bioRxiv preprint doi: https://doi.org/10.1101/052118; this version posted May 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Denoising genome-wide histone ChIP-seq with convolutional neural networks

Pang Wei Koh*, Emma Pierson*, and Anshul Kundaje

Departments of Computer Science and Genetics, Stanford University

pangwei@cs.stanford.edu, {emmap1, akundaje}@stanford.edu

Abstract

Chromatin immunoprecipitation sequencing (ChIP-seq) experiments targeting histone modifications are commonly used to characterize the dynamic epigenomes of diverse cell types and tissues. However, suboptimal experimental parameters such as poor ChIP enrichment, low cell input, low library complexity, and low sequencing depth can significantly affect the quality and sensitivity of histone ChIP-seq experiments. We show that a convolutional neural network trained to learn a mapping between suboptimal and high-quality histone ChIP-seq data in reference cell types can overcome various sources of noise and substantially enhance signal when applied to low-quality samples across individuals, cell types, and species. This approach allows us to reduce cost and increase data quality. More broadly, our approach – using a high-dimensional discriminative model to encode a generative noise process – is generally applicable to biological problems where it is easy to generate noisy data but difficult to analytically characterize the noise or underlying data distribution.

1 Introduction

Distinct histone modifications are associated with different classes of functional genomic elements. Chromatin immunoprecipitation sequencing (ChIP-seq) experiments targeting histone modifications are commonly used to profile the context-specific epigenomes of populations of cells [1, 2, 3]. However, high-quality histone ChIP-seq experiments typically require high-quality antibodies, high ChIP enrichment, millions of cells, and deep sequencing [4]. These experimental conditions are often difficult, costly, or even impossible to attain, resulting in low sensitivity and specificity of measurements especially in low input samples such as rare populations of primary cells and tissues [5, 6]. To overcome these limitations, we train a convolutional neural network (CNN) [7, 8] to learn a generalizable mapping between suboptimal and high-quality ChIP-seq data. The model substantially attenuates three primary sources of noise – due to low sequencing depth, low cell input, and low ChIP enrichment – enhancing signal in low-quality samples across individuals, cell types, and species.

2 Model

CNNs have recently been used to successfully predict regulatory sequence determinants of DNA and RNA binding proteins [9, 10] and chromatin accessibility [11]. Here, we use them to learn a mapping between multiple noisy ChIP-seq experiments on several histone marks and a high-quality ChIP-seq experiment for one target histone mark by training on pairs of noisy $X^{(n)}$ and high-quality $Y^{(n)}$ in the same cellular contexts, where n indexes over the set of training examples. In each pair $(X^{(n)}, Y^{(n)})$, $Y^{(n)}$ is a scalar representing binary peak calls or continuous signal of the target histone mark from a high-quality experiment at a specific genomic position, and $X^{(n)}$ is a matrix, where $X^{(n)}_{ij}$ is the continuous signal of the jth histone mark from the noisy experiment at the ith genomic coordinate in a 25,025 bp window centered at the genomic position at which $Y^{(n)}$ was measured. (See Supp. Info. for more information on model training and hyperparameter selection.) X is fed through a convolutional layer, a rectified linear unit (ReLU), and then a dense layer to predict Y.

For each target mark, we train two separate CNNs to accomplish two tasks: a regression task to predict histone ChIP signal (i.e., the fold enrichment of ChIP reads over input DNA control) and a binary classification task to predict the presence or absence of a significant histone mark peak. To predict fold-enrichment, the dense layer uses a ReLU and we minimize mean-squared error; to call peaks, the dense layer uses a sigmoid activation (because the output is binary) and we minimize cross-entropy loss.

*These authors contributed equally to this work.

bioRxiv preprint doi: https://doi.org/10.1101/052118; this version posted May 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

