fluff: exploratory analysis and visualization of high-throughput sequencing data

Georgios Georgiou[1] **and Simon J. van Heeringen**[1]

[1]**Radboud University, Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, 6500HB Nijmegen, The Netherlands**

ABSTRACT

Summary: In this application note we describe fluff, a software package that allows for simple exploration, clustering and visualization of high-throughput sequencing data mapped to a reference genome. The package contains three command-line tools to generate publication-quality figures in an uncomplicated manner using sensible defaults. Genome-wide data can be aggregated, clustered and visualized in a heatmap, according to different clustering methods. This includes a predefined setting to identify dynamic clusters between different conditions or developmental stages. Alternatively, clustered data can be visualized in a bandplot. Finally, fluff includes a tool to generate genomic profiles. As command-line tools, the fluff programs can easily be integrated into standard analysis pipelines. The installation is straightforward and documentation is available at http://fluff.readthedocs.org.
Availability: fluff is implemented in Python and runs on Linux. The source code is freely available for download at https://github.com/simonvh/fluff.
Contact: s.vanheeringen@science.ru.nl

Keywords: ChIP-seq, clustering, next-generation sequencing, high-throughput sequencing, visualization, Python

INTRODUCTION

The advances in sequencing technology and the reduction of costs have led to a rapid increase of High-Throughput Sequencing (HTS) data. Applications include chromatin immunoprecipitation followed by high-throughput deep sequencing (ChIP-seq; Robertson et al. (2007)) to determine the genomic location of DNA-associated proteins, chromatin accessibility assays (Buenrostro et al., 2013; Hesselberth et al., 2009) and bisulfite sequencing to assay DNA methylation (Lister et al., 2009). The integration of these diverse data allow identification of the epigenomic state, for instance in different tissues (Martens and Stunnenberg, 2013; Roadmap Epigenomics Consortium et al., 2015) or during development (Hontelez et al., 2015). However, the scale and complexity of these datasets call for the use of computational methods that facilitate data exploration and visualization.

Various options exist to explore and visualize HTS data mapped to a reference genome, for instance in aggregated form such as heatmaps and average profiles. These include general purpose modules for specific programming languages (Huber et al., 2015), dedicated HTS modules (Dale et al., 2014; Statham et al., 2010; Akalin et al., 2015), command-line tools (Shen et al., 2014; Giannopoulou and Elemento, 2011), web tools (Ramírez et al., 2014), stand-alone applications (Ramírez et al., 2014; Ye et al., 2011) and tools that depend on other software for visualization (Heinz et al., 2010). Here, we present fluff, a Python package for visual, reference-based HTS data exploration. It includes command-line applications to both cluster and visualize aggregated signals in genomic regions, as well as to create genome browser-like profiles. The scripts can be included in analysis pipelines and accept commonly used file formats. The fluff applications are pitched at the beginner to intermediate user. They have sensible defaults, yet allow for customizable creation of high-quality, publication-ready figures.

METHODS

General

Detailed documentation, including tutorials, is available at http://fluff.readthedocs.org. Fluff is implemented in Python and uses several previously published modules (Brewer (2016); Anders et al. (2015); Dale et al. (2011); Quinlan and Hall (2010); Li et al. (2009); de Hoon et al. (2004), see Supplemental Information). All fluff tools support indexed BAM, bigWig or (tabix-indexed) BED, WIG or bedGraph files as input. A large selection of major image formats are supported as output. The fluff tools were developed to explore ChIP-seq data, however, they will work with any type of data where (spliced) reads can be mapped to a genomic reference. For instance DNA methylation profiles from bisulfite-sequencing or RNA-seq data (Supplemental Figure 1) can also be visualized.

Normalization Normalization of sequencing data is critical for downstream analysis and various methods have been proposed (see for instance Angelini et al. (2015) and Bailey et al. (2013) for an overview of ChIP-seq normalization methods). For visualization, the most important factor is the sequencing read depth. Therefore fluff has the option to normalize to the total number of mapped reads. Alternatively, averaged signal files such as bigWig tracks that are processed or normalized by a different method can be used as input.

