicated experiments are available, they remain a gold standard. However, with the reality of budgetary and sample limitations, our work establishes imputed data as an important complement to experimental studies. For any fixed number of budgeted experiments, imputation allows projects to explore a larger diversity of samples, assays or conditions, and to increase robustness by leveraging automatically learned correlations in these datasets, rather than relying solely on direct experimental profiling and replicates to increase robustness.

Moreover, the combined use of observed and imputed data opens many new applications that were previously not possible. Imputed data can be used as a prior expectation for an experiment, against which observed data can be compared and benchmarked. We demonstrated two applications of such comparisons, using global discrepancies between observed and imputed data as a QC metric, and identifying surprising locations that we found enriched for regulator targets. For QC in particular, we showed that low agreement between imputed and observed data revealed problematic datasets that were missed by many existing metrics that focus on signal-to-noise properties of the data, and thus can miss sample mix-ups, cross-reacting antibodies or other experimental errors. With more densely sampled epigenomic datasets, we expect that next-generation QC metrics will increasingly exploit imputation-like measures, such as our stringent baselines defined earlier or the more sophisticated agreement with ChromImpute.

Our work also has implications for experiment prioritization for large-scale epigenomic mapping efforts. The Roadmap Epigenomics project mapped a set of six histone marks at highest depth: H3K4me1, H3K4me3, H3K27me3, H3K9me3, H3K36me3 and H3K27ac. Our results validate this strategy, as H3K27me3, H3K9me3 and H3K36me3 could not be imputed effectively using same-sample data even if every other mark in the same sample was mapped, and H3K4me1, H3K4me3 and H3K27ac all had substantial unique information that could not be predicted from just using same-sample features of the other five marks. Our results support possibly extending this set with H3K18ac, which led to better imputation of non-H3K27ac acetylations and with H3K79me2, which led to better capture of transcription-associated marks. The evidence shows both marks are important in their own right, H3K18ac in pathogen response[41] and cancer[42–45], and H3K79me2 in epigenetic memory[46], development and cancer[47].

It is also important to recognize limitations of the imputation approach. If the presence of mark signal is highly specific to one or a few samples, and it does not correlate with other marks mapped in the sample or has a different correlation structure than in samples used for training, then it would not be possible to accurately impute the mark at those locations. When the target mark has been mapped in only a few samples, the features pertaining to the same mark in other samples may be less informative or more biased. For example, imputation of transcription factor binding may be more challenging, as their correlation structure with other marks can vary greatly across samples, depending on whether a transcription factor is active or not, and most have been mapped in only a limited number of samples. A limitation of our current framework when imputing datasets across individuals is that we do not currently incorporate genetic variation as an input, and this is potentially an important area of future development given that datasets on chromatin marks and genotype across individuals are becoming increasingly available[48–50]. For tissue samples that reflect mixtures of multiple cell types, our imputed maps will most likely reflect the same mixture as the observed data, though deconvolution of mixed samples is a potentially important direction for future work.

Lastly, our paper contributes, to our knowledge, the most comprehensive epigenomic resource to date, including 4,315 imputed datasets across 127 samples and 34 marks (of which only 26% have been experimentally profiled). The remaining 74% (3,193 datasets) exist only as imputed data, dramatically expanding the number, diversity and completeness of even the most complete existing set of epigenomic maps. We also provide an annotation of 25 chromatin states based on 12 imputed marks across 127 samples, and of 50 chromatin states based on 29 epigenomic marks across 7 samples, providing the most comprehensive collection of regulatory annotations across the human genome to date. As our initial analyses demonstrate, the resulting annotation of the noncoding portion of the human genome can increase the power of future studies of gene regulation, cellular differentiation, genetic variation and human disease.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. All imputed signal datasets and peak calls and chromatin states based on imputed data are available from http://compbio.mit.edu/roadmap/. The ChromImpute software is available at http://www.biolchem.ucla.edu/labs/ernst/ChromImpute/ and source code is provided as **Supplementary File 1** and maintained at https://github.com/jernst98/ChromImpute.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank A. Kundaje, W. Meuleman, M. Bilenky and members of the NIH Roadmap Epigenomics Data Analysis and Coordination Center for data processing. We thank P. Kheradpour for advice on the motif analysis. We thank M. Eaton for generating the chromatin state segmentation visualization. We thank T. Wang for hosting the imputed data on the WashU Epigenome Browser. We thank N. Rajagopal, B. Ren and members of the Kellis laboratory and Roadmap Epigenomics Consortium for discussions related to this work. We thank the NIH Roadmap Epigenomics and ENCODE consortia for generating the data used in this paper. Funding for this work provided by National Science Foundation CAREER Award #1254200 and an Alfred P. Sloan Fellowship to J.E. and by US National Institutes of Health through National Human Genome Research Institute grants RC1HG005334 and R01HG004037 to M.K.

