Improving reference epigenome catalogs by computational prediction

Peter Ebert & Christoph Bock

Bioinformatic imputation of epigenomic marks promises to supplement catalogs of experimental data.

A large international effort is under way to map the human epigenome— chemical modifications of chromatin and DNA that control cellular identity (http://ihec-epigenomes.org/). With contributions from ENCODE[1], Roadmap Epigenomics[2], BLUEPRINT[3], DEEP (http://www.deutsches-epigenom-programm.de) and other projects, epigenome maps for at least 1,000 cell types are being generated. Barring a breakthrough in technology, however, it is unlikely that the epigenomics community will be able to measure every epigenomic modification in every sample and experimental condition of interest. In this issue, Ernst and Kellis[4] describe an approach that should ease the burden on experimentalists by allowing epigenomic marks to be statistically imputed—that is, predicted—from measured data. Epigenome mapping by combining experiments for the most informative marks with bioinformatic imputation of additional marks could massively increase our ability to analyze many epigenomic marks across multiple cell types.

The new method, called ChromImpute, is based on the observation that individual epigenomic marks follow similar patterns in related cell types, and that different epigenomic marks in the same cell type tend to show reproducible patterns of correlation and anticorrelation. It should therefore be possible to predict missing experimental data for a specific epigenomic mark in a given cell type by combining experimental data for this mark in other cell types and data for other marks in the given cell type. More broadly speaking, the method presented by Ernst and Kellis[4] builds on the proven value of bioinformatic prediction not only in computational epigenetics[5] but also in related areas such as disease genetics[6] and systems biology[7].

The authors train ensembles of regression trees—a statistical learning algorithm that combines many regression models to increase accuracy to predict epigenomic marks. They mainly use experimental data for seven histone modifications (H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3, H3K27ac and H3K9ac) and DNase hypersensitivity as basis for the predictions, and impute a total of 31 histone marks, DNase hypersensitivity, DNA methylation and RNA-seq profiles. The Roadmap Epigenomics and ENCODE projects mapped these eight epigenomic marks in many different cell types, rendering them a solid foundation for epigenome prediction. The resulting statistical learning models provide a quantitative description of how a specific epigenomic mark in a given cell type correlates with other epigenome marks in the same cell type and with the same mark in related cell types. This makes it possible to predict genome-wide maps at high resolution for a given mark based on other marks that have been experimentally measured in the cell type of interest (**Fig. 1**).

To validate their method, Ernst and Kellis[4] predict genome-wide tracks for up to 34 epigenomic marks in 127 different samples, and they compare the predictions with experimental data available for about a quarter of these sample-mark combinations. Good agreement with the experimentally mapped data is observed for most epigenomic marks, with Pearson correlation coefficients of ~0.8 for H3K27ac and DNase hypersensitivity, ~0.7 for H3K4me1 and H3K36me3, and ~0.6 for H3K27me3 and H3K9me3. The best results—with correlation coefficients of ~0.9—are seen for marks that tend to be highly correlated with other marks in the same cell type (H3K4me2) or that are generally similar across cell types (H3K4me3 and DNA methylation).

The authors also compare imputed and observed epigenome tracks in terms of their agreement with annotated gene models, their ability to detect similarities between related cell types, and the identification of biologically most relevant tissue types for disease-associated genetic variants. In almost all cases, the imputed data correlate better with biologically interesting features than do the observed data. This leads the authors to conclude that epigenome prediction not only is useful for inferring epigenomic marks that were not measured experimentally but also might provide a broadly relevant strategy for obtaining higher quality epigenome tracks based on the combination of imputed and experimental data for a given cell type.

The authors also explore other applications for ChromImpute. They show that a comparison between observed and imputed epigenome tracks can detect low-quality samples with better sensitivity than existing computational methods for ChIP-seq quality control. This could provide a scalable alternative to quality control by manual inspection, which is so far the most sensitive method for detecting problematic samples.