Program descriptions

Heatmaps Visualization of HTS data as heatmaps, where rows represent different genomic regions, can highlight important aspects of the data, like differential enrichment or positional patterns for specific groups of features. In addition, it allows for comparison between multiple regions within the same or between different experiments. The *fluff heatmap* tool visualizes HTS data on basis of list of genomic coordinates. The data can optionally be clustered using either k-means or hierarchical clustering. For clustering, the read counts in the bins are normalized to the 75 percentile. The distance can be calculated using either the Euclidean distance or Pearson correlation similarity.

If the regions in the input file are not strand-specific, different clusters might represent the same strand-specific profile in two different orientations. Clusters that are mirrored relative to the center can optionally be merged. Here, the similarity is based on the chi-squared p-value of the mean profile per cluster.

One important use case for clustering is the ability to identify dynamic patterns, for instance during different time points or conditions. For this purpose, clustering on the binned signal is not ideal. Therefore, *fluff heatmap* provides the option to cluster genomic regions based on a single value derived from the number of reads in the feature centers (+/- 1kb). In combination with the Pearson correlation metric, this allows for efficient retrieval of dynamic clusters. The difference is illustrated in Figure 2.

Bandplots In heatmaps, more subtle patterns can be difficult to detect, as the dynamic range of signal intensities is not well-reflected in the color scale. Therefore, as an alternative to a heatmap, *fluff bandplot* plots the average profiles in small multiples (Shoresh and Wong, 2012). Here, the spatial encoding of the signal allows for more accurate comparison of values (Gehlenborg et al., 2012). The median enrichment is visualized as a black line with the 50th and 90th percentile as a dark and light colour respectively.

Profiles. Genome browsers are unrivaled for data exploration and visualization in a genomic context. However, it can be useful to create profiles of HTS data in genomic intervals using a consistent command-line tool, that can optionally be automated. The *fluff profile* tool can plot summarized profiles from one or more profiles, together with (gene) annotation from a BED12-formatted file.

Analysis

In short, FASTQ files were downloaded from NCBI GEO (Edgar et al., 2002) and mapped to the human genome (hg19) using bwa (Li and Durbin, 2009). Duplicate reads were marked using bamUtil (http://genome.sph.umich.edu/wiki/BamUtil). All BAM files from replicate experiments were merged. Peaks were called using MACS2 (Zhang et al., 2008) with default settings. See Supplemental Information for specific details and accession numbers.

RESULTS

Demonstrating fluff: dynamic enhancers during macrophage differentiation

To illustrate the functionality of fluff we visualized previously published ChIP-seq data (Saeed et al., 2014). Here, the epigenomes of human monocytes and in vitro-differentiated naïve, tolerized, and