AUTHOR CONTRIBUTIONS

J.E. and M.K. developed the method, analyzed the results and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
2. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
3. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
4. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
5. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).

6. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
7. Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
8. Ziller, M.J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).
9. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
10. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 human reference epigenomes. *Nature* doi:10.1038/nature14248 (18 February 2015).
11. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
12. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
13. Bock, C. *et al.* CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet.* **2**, e26 (2006).
14. Das, R. *et al.* Computational prediction of methylation status in human genomic sequences. *Proc. Natl. Acad. Sci. USA* **103**, 10713–10716 (2006).
15. Yuan, G.-C. Targeted recruitment of histone modifications in humans predicted by genomic sequences. *J. Comput. Biol.* **16**, 341–355 (2009).
16. Fan, S., Zhang, M.Q. & Zhang, X. Histone methylation marks play important roles in predicting the methylation status of CpG islands. *Biochem. Biophys. Res. Commun.* **374**, 559–564 (2008).
17. Zheng, H., Wu, H., Li, J. & Jiang, S.-W. CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome. *BMC Med. Genomics* **6**, S13 (2013).
18. Stevens, M. *et al.* Estimating absolute methylation levels at single CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res.* **23**,1541–1553 (2013).
19. Capra, J.A. & Kostka, D. Modeling DNA methylation dynamics with approaches from phylogenetics. *Bioinformatics* **30**, i408–i414 (2014).
20. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).
21. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
22. Hoffman, M.M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).
23. Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. USA* **107**, 2926–2931 (2010).
24. Lasserre, J., Chung, H.-R. & Vingron, M. Finding associations among histone modifications using sparse partial correlation networks. *PLoS Comput. Biol.* **9**, e1003168 (2013).
25. Yu, H., Zhu, S., Zhou, B., Xue, H. & Han, J.-D.J. Inferring causal relationships among different histone modifications and gene expression. *Genome Res.* **18**, 1314–1324 (2008).
26. Zhou, J. & Troyanskaya, O.G. Global quantitative modeling of chromatin factor interactions. *PLoS Comput. Biol.* **10**, e1003525 (2014).
27. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, 2009).
28. Zhou, X. *et al.* The Human Epigenome Browser at Washington University. *Nat. Methods* **8**, 989–990 (2011).
29. Raney, B.J. *et al.* Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**, 1003–1005 (2014).
30. Harris, R.A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.* **28**, 1097–1105 (2010).
31. Bolstad, B.M., Irizarry, R.A., Åstrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
32. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
33. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
34. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).
35. Fejes, A.P. *et al.* FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**, 1729–1730 (2008).
36. Landt, S.G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
37. Sanyanusin, P. *et al.* Mutation of the PAX2 gene in a family with optic nerve colobomas, renal anomalies and vesicoureteral reflux. *Nat. Genet.* **9**, 358–364 (1995).
38. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
39. Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* **21**, 1757–1767 (2011).
40. Bernstein, B.E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
41. Eskandarian, H.A. *et al.* A role for SIRT2-dependent histone H3K18 deacetylation in bacterial infection. *Science* **341**, 1238858 (2013).
42. Barber, M.F. *et al.* SIRT7 links H3K18 deacetylation to maintenance of oncogenic transformation. *Nature* **487**, 114–118 (2012).
43. Ferrari, R. *et al.* Epigenetic reprogramming by adenovirus e1a. *Science* **321**, 1086–1088 (2008).
44. Horwitz, G.A. *et al.* Adenovirus small e1a alters global patterns of histone modification. *Science* **321**, 1084–1085 (2008).
45. Seligson, D.B. *et al.* Global histone modification patterns predict risk of prostate cancer recurrence. *Nature* **435**, 1262–1266 (2005).
46. Kouskouti, A. & Talianidis, I. Histone modifications defining active genes persist after transcriptional and mitotic inactivation. *EMBO J.* **24**, 347–357 (2005).
47. Nguyen, A.T. & Zhang, Y. The diverse functions of Dot1 and H3K79 methylation. *Genes Dev.* **25**, 1345–1358 (2011).
48. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).
49. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747–749 (2013).
50. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744–747 (2013).

