Supplementary Information: Denoising genome-wide histone ChIP-seq with convolutional neural networks

1 Dataset preparation

Fold change calculation and peak calling. For each experiment, we used align2rawsignal [1] to generate signal tracks and MACS2 [2] to call peaks, as implemented in the *AQUAS* package [3]. For the signal track, we used fold change relative to the expected uniform distribution of reads after an inverse hyperbolic sine transformation [4]. For computational efficiency, we binned the genome into 25bp segments, averaging the signal in each segment.

Histone marks used. We used different sets of input and output histone marks for different experiments, depending on what was available in their respective datasets. For the same cell type, different individual experiments (using lymphoblastoid cell lines), we trained and tested on H3K4me1, H3K4me3, H3K27ac, H3K27me3, and H3K36me3; we used the same data for the low-ChIP-enrichment experiments. For the different cell type, different individual experiments (using the uniformly-processed Roadmap Epigenomics Consortium datasets [5]), we trained and tested on H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3. For all of the above experiments, we also used data from the control experiments (no antibody) as input. Lastly, for the low-cell-input experiments, we used H3K4me3, H3K9me3, and H3K27me3 from the ULI-NChIP-seq dataset and H3K4me3 and H3K27ac from the MOWChIP-seq dataset.

Low-cell-input datasets. The ULI-NChIP-seq [6] and MOWChIP-seq [7] papers provided several datasets corresponding to different numbers of input cells used. For each protocol, we used the datasets with the lowest number of input cells as the noisy input data (ULI-NChIP-seq: 10^3 cells for H3K9me3 and H3K27me3, $5x10^3$ cells for H3K4me3; MOWChIP-seq: 10^2 cells) and the datasets with the highest number of input cells as the gold-standard, high-quality data (ULI-NChIP-seq: 10^6 cells for H3K9me3, 10^5 cells for H3K4me3 and H3K27me3; MOWChIP-seq: 10^4 cells). The ULI-NChIP-seq data only had matching low- and high-input experiments for a single cell type, so we divided it into chr5-19 for training, chr3-4 for validation, and chr1-2 for testing.

2 Model selection and training

We trained our models on 50,000 positions randomly sampled from peak regions of the genome and 50,000 positions sampled from non-peak regions, sampling from each autosome with equal likelihood. We defined peak regions using the output mark of interest and with the high-quality data. (Further increasing dataset size did not increase performance, and subsampling increased efficiency.) We selected the dataset to be balanced because a uniformly drawn dataset would have had very few peaks and peak regions are particularly important to predict well. We used the *Keras* package [8] for training and AdaGrad [9] as the optimizer.

We evaluated model performance on the entire genome using the Pearson correlation and MSE to evaluate fold change prediction and AUPRC to evaluate peak prediction. We chose model hyperparameters through hold-out validation on the low-sequencing-depth denoising task with GM12878 as the training cell line [10], holding out a random 20% subset of the training data for validation. A sequence length of 1001 bins (ie, 500 bins – 12,500 bp – on either side of the sequence position we were predicting) with 6 convolutional filters each 51 bins in length yielded optimal validation performance. To be sure that our model architecture generalized, we used the same architecture and hyperparameters for all denoising tasks without any further tuning.

3 Genome browser tracks

We uploaded the output of our model, together with the original noisy and high-quality data, to the WashU epigenome browser:

- Low sequencing depth (same cell type, different individual): GM18526 (trained on GM12878).
- Low sequencing depth (different cell type, different individual): Mesenchymal stem cells (E026) (trained on data from monocytes (E029)); foreskin fibroblasts (E056) trained on E029.
- Low cell input: Hematopoetic stem and progenitor cells from mouse fetal liver, MOWChIP-seq (trained on GM12878, MOWChIP-seq); Mouse embryonic stem cells, ULI-NChIP-seq, chr1-2 (trained and tested on different chromosomes in the same sample).
- Low ChIP enrichment: GM18526 mixed with 90% control reads (trained on GM12878).