In addition, they prioritize epigenomic marks based on how informative their predictions are for inferring other marks. These results indicate that the Roadmap Epigenomics project would benefit by including experimental data for additional marks (H3K18ac and H3K79me2 are suggested) to the core marks it covers. The value of adding at least one acetylation mark is supported by data from many labs,

Peter Ebert and Christoph Bock are at the Max Planck Institute for Informatics, Saarbrücken, Germany. Peter Ebert is also at the Graduate School of Computer Science, Saarland University, Saarbrücken, Germany. Christoph Bock is also at the CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria, and the Department of Laboratory Medicine, Medical University of Vienna, Vienna, Austria.
e-mail: cbock@cemm.oeaw.ac.at

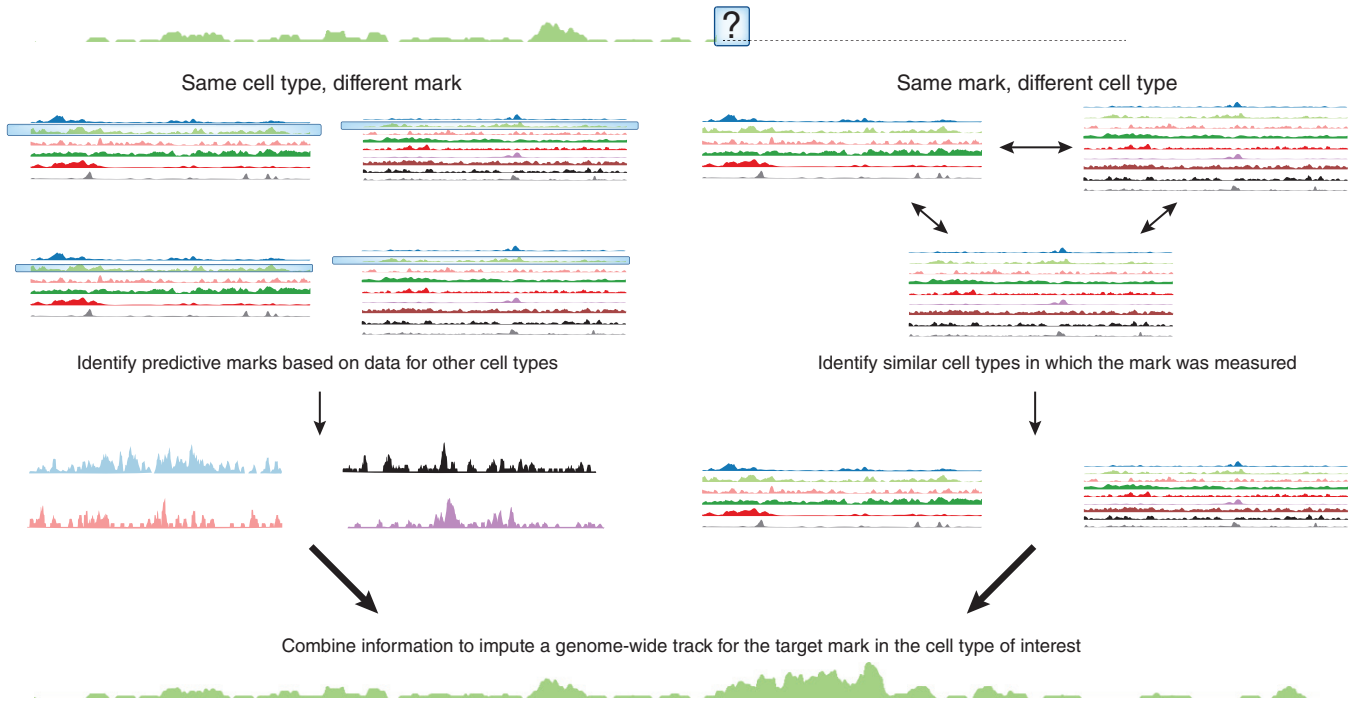

Figure 1 The ChromImpute workflow. Two sources of information are harnessed to predict the genome-wide profile of epigenomic marks that have not been experimentally measured. First, the correlations between different marks are analyzed in all available samples. This knowledge can then be applied to calculate the genome-wide profiles of additional marks based on those marks that have been experimentally determined. Second, the correlation analysis between different marks in the same sample is complemented by an analysis of the same mark in the overall most related samples for which the epigenomic mark of interest has been experimentally measured. Combining the two sources of information yields a high-accuracy estimate of the genome-wide profile of a given mark.