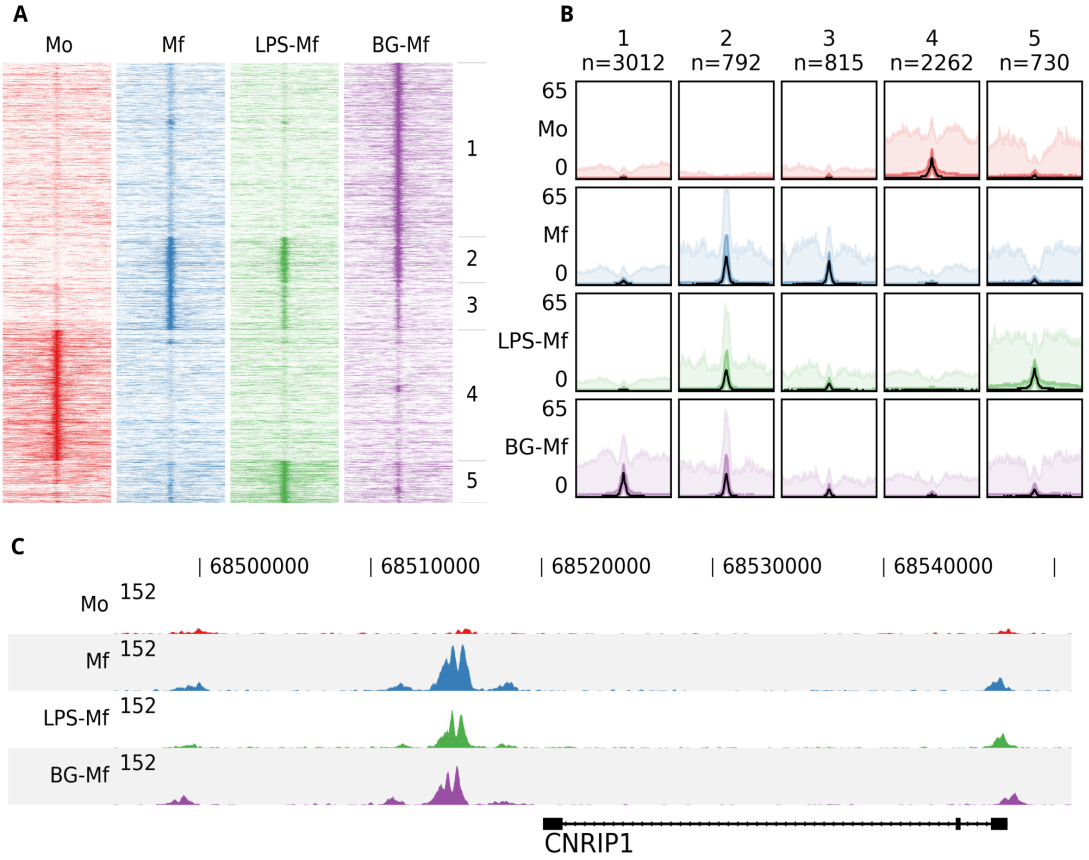

Figure 1. An example of the fluff output. All panels were generated by the fluff command-line tools and were not post-processed or edited. (A) Heatmap showing the results of k-means clustering (k=5, metric=Pearson) of dynamic H3K27ac regions in monocytes (Mo), naïve macrophages (Mf), tolerized (LPS-Mf) and trained cells (BG-Mf) (Saeed et al., 2014). ChIP-seq read counts are visualized in 100-bp bins in 24-kb regions. (B) Bandplot showing the average profile (median: black, 50 percent: dark color, 90 percent: light color) of the clusters as identified in Fig. 1A. (C) The H3K27ac ChIP-seq profiles at the CNRIP1 gene locus, which shows a gain of H3K27ac in Mf, LPS-Mf and BG-Mf relative to Mo.

trained macrophages were analyzed, with the aim to understand the epigenetic basis of innate immunity. Circulating monocytes (Mo) were differentiated into three macrophages states: to macrophages (Mf), to long-term tolerant cells (LPS-Mf) by exposition to lipopolysaccharide and to trained immune cells (BG-Mf) by priming with β-glucan. We used fluff heatmap to cluster and visualize the signal of histone 3 lysine 27 acetylation (H3K27ac), which is located at active enhancers and promoters (Fig. 1A). The input consisted of a BED file with 7,611 differentially regulated enhancers (Supplemental Table 1) and four BAM files, for each of the monocytes and three types of macrophages. Using k-means clustering (k = 5) with the Pearson correlation metric, the heatmap recapitulates the H3K27ac dynamics as described (Saeed et al., 2014).

While heatmaps are often used for visualization of signals over genomic features, either clustered or ordered by signal intensity, it can be difficult to distinguish relative levels of individual clusters. Figure 1B shows an alternative visualization of average enrichment profiles in small multiples. The same clusters as in Fig. 1A are plotted using *fluff bandplot*. Shown are the median (black line), along with the 50th (darker color) and 90th percentile (lighter color) of the data. This allows for more detailed comparisons.

Finally, we illustrate *fluff profile*, which can visualize one or more genomic regions (Fig. 1C). This figure highlights the CNRIP1 gene from cluster 2, which shows a consistent increase of H3K27ac from Mo to Mf, LPS-Mf and BG-Mf. The signal profiles are directly generated from the BAM files.