ONLINE METHODS

Signal tracks. For the histone mark and DNase signal tracks we used the version of the reference epigenomes signal tracks based on the $-\log_{10} P$ value of enrichment relative to input control based on a Poisson distribution from (Roadmap Epigenomics Consortium *et al.*, 2015)[10], available through http://compbio.mit.edu/roadmap/. Some of these reference epigenomes are based on multiple biological samples that were pooled, but we refer to each reference epigenome as a 'sample' here. We only used the signal for chromosomes 1-22 and X. For the RNA-seq data we converted the uniformly processed unstranded signal tracks, also available from the same site, to normalized RPKM values, then added one, and then took the log base 2 value. The normalized RPKM values were computed based on multiplying the unnormalized signal value by 10^9 then dividing by the product of the read length and the number of exonic reads, excluding the mitochondria, ribosome and the top 0.5% of signal values[10]. We converted these signal tracks for the histone marks, DNase and RNA-seq data to a 25-bp resolution by taking the base level average of signal overlapping each 25-bp bin. For the DNA methylation we used the uniformly processed whole genome bisulfite data[10], which provided a fraction methylated value at each base within all CpGs that had more than three reads covering it. We filled in missing values for bases within CpGs by replacing them with the genome average for DNA methylation when training and the chromosome average when applying the predictors as this step was done on each chromosome independently.

We selected the $-\log_{10} P$ value signal tracks rather than the fold-change tracks for histone marks and DNase as they were designated the primary signal tracks for analyses in (Roadmap Epigenomics Consortium *et al.*, 2015)[10] on the basis of having better signal-to-noise properties. In particular, both sets of tracks were generated based on downsampling highly sequenced datasets to the same sequencing depth, thus in the $-\log_{10} P$ value track, no dataset had a disproportionately high signal simply due to being sequenced very deeply, whereas on the other hand under-sequenced datasets were included and in some cases had locations with high fold-change signals that were the result of noise and did not have values as relatively high on the $-\log_{10} P$-value track. Additionally focusing on the $-\log_{10} P$-value tracks is more consistent with the basis of the default binarization of ChromHMM[21] used for the chromatin state learning.

ChromImpute method. The ChromImpute method predicts the signal of a target mark in a target sample based on two classes of features: (i) other marks mapped in the same sample and (ii) the target mark in other samples. Predictors that integrate these features are trained based on each sample for which we have the target mark available, excluding the target sample. The ensemble of trained predictors are then each applied in the target sample and their predictions are averaged to obtain the final predictions. The ensemble approach would be expected to tend to average out biases associated with any one predictor.

Formally, let $o_{c,m,p}$ represent the observed value of mark m in sample c at position p. Let $M_{c,m}$ denote the set of marks in sample c among those eligible to be used to predict mark m. Let C_m denote the set of samples in which mark m has been mapped. Let m^t denote the target mark and c^t the target sample. To predict mark m^t in sample c^t for each sample $c^{t'} \in C_{m^t} \backslash \{c^t\}$, we separately define features. For a sample $c^{t'}$ we let M_I denote $M_{c^t,m^t} \cap M_{c^{t'},m^t} \backslash \{m^t\}$, which is the subset of common marks between c^t and $c^{t'}$ that can be used to predict the target mark m^t, and then define the two classes of features to predict the signal of mark m^t in sample $c^{t'}$ at a target genomic position p.

1. Features based on the set of other marks mapped in the same sample. We define features $s_{m,n}$ for each mark $m \in M_I$ and each value of n such that $n = 500i$ or $n = 25i$ for integer values of $i = -20,\ldots,20$. The feature $s_{m,n}$ is assigned a value $o_{c^{t'},m,p+n}$. In our notation $p+n$ refers to a position on the same chromosome as p, but a base position shifted by n. This corresponds to having features at the target position and every 25 bp within 500 bp, and every 500 bp within 10,000 bp both upstream and downstream of the target position.
2. Features based on the target mark in other samples. We define features $f_{m,g,k}$ for each mark $m \in M_I$, $g \in \{local, global\}$, and $k = 1,\ldots,\min(10,|C_I|)$ where we define C_I to be $C_{m^t} \cap C_m \backslash \{c^{t'}, c^t\}$. C_I corresponds to all samples having the target mark and the mark that will be used for determining similar samples excluding the overall target sample and the sample being used for training the predictor. $f_{m,g,k}$ has the value $\frac{1}{k}\sum_{j=1}^{k} o_{c_j,m,p}$ where c_j is the sample of C_I that is in the ranked position j when each sample $c \in C_I$ is ordered in increasing value of $d_{m,g}(c^{t'}, c)$. If $g = global$, then $d_{m,g}(c^{t'}, c) = 1 - \rho(o_{c^{t'},m}, o_{c,m})$ where ρ is the Pearson correlation coefficient applied to the genome-wide signal of mark m in samples $c^{t'}$ and c. If $g = local$, then at the position p