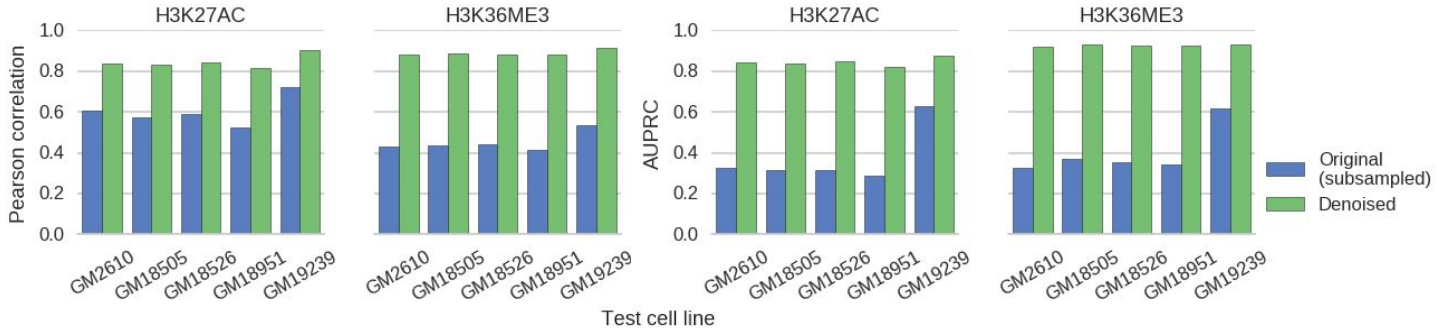


Figure 1: Low sequencing depth experiments on LCL cell lines derived from different individuals. Compared to the signal derived from subsampled reads, the denoised signal shows greater correlation with the full signal (left) and more accurate peak-calling (right) across all cell lines. Here, we show a representative narrow mark (H3K27ac) and broad mark (H3K36me3). Results on other marks are similar except for H3K27me3, where baseline peak calling performance is particularly poor, though the denoised output is still significantly better. Full results are in the Supp. Info.

3 Experiments

We tested our model on three distinct sources of noise: low sequencing depth, low cell input, and low ChIP enrichment. In all cases, the inputs to our model were noisy signal measurements of multiple histone marks (for marks used, see Supp. Info.), and we trained separate models to predict the high-quality signal and peak calls for each target mark. For the regression tasks (predicting signal), we evaluated performance by computing the Pearson correlation and mean squared error (MSE) between the predicted and measured high-quality signal profiles. We compared this to the baseline performance obtained by directly comparing the noisy and high-quality signal profiles of the target mark. For the classification tasks (predicting presence or absence of a peak), we used the area under the precision-recall curve (AUPRC) obtained by comparing our model's output to peaks called by the MACS2 peak caller [12] on the high-quality signal for the target mark. We compared the AUPRC of our model to a baseline obtained by comparing MACS2 peaks on the noisy data for the target mark to those obtained from the high-quality data for the target mark (see Supp. Info. for further details on dataset preparation and model training).

3.1 Low sequencing depth

A minimum of 40-50M reads is recommended for optimal sensitivity for histone ChIP-seq experiments in human samples targeting most canonical histone marks [4]. Adhering to this standard may often be infeasible due to cost and/or low complexity libraries from low input samples. Motivated by these constraints, we tested whether our model could recover high-read depth signal from low-read depth experiments.

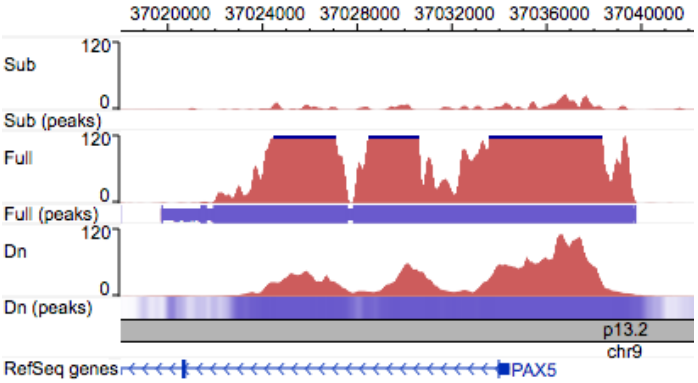
Same cell type across different individuals. We evaluated models on lymphoblastoid cell lines (LCLs) derived from six individuals of diverse ancestry (European (CEU), Yoruba (YRI), Japanese, Han Chinese, San) [13]. We used the CEU-derived cell line (GM12878) to train our model to reconstruct the high-depth signal (100M+ reads per mark) from a simulated noisy signal derived by subsampling 1M reads per mark. On the other five cell lines, our model significantly improved Pearson correlation between the full and noisy signal (Fig. 1, left) and the accuracy of peak calling (Fig. 1, right). Using just 1M reads per mark, the predicted output of our model was equivalent in quality to signal derived from 30M reads (H3K27ac) and 45M reads (H3K36me3) (Supp. Info.). Fig. 2a shows an example of the model correctly recovering H3K27ac structure at the promoter of the *PAX5* gene, a master transcription factor required for differentiation into the B-lymphoid lineage [14].

Table 1: Denoising results on differential regions (diff. reg.) 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. Peak-calling results on H3K27me3 are omitted due to the lack of peak calls in differential regions; all results on H3K36me3 are omitted due to low number of differential regions.