Identification and visualization of dynamic patterns

Most applications that cluster HTS data for heatmap visualization use a binning approach, followed by clustering using the Euclidean distance. The implicit effect is that the bins are clustered on basis of the spatial patterns relative to the region of interest. Often, this is the desired result, for instance when clustering the ChIP-seq enrichment patterns of different histone modifications at the transcription start sites of genes. However, for other analyses this clustering approach does not suffice. An example could be the ChIP-seq profiles of specific histone modifications correlated to the activity of a regulatory element, such as H3K4me3 at promoters or H3K27ac at enhancers. In this case, a relevant objective is to identify the clusters associated with differential activation dynamics. As illustration, we visualized the H3K27ac enrichment profile at DNaseI hypersensitive sites in human embryonic stem (ES) cells differentiated into different lineages (Xie et al., 2013). Here, H1 ES cells were differentiated into mesendoderm, neural progenitor cells, trophoblast-like cells, and mesenchymal stem cells. We first clustered the H3K27ac profiles at regulatory elements on chromosome 1 using the standard approach, based on comparing all the bins using the Euclidean distance metric (Fig. 2A).

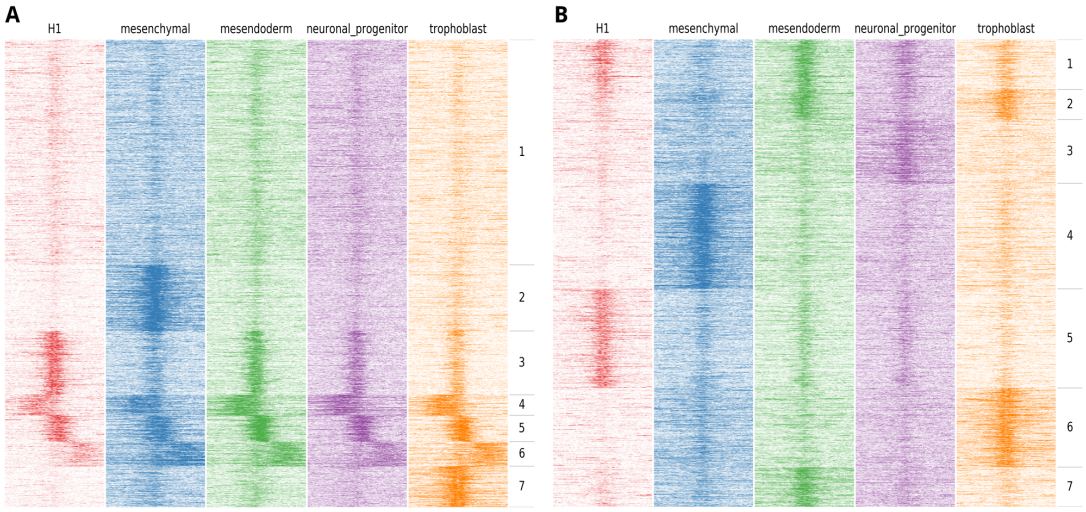

Figure 2. Example of the output of *fluff heatmap* using standard clustering compared to using the dynamics option. Shown are the H3K27ac ChIP-seq read counts in 100bp bins in 20kb around the DNaseI peak summit in human H1 ES cell-derived cells. (A) Heatmap showing the results of k-means clustering of all bins (k=7, metric=Euclidean) (B) Heatmap showing the results of k-means clustering in 2kb regions centered at the peak summit (k=7, metric=Pearson).

Here, we identify two clusters with high enrichment (cluster 3 and cluster 5), a cluster with relatively low, narrow enrichment (cluster 1), and two clusters with broad enhancer domains (cluster 4 and 6). However, only two strong dynamic clusters are identified, cluster 2, which shows enhancers specifically activated in mesenchymal stem cells and cluster 7 which shows enhancers specifically activated in trophoblast-like stem cells. Figure 2B shows an alternative clustering approach implemented in *fluff heatmap*. Here the regions were clustered on basis of the Pearson correlation of read counts in the center of the region (extended to 2kb). This shows a completely different picture and we now can identify enhancers specific to H1 ES cells (cluster 5), mesenchymal (cluster 4), mesendoderm (cluster 7), neuronal progenitor (cluster 3) and trophoblast cells (cluster 6). These lineage-specific enhancer dynamics were not visible in the clustering in Figure 2A.