$$d_{m,g}(c^{t'},c) = \sum_{i=-20}^{20} (o_{c^{t'},m,p+25i} - o_{c,m,p+25i})^2$$

which uses the signal at target position and every 25-bp interval within 500 bp to determine the nearest samples. Ties for the nearest sample based on local distance were broken arbitrarily.

We construct feature vectors by combining all the $s_{m,n}$ and $f_{m,g,k}$ features defined above. Features when applying a predictor in sample c^t trained on sample $c^{t'}$ are defined as above except $c^{t'}$ is interchanged with c^t.

The specific predictors we used were regression trees[27]. Formally we define a regression tree, T, to have a set of split nodes S and a set of leaf nodes N. A split node $s \in S$ can be represented by the 4-tuple (f, v, l, r) where f is a feature used to the split the data, v is the value of feature f on which the split is based, and l and r are nodes in $S \cup N$. A leaf node $n \in N$ can be represented by a 1-tuple (e) which is the prediction value associated with the node. In addition one node $w \in S \cup N$ is designated as the root of the tree. We let u denote a vector of feature values for which an output prediction should be generated. To generate a prediction we start by setting a variable z to the root node w, and then while z is not a leaf node, if $u.(z.f) \leq z.v$ we let $z = z.l$ and otherwise $z = z.r$ where $u.x$ refers to feature x of vector u. Once z is a leaf node the prediction of $z.e$ is made.

We train regression trees for a mark m^t based on sample $c^{t'}$ for a set of sampled positions P recursively. We define a node creation procedure that takes as input a set X of positions and identifies a feature, f, and split value, v, on which to split the positions. In the procedure we define the sets

$$X_{L_{f,v}} = \left\{p \in X \middle| u_{c^{t'},m^t,p}.f \leq v\right\} \text{ and } X_{R_{f,v}} = \left\{p \in X \middle| u_{c^{t'},m^t,p}.f > v\right\}$$

where $u_{c^{t'},m^t,p}.f$ corresponds to the feature value f of the feature vector for position p as defined above when considering m^t based on sample $c^{t'}$. If the set $\{(f,v) | \, |X_{L_{f,v}}| \geq 20 \wedge |X_{R_{f,v}}| \geq 20\}$ is empty, meaning there is no split that can be created with both subsets of the partition containing at least 20 data points, a constraint intended to reduce overfitting, then we create a leaf node n where the associated output prediction of the node $n.e$ is set to $\frac{1}{|X|}\sum_{p \in X} o_{c^{t'},m^t,p}$, that is, the average value of mark m^t in sample $c^{t'}$ at all positions in X; otherwise, we create a split node s and set $s.f$ and $s.v$ to f and v, respectively, based on

$$\underset{\{(f,v) | |X_{L_{f,v}}| \geq 20 \wedge |X_{R_{f,v}}| \geq 20\}}{\operatorname{argmin}} \left(\sum_{p \in X_{L_{f,v}}} \left(o_{c^{t'},m^t,p} - \frac{1}{|X_{L_{f,v}}|} \sum_{p' \in X_{L_{f,v}}} o_{c^{t'},m^t,p'}\right)^2 + \sum_{p \in X_{R_{f,v}}} \left(o_{c^{t'},m^t,p} - \frac{1}{|X_{R_{f,v}}|} \sum_{p' \in X_{R_{f,v}}} o_{c^{t'},m^t,p'}\right)^2 \right)$$

This chooses a split that minimizes the squared error of the resulting output prediction subject to the constraint that both subsets of the partition have at least 20 data points. We then set $s.l$ and $s.r$ to the nodes created by applying the node creation procedure to the set of positions $X_{L_{f,v}}$ and $X_{R_{f,v}}$, respectively. Ties for the best split feature and value were broken randomly. Input data were rounded to the nearest tenth, for generating features, training and applying the predictors, and only those values present in the training data were considered as split values. DNA methylation values were treated as percentages for the purposes of this rounding, but the final output for DNA methylation was reported as a fraction. The node creation procedure is initially called with all positions in P, which creates the root node.