4 Additional results

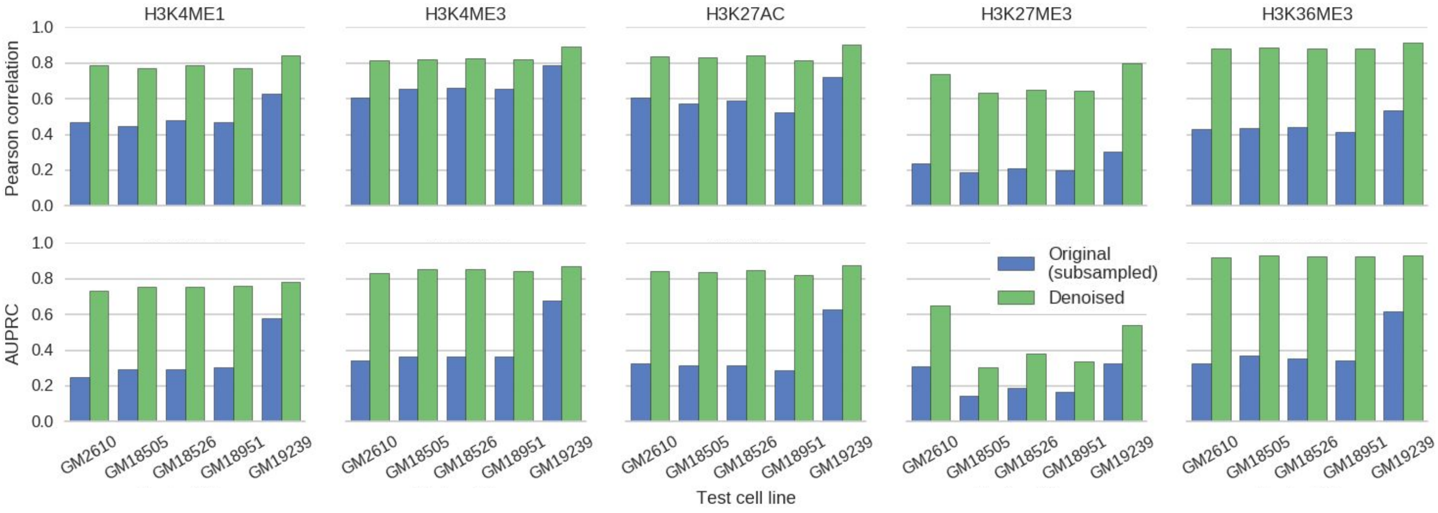

Figure 1: Low sequencing depth experiments on LCL cell lines derived from different individuals. Compared to the signal derived from subsampled reads, the denoised signal shows greater correlation with the full signal (top) and more accurate peak-calling (bottom) across all cell lines.

Table 1: Denoising results on peak regions between test cell line GM18526 and training cell line GM12878. Performance reported is improvement of the denoised model over baseline (original, subsampled reads) on the test cell line.

	MSE (Peaks)	Pearson R (Peaks)
H3K4me1	**-86%** (3.69 → 0.49)	**+56%** (0.44 → 0.70)
H3K4me3	**-83%** (2.93 → 0.50)	**+11%** (0.78 → 0.87)
H3K27ac	**-87%** (3.36 → 0.43)	**+28%** (0.65 → 0.83)
H3K27me3	**-90%** (2.20 → 0.21)	**+103%** (0.18 → 0.36)
H3K36me3	**-93%** (3.78 → 0.25)	**+120%** (0.32 → 0.70)

Table 2: Cross cell-type experiments. Rows are train cell type; columns are test cell type; performance reported is improvement in AUPRC on peak calling task (with baseline and denoised AUPRC in parentheses) averaged across all histone marks used in cross-cell type experiments (H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3).