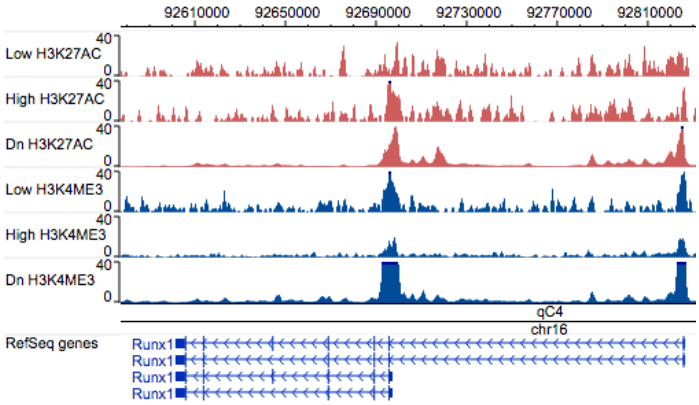
	MSE (diff. reg.)	Pearson R (diff. reg.)	AUPRC (diff. reg.)
H3K4me1	**-85%** (4.01 → 0.57)	**+59%** (0.37 → 0.59)	**+03%** (0.93 → 0.97)
H3K4me3	**-75%** (2.88 → 0.70)	**+14%** (0.63 → 0.72)	**+09%** (0.79 → 0.87)
H3K27ac	**-86%** (3.43 → 0.48)	**+39%** (0.55 → 0.77)	**+05%** (0.91 → 0.96)
H3K27me3	**-80%** (0.78 → 0.15)	**+106%** (0.14 → 0.30)	-

We confirmed our model was not simply memorizing the profile of the training cell line (GM12878) and copying it to the test cell lines by examining differential regions, called by DESeq [15], between GM12878 and the other cell lines [13]. Our model improved correlation and peak-calling even in those regions (Table

bioRxiv preprint doi: https://doi.org/10.1101/052118; this version posted May 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



(a) Low sequencing depth. H3K27ac tracks on GM18526 at the PAX5 promoter. The model recovers the structure of the three peaks and correctly calls a broad peak, whereas the subsampled data misses the structure and the peak call. Sub = subsampled data (1M reads), Full = original data (142M reads), Dn = denoised model output.



(b) Low cell input. H3K27ac and H3K4me3 signal across the Runx1 gene in mouse HSPCs. The model was trained on MOWChIP-seq data generated from human LCL (GM12878) and captures two strong peaks at the promoters of the two isoform classes and removes much of the intervening noise. Low = 10^2 cells, High = 10^4 cells, Dn = denoised model output.

Figure 2: Genome browser tracks comparing model output with baseline and gold-standard data.

1). Similarly, our model also improved correlation on the regions of the genome with enriched signal (i.e. called as statistically significant peaks; see Supp. Info.).

Different cell type across different individuals. We next assessed if a model trained on one cell type in one individual could denoise low-sequencing-depth data from a different cell type in a different individual. As above, the model was trained to output high-depth data (30M reads) from low-depth data (1M reads). We used histone ChIP-seq data spanning T-cells (E037), monocytes (E029), mesenchymal stem cells (MSCs, E026), and fibroblasts (E056) from the Roadmap Epigenomics Consortium [3]. Our model substantially improved the quality of the low-depth signal on the test cell type for all pairs of cell types (Table 2), showing that it can denoise low-depth data on a cell type even if high-depth training data for that cell type is not available.

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

	Monocytes	MSCs	Fibroblasts
T-cells	**+33%** (0.51 → 0.67)	**+58%** (0.44 → 0.70)	**+78%** (0.36 → 0.65)
Monocytes	-	**+59%** (0.44 → 0.70)	**+79%** (0.36 → 0.65)
MSCs	-	-	**+81%** (0.36 → 0.66)

Linear baselines. We compared our CNN model to a linear/logistic regression baseline for signal denoising and peak calling, respectively. When evaluated in the same cell type, different individual setting, the CNN model achieved only slightly better performance when evaluated across the entire genome, but a 3x lower MSE on peaks and a 2x lower MSE on differential regions. This implies that the CNN is better able to learn to match the exact values of the signal tracks on "difficult" regions (i.e., where there is the greatest deviation from the training signal), even though the linear model matches the rough shape.