To make a prediction in sample c^t for mark m^t at position p we compute

$$\frac{1}{b|C_{m_t} \backslash \{c_t\}|} \sum_{c_{t'} \in C_{m_t} \backslash \{c_t\}} \sum_{i=1}^{b} T_{c_{t'},m_t,P_i}(u_{c_t,m_t,p})$$

where b is number of sets of sampled positions and $T_{c_{t'},m_t,P_i}(u_{c_t,m_t,p})$ denotes the prediction made by the regression tree trained on sample $c^{t'}$ to predict mark m^t using the set of sampled positions P_i when applied to the feature vector defined as above for predicting mark m^t in sample c^t at position p.

Each set of positions for training contained 100,000 randomly sampled positions. We used one set of positions for training, with two exceptions. We trained predictors for the tier 3 marks in the primary imputation and for all marks in the imputation restricted to the seven samples with deep coverage of many marks

(E003, E004, E005, E006, E007, E008, E017)[10] on the basis of three independent 100,000 sampled positions, as we had a limited number of different samples on which to train predictors. If the set of features that could be defined for a target sample in training is empty, which happened during evaluation of predictive performance when holding out some features, we excluded that predictor from the ensemble.

All predictions except for DNA methylation were at a 25-bp resolution. For DNA methylation we made base predictions just at the positions of CpGs, but the features based on other marks were still computed at a 25-bp resolution. We did not make explicit predictions for positions within the first and last 10 kb of each chromosome, and instead 0 was used as the signal value there except for DNA methylation where it was 0.5.

For the primary imputation the tier assignments of marks determined which marks were eligible to be used to impute other marks (**Supplementary Fig. 2**), and we made predictions across chr1-22 and chrX. For the purpose of evaluating imputation performance with subsets of features and marks unbiased by the deep sample coverage of certain marks, we did a separate set of imputations using only the seven samples with deep mark coverage. For this set of imputations we treated the tier 1–3 marks in the same way, and the method could use any of the available marks within these tiers to predict any other mark. For these evaluations we made predictions only on chr10.

In order to handle the computational demands of training an ensemble of predictors and then applying them to generate genome-wide predictions for more than 4,000 datasets we first wrote out to disk for the randomly sampled positions feature instances for each observed mark and sample. The set of feature instances for a mark and sample written out were sufficient to be used to train predictors based on the sample for the goal of predicting the mark in any other sample. Depending on the overall target sample, different subsets of the features would be used, consistent with what is described above, but this step allowed significant reuse of computation and memory when imputing the same mark across multiple samples. Once the training instances were written out, different predictors could be trained in parallel. Applying the predictors to impute genome-wide values was parallelized over different samples, marks and chromosomes. To more efficiently compute the ordering of the locally nearest samples at each position when making genome-wide predictions, a computationally demanding step, we leveraged information on the ordering of the nearest samples at the previously considered position, which would often be highly similar.

Comparison with linear regression, nearest neighbor and single sample training predictions. For the linear regression and nearest-neighbor comparison, we limited the predictions to chr10. The linear regression was the weka (v.3.7.3)[51] implementation with a ridge regularization parameter set to 1. For the comparison with nearest-neighbor approaches we used up to the ten nearest neighbors defined by H3K4me1 and for both the local and global distance as defined above. We selected H3K4me1 as it was defined in all samples and associated with more sample-specific patterns[3,4]. For predicting H3K4me1 we used H3K4me3 instead. Similarly for the comparison with training based on a single nearest sample, we selected the nearest sample based on global H3K4me1 correlation, except using H3K4me3 when predicting H3K4me1.

Gene annotations, expression, conserved elements. For gene annotation enrichments we used a modified version of the GENCODE 10 gene annotations[52] that only included long transcripts as used in (Roadmap Epigenomics Consortium *et al.*, 2015)[10]. For defining a set of expressed genes in each sample we combined the protein coding genes and noncoding RNA sets selecting those genes that had an RPKM ≥ 0.5 as processed in (Roadmap Epigenomics Consortium *et al.*, 2015)[10]. The evolutionarily conserved elements were the hg19 liftover of the SiPhy-pi conserved elements previously reported[38,53].