and H3K27ac has been included as a required histone mark for the International Human Epigenome Consortium. Finally, the authors use imputed data for 127 samples to calculate a comprehensive segmentation map of regulatory elements for the human genome, and they make their epigenome predictions available for download and visual exploration using the UCSC Genome Browser and the WashU Epigenome Browser.

Overall, this study firmly establishes bioinformatic prediction of epigenome maps as a suitable method for computationally completing, extending and polishing reference epigenomes. ChromImpute will be particularly useful for large-scale reference epigenome consortia that find it difficult to fill in the last few holes in otherwise complete catalogs. It will also aid researchers who study epigenomic marks that are not commonly examined by the international consortia because it allows prediction of genome-wide profiles for a large number of cell types using training data from just a handful of samples in which these marks were profiled.

In light of these exciting results, should we cut back on experimental epigenome mapping and rely instead on bioinformatic prediction? Of course not, for many reasons. First, ChromImpute requires at least some experimental epigenome data for the cell type of interest to predict comprehensive cell type–specific epigenomes. Second, combining and comparing imputed and observed data in the same cell type makes it possible to detect deviations from the predictions that are specific for a given cell type and epigenomic mark, which could provide a powerful strategy for identifying relevant regulatory mechanisms. Third, Ernst and Kellis[4] show for the epigenomes of pluripotent stem cells that a large number of biological replicates ultimately trumps the accuracy of bioinformatic prediction, thus indicating that the main strength of the imputation method lies in its ability to reduce the experimental noise in the data. Fourth, there is a danger that imputation may 'average out' biologically relevant heterogeneity, which has been observed in single-cell DNA methylation maps[8,9] and could even provide a predictive biomarker for cancer risk[10]. Fifth, the power of imputed data continues to increase with the number of experimentally profiled epigenomic marks, and in order to obtain high-quality imputed data sets it is still necessary to measure several epigenomic marks in each cell type of interest.

So rather than going for either experimental mapping or bioinformatic prediction, we should combine the power of both. For the biologically most important and experimentally accessible samples, comprehensive experimental mapping in large numbers of biological replicates continues to be the method of choice, a strategy that also accounts for the inherent biological variability of primary samples. In addition, focused epigenome mapping combined with bioinformatic prediction will make it possible to scale the power of comprehensive epigenome analysis to disease studies in which sample material may be insufficient for comprehensive epigenome mapping (for example, primary cancer cell samples) or for which it is prohibitively costly to study more than one epigenomic mark (for example, samples from large-scale population studies).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. ENCODE Project Consortium. *Science* **306**, 636–640 (2004).
2. Bernstein, B.E. *et al. Nat. Biotechnol.* **28**, 1045–1048 (2010).
3. Adams, D. *et al. Nat. Biotechnol.* **30**, 224–226 (2012).
4. Ernst, J. & Kellis, M. *Nat. Biotechnol.* **33**, 364–376 (2015).
5. Bock, C. & Lengauer, T. *Bioinformatics* **24**, 1–10 (2008).
6. Marchini, J. & Howie, B. *Nat. Rev. Genet.* **11**, 499–511 (2010).
7. Stolovitzky, G., Monroe, D. & Califano, A. *Ann. NY Acad. Sci.* **1115**, 1–22 (2007).
8. Farlik, M. *et al. Cell Rep.* **10**, 1386–1397 (2015).
9. Smallwood, S.A. *et al. Nat. Methods* **11**, 817–820 (2014).
10. Teschendorff, A.E. *et al. Genome Med.* **4**, 24 (2012).