	Monocytes	MSCs	Fibroblasts
T-cells	**116%** (0.31 → 0.66)	**136%** (0.31 → 0.72)	**94%** (0.35 → 0.69)
Monocytes	-	**139%** (0.31 → 0.73)	**94%** (0.35 → 0.69)
MSCs	-	-	**100%** (0.35 → 0.71)

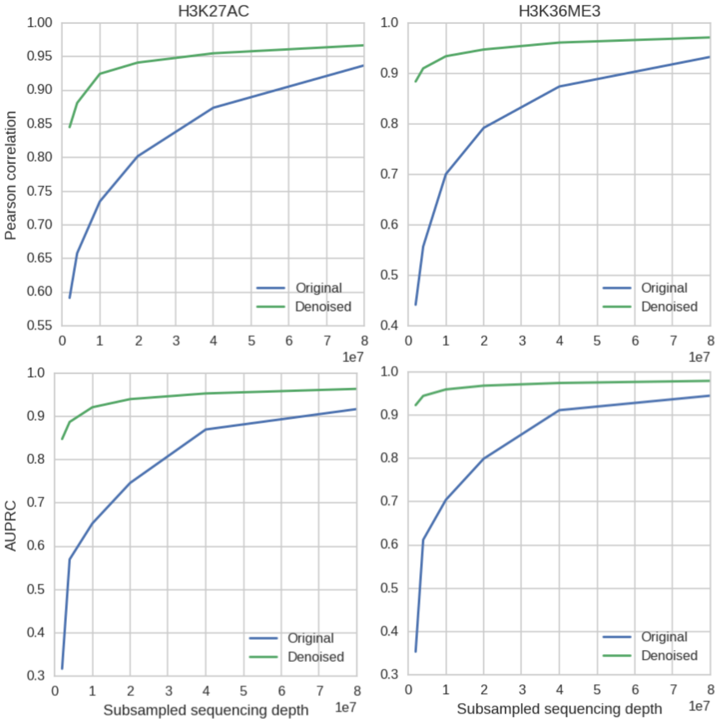

Figure 2: The effect of sequencing depth on baseline and denoising performance in the same cell type, different individual setting (trained on GM12878, tested on GM18526). With 1M reads, the denoised H3K27ac data is equivalent in quality to a dataset with 30M+ reads, and the H3K36me3 data is equivalent in quality to a dataset with 45M reads.

References

[1] A. Kundaje, “align2rawsignal,” 2013. [Online]. Available: https://code.google.com/archive/p/align2rawsignal/

[2] J. Feng, T. Liu, B. Qin, Y. Zhang, and X. S. Liu, “Identifying ChIP-seq enrichment using MACS.” *Nature Protocols*, vol. 7, no. 9, pp. 1728–40, 9 2012. [Online]. Available: http://dx.doi.org/10.1038/nprot.2012.101

[3] J.-W. Lee and A. Kundaje, “AQUAS TF ChIP-seq pipeline,” 2016. [Online]. Available: https://github.com/kundajelab/TF{_}chipseq{_}pipeline

[4] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble, “Unsupervised pattern discovery in human chromatin structure through genomic segmentation.” *Nature Methods*, vol. 9, no. 5, pp. 473–6, 5 2012. [Online]. Available: http://dx.doi.org/10.1038/nmeth.1937

[5] R. E. Consortium *et al.*, “Integrative analysis of 111 reference human epigenomes,” *Nature*, vol. 518, no. 7539, pp. 317–330, 2 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14248

[6] J. Brind’Amour, S. Liu, M. Hudson, C. Chen, M. M. Karimi, and M. C. Lorincz, “An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations,” *Nature Communications*, vol. 6, p. 6033, 1 2015. [Online]. Available: http://www.nature.com/ncomms/2015/150121/ncomms7033/full/ncomms7033.html

[7] Z. Cao, C. Chen, B. He, K. Tan, and C. Lu, “A microfluidic device for epigenomic profiling using 100 cells.” *Nature Methods*, vol. 12, no. 10, pp. 959–62, 10 2015. [Online]. Available: http://www.nature.com.laneproxy.stanford.edu/nmeth/journal/v12/n10/full/nmeth.3488.html

[8] François Chollet, “Keras,” 2015. [Online]. Available: https://github.com/fchollet/keras

[9] J. Duchi, E. Hazan, and Y. Singer, “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization,” *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011. [Online]. Available: http://www.jmlr.org/papers/v12/duchi11a.html

[10] M. Kasowski *et al.*, “Extensive variation in chromatin states across humans.” *Science (New York, N.Y.)*, vol. 342, no. 6159, pp. 750–2, 11 2013. [Online]. Available: http://www.sciencemag.org/content/342/6159/750