Denoising and imputation. We varied the set of input marks for the CNN model in same cell type, different individual setting: first, using the noisy target mark as the only input mark (pure denoising), and second, using all noisy marks but the target mark (imputation with noise). In the denoising case, Pearson correlation dropped by 0.03 points and AUPRC dropped by 0.05, on average, compared to when all marks were used as input; thus, additional marks provided some information, but the denoised signal was still substantially better than the original subsampled signal. In the imputation case, performance dropped somewhat on the narrow marks (H3K4me1, H3K4me3, H3K27ac; −0.12 correlation, −0.13 AUPRC) and dropped significantly on the broad marks (H3K27me3, H3K36me3; −0.29 correlation, −0.30 AUPRC). The gap in correlation was even larger within peak regions. Thus, having a noisy version of the target mark substantially boosts recovery of the high-quality signal.

bioRxiv preprint doi: https://doi.org/10.1101/052118; this version posted May 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

3.2 Low cell input

Conventional ChIP-seq protocols require a large number of cells to reach the necessary sequencing depth and library complexity [5, 6], precluding profiling when input material is limited. Several ChIP-seq protocols were recently developed to address this problem. We study ULI-NChIP-seq [5] and MOWChIP-seq [6], which use low cell input (10^2-10^3 cells) to generate signal that is highly correlated, when averaged over bins of size 2-4kbp, with experiments with high cell input. However, at a finer scale of 25bp, the low-input signals from both protocols are poorly correlated with the high-input signals (Table 3). We thus trained our model to recover high-resolution, high-cell-input signal from low-cell-input signal specific to each protocol. For ULI-NChIP-seq, we used a single mouse embryonic stem cell dataset [5]. For MOWChIP-seq, we trained on data from the human LCL GM12878 and tested on hematopoietic stem and progenitor cells (HSPCs) from mouse fetal liver [6]. Our model successfully denoised the low-input signal from both protocols (Table 3). Fig. 2b illustrates our model denoising MOWChIP-seq signal across the *Runx1* gene, a key regulator of HSPCs [16].

Table 3: Low-cell-input experiments. We report improvement of the denoised model output over baseline (original low-input experiments), as compared to high-input experiments.

		MSE	Pearson R	AUPRC
ULI-N	H3K4me3	**-61%** (1.39 → 0.54)	**+208%** (0.13 → 0.41)	**+42%** (0.27 → 0.38)
	H3K9me3	**-46%** (0.51 → 0.27)	**+28%** (0.41 → 0.53)	**+24%** (0.29 → 0.36)
	H3K27me3	**-41%** (0.68 → 0.40)	**+57%** (0.34 → 0.54)	**+25%** (0.36 → 0.45)
MOW	H3K4me3	**-42%** (1.18 → 0.68)	**+122%** (0.14 → 0.31)	**+25%** (0.20 → 0.25)
	H3K27ac	**-21%** (1.44 → 1.14)	**+159%** (0.09 → 0.24)	**+47%** (0.17 → 0.25)

3.3 Low-enrichment ChIP-seq

Histone ChIP-seq experiments use antibodies to enrich for genomic regions associated with the target histone mark. When an antibody with low specificity or sensitivity for the target is used, the resulting ChIP-seq data will be poorly enriched for the target mark. This is a major source of noise [2]. We simulated results from low-enrichment experiments by corrupting GM12878 and GM18526 LCL data [13]. For each histone mark profiled in those cell lines, we kept only 10% of the actual reads and replaced the other 90% with reads taken from the control ChIP-seq experiment, which was done without the use of any antibody; this simulates an antibody with very low specificity. Training on GM12878 and testing on GM18526, our model accurately recovered high-quality, uncorrupted signal from the corrupted data (Table 4).

Table 4: Low-enrichment experiments. We report improvement of the denoised model output over baseline (low-enrichment experiments), as compared to high-enrichment experiments.

	MSE	Pearson R	AUPRC
H3K4me1	**-75%** (0.35 → 0.09)	**+42%** (0.64 → 0.91)	**+119%** (0.42 → 0.92)
H3K4me3	**-86%** (0.44 → 0.06)	**+54%** (0.58 → 0.91)	**+74%** (0.54 → 0.95)
H3K27ac	**-70%** (0.37 → 0.11)	**+37%** (0.65 → 0.90)	**+83%** (0.51 → 0.94)
H3K27me3	**-61%** (0.27 → 0.10)	**+88%** (0.42 → 0.78)	**+178%** (0.18 → 0.49)
H3K36me3	**-82%** (0.36 → 0.06)	**+47%** (0.65 → 0.95)	**+82%** (0.53 → 0.98)