Signal heatmap clustering. The signal heatmaps were generated by first randomly selecting 2,000 25-bp intervals in the genome, which form one dimension of each matrix. The other dimension corresponds to different samples in which the mark was observed. The ordering of elements in both dimensions of the matrix were determined using the Matlab implementation of hierarchical clustering and optimal leaf ordering[54] applied to the observed data. Correlation distance was used except to cluster the rows for DNA methylation, H3K23me3, H4K5ac and RNA-seq where Euclidean distance was used because of zero variance rows. The imputed data matrix is based on using the same ordering of rows and columns as generated based on the observed data.

Chromatin states based on imputed data. Chromatin states were inferred on the imputed data using ChromHMM[21]. The data were binarized at a 200-bp resolution by averaging the eight 25-bp intervals overlapping and using an average signal threshold of 2. Two types of models were inferred. One model used the 12 tier-1 and 2 marks across all 127 samples. The second model was based on all tier 1–3 marks imputed in all the seven samples with deep mark coverage, where we had a more confident imputation of the tier 3 marks. Both posterior probabilities soft-assignments for each state and hard assignments based on the maximum posterior were produced, but all the chromatin state analyses were based on the hard assignments. Chromatin states based on the observed data were obtained from (Roadmap Epigenomics Consortium *et al.*, 2015)[10].

The chromatin state assignment recovery based on the maps of a subset of marks was determined using the *EvalSubset* command of ChromHMM[21]. This is similar to a procedure previously described[20], but based on hard assignments.

Single mark peak calls. Macs2 (version 2.0.10)[55] was used to call peaks on the imputed signal data. The bdgpeakcall command was used to generate narrowPeaks whereas the bdgbroadcall command was used to generate gappedPeaks with the '-c' cutoff flag was set to 2. These peak calls were compared to corresponding peak calls based on the observed data obtained from (Roadmap Epigenomics Consortium *et al.*, 2015)[10] that were also generated based on Macs2 but based on the callpeak command applied to aligned reads.

Comparison with GWAS analysis. We obtained the contents of the NHGRI GWAS Catalog[33] on September 12, 2014 through the UCSC Genome Browser[56]. We grouped entries into studies based on a unique combination of PubMed ID and trait combination. We filtered the set of SNPs in each study such that no two SNPs were within 1 Mb of each other on the same chromosome. We did this by ranking the SNPs in a study based on their *P* value significance, and then filtering a SNP if it was within 1 Mb of any higher ranked SNP that was not filtered. We tested the significance of the signal level for observed and separately imputed data associated with a set of SNPs in a study compared to all other GWAS catalog SNPs after the filtering using a Mann-Whitney *U* Test as implemented in the Apache Commons Math 3.3 library. For each mark and separately for the observed and imputed data, we computed estimated false discovery rates (FDRs) at each *P* value threshold controlling for testing multiple study and sample combinations. We did this by generating 100 random permutations of the study assignments among the set of filtered SNPs across all studies, and then recomputed the significance of the signal associations. The FDRs corresponding to a *P* value were estimated by computing the average number of sample-study combinations that reached that significance threshold for a permuted catalog divided by the total number of combinations that reached the significance threshold based on the actual catalog. If a less significant *P* value had an initial lower FDR estimate than a more significant *P* value, then the more significant *P* value also received that lower FDR estimate. We displayed the first ten permutations generated in the *P* value comparison plots. For the comparison of the most significant imputed sample with the average signal, the FDR for the average signal needed only to control for testing multiple studies as there were no sample-specific predictions. In this specific comparison the FDR for the imputed data were determined as above, but by only considering the most significant *P* value across all samples for a specific study for both the actual and each randomized catalog.

Motif analysis. The motif analysis was conducted for each sample in which there were DNase data available. The foreground for the enrichment was those locations that had a DNase signal above 5 in the observed data and below 1 in the imputed data. The background for the enrichment was restricted to all locations, which had an observed DNase signal above 5. An additional analysis was done where the foreground was all locations that had observed a DNase signal above 5, with a genome-wide background. The motif analysis was conducted using a previously described software and assembled compendium of motifs[57].

doi:10.1038/nbt.3157

51. Hall, M. *et al.* The WEKA Data Mining Software: an update. *SIGKDD Explor.* **11**, 10–18 (2009).
52. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
53. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54–i62 (2009).
54. Bar-Joseph, Z., Gifford, D.K. & Jaakkola, T.S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **17**, S22 (2001).
55. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
56. Karolchik, D. *et al.* The UCSC Genome Browser Database: 2014 update. *Nucleic Acids Res.* **42**, D764–D770 (2014).
57. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).