4 Discussion

We use paired noisy and high-quality samples to substantially improve the quality of new, noisy ChIP-seq data. Our approach transfers information from generative noise processes (e.g., mixing in control reads to simulate low-enrichment, or performing low-input experiments) to a flexible discriminative model that can be used to denoise new data. A similar approach can be used in many other biological assays where it is impossible to analytically characterize the noise or the overall data distribution, but possible to generate noisy versions of high-quality samples through experimental or computational perturbation. Our work is conceptually related to the existing literature on structured signal recovery, in particular supervised denoising in images [7, 17, 18] and speech [19]. It complements other efforts to impute genomic data [20]; whereas those methods use high-quality data of one type to impute data of another (e.g., using high-quality H3K27ac signal to impute H3K4me3), we take in low-quality signal of the same type and denoise it. There are many ways to build upon this work. We assume that the noise parameters in the test data are known in advance; in some cases (e.g., the low-cell-input setting) this is true, but in others (e.g., the low-enrichment setting) it is not. Training a single model over various settings of the noise parameters or generation process would make the model more robust. Future work might also make use of cell type similarity information; if cell types A and B are known to be related, their histone signals may be similar as well. These lines of research may further enhance the improvements in data quality we achieve through computational denoising of very noisy but highly structured biological data.

bioRxiv preprint doi: https://doi.org/10.1101/052118; this version posted May 7, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Acknowledgments

We thank Jin-Wook Lee for his assistance with the *AQUAS* pipeline and Kyle Loh, Irene Kaplow, and Nasa Sinnott-Armstrong for their helpful feedback and suggestions.

References

[1] J. Ernst and M. Kellis, "ChromHMM: automating chromatin-state discovery and characterization." *Nature Methods*, vol. 9, no. 3, pp. 215–6, 3 2012. [Online]. Available: http://dx.doi.org/10.1038/nmeth.1906

[2] S. G. Landt *et al.*, "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia." *Genome Research*, vol. 22, no. 9, pp. 1813–31, 9 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3431496{&}tool=pmcentrez{&}rendertype=abstract

[3] 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

[4] Y. L. Jung *et al.*, "Impact of sequencing depth in ChIP-seq experiments." *Nucleic Acids Research*, vol. 42, no. 9, p. e74, 5 2014. [Online]. Available: http://nar.oxfordjournals.org/content/early/2014/03/05/nar.gku178

[5] 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

[6] 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

[7] V. Jain and S. Seung, "Natural Image Denoising with Convolutional Networks," in *Advances in Neural Information Processing Systems*, 2009, pp. 769–776. [Online]. Available: http://papers.nips.cc/paper/3506-natural-image-denoising-with-convolutional-networks

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-w

[9] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, no. 8, pp. 831–838, 7 2015. [Online]. Available: http://dx.doi.org/10.1038/nbt.3300

[10] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model." *Nature Methods*, vol. 12, no. 10, pp. 931–934, 8 2015. [Online]. Available: http://dx.doi.org/10.1038/nmeth.3547

[11] D. R. Kelley, J. Snoek, and J. Rinn, "Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks." Tech. Rep., 10 2015. [Online]. Available: http://biorxiv.org/content/early/2016/02/18/028399.abstract

[12] 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

[13] 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

[14] S. L. Nutt, B. Heavey, A. G. Rolink, and M. Busslinger, "Commitment to the B-lymphoid lineage depends on the transcription factor Pax5." *Nature*, vol. 401, no. 6753, pp. 556–62, 10 1999. [Online]. Available: http://dx.doi.org/10.1038/44076

[15] S. Anders and W. Huber, "Differential expression analysis for sequence count data." *Genome Biology*, vol. 11, no. 10, p. R106, 1 2010. [Online]. Available: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r106

[16] T. E. North *et al.*, "Runx1 Expression Marks Long-Term Repopulating Hematopoietic Stem Cells in the Midgestation Mouse Embryo," *Immunity*, vol. 16, no. 5, pp. 661–672, 5 2002. [Online]. Available: http://www.cell.com/article/S1074761302002960/fulltext

[17] J. Xie, L. Xu, and E. Chen, "Image Denoising and Inpainting with Deep Neural Networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 341–349. [Online]. Available: http://papers.nips.cc/paper/4686-image-denoising

[18] A. Mousavi, A. B. Patel, and R. G. Baraniuk, "A Deep Learning Approach to Structured Signal Recovery," 8 2015. [Online]. Available: http://arxiv.org/abs/1508.04065

[19] A. Maas and Q. Le, "Recurrent Neural Networks for Noise Reduction in Robust ASR." *INTERSPEECH*, pp. 3–6, 2012. [Online]. Available: https://research.google.com/pubs/pub45168.html

[20] J. Ernst and M. Kellis, "Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues," *Nature Biotechnology*, vol. 33, no. 4, pp. 364–376, 2 2015. [Online]. Available: http://dx.doi.org/10.1038/nbt.3157