CONCLUSION

The analysis of multi-dimensional genomic data requires methods for data exploration and visualization. We provide fluff, a Python package that contains several command-line tools to generate figures for use in high-throughput sequencing analysis workflows. We aim to fill the gap between powerful, flexible libraries that require programming skills on the one hand, and intuitive, graphical programs with limited customization possibilities on the other hand. These tools were developed based on a need for straight-

forward analysis and visualization of ChIP-seq data and have been successfully applied in a variety of projects (Menafra et al., 2014; van den Boom et al., 2016; Kouwenhoven et al., 2015). In conclusion, fluff helps to interpret genome-wide experiments by efficient visualization of sequencing data.

ACKNOWLEDGMENTS

This study makes use of data generated by the Blueprint Consortium. A full list of the investigators who contributed to the generation of the data is available from www.blueprint-epigenome.eu. Additionally, this study used data provided by the NIH Roadmap Epigenomics Consortium (http://nihroadmap.nih.gov/epigenomics/) .

REFERENCES

Akalin, A., Franke, V., Vlahoviček, K., Mason, C. E., and Schübeler, D. (2015). Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics*, 31(7):1127–1129.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.

Angelini, C., Heller, R., Volkinshtein, R., and Yekutieli, D. (2015). Is this the right normalization? a diagnostic tool for ChIP-seq normalization. *BMC Bioinformatics*, 16:150.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.*, 9(11):e1003326.

Brewer, C. (2016). ColorBrewer: Color advice for maps. http://www.colorbrewer2.org, accessed: 2016-3-15.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, 10(12):1213–1218.

Dale, R. K., Matzat, L. H., and Lei, E. P. (2014). metaseq: a python package for integrative genome-wide analysis reveals relationships between chromatin insulators and associated nuclear mRNA. *Nucleic Acids Res.*, 42(14):9158–9170.

Dale, R. K., Pedersen, B. S., and Quinlan, A. R. (2011). Pybedtools: a flexible python library for manipulating genomic datasets and annotations. *Bioinformatics*, 27(24):3423–3424.

de Hoon, M. J. L., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. *Bioinformatics*, 20(9):1453–1454.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210.

Gehlenborg, N., Nils, G., and Bang, W. (2012). Points of view: Heat maps. *Nat. Methods*, 9(3):213–213.

Giannopoulou, E. G. and Elemento, O. (2011). An integrated ChIP-seq analysis platform with customizable workflows. *BMC Bioinformatics*, 12:277.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, 38(4):576–589.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, 6(4):283–289.

Hontelez, S., van Kruijsbergen, I., Georgiou, G., van Heeringen, S. J., Bogdanovic, O., Lister, R., and Veenstra, G. J. C. (2015). Embryonic transcription is controlled by maternally defined chromatin state. *Nat. Commun.*, 6:10148.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., and Morgan, M. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nat. Methods*, 12(2):115–121.

Kouwenhoven, E. N., Oti, M., Niehues, H., van Heeringen, S. J., Schalkwijk, J., Stunnenberg, H. G., van Bokhoven, H., and Zhou, H. (2015). Transcription factor p63 bookmarks and regulates dynamic enhancers during epidermal differentiation. *EMBO Rep.*, 16(7):863–878.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322.

Martens, J. H. A. and Stunnenberg, H. G. (2013). BLUEPRINT: mapping human blood cell epigenomes. *Haematologica*, 98(10):1487–1489.

Menafra, R., Brinkman, A. B., Matarese, F., Franci, G., Bartels, S. J. J., Nguyen, L., Shimbo, T., Wade, P. A., Hubner, N. C., and Stunnenberg, H. G. (2014). Genome-wide binding of MBD2 reveals strong preference for highly methylated loci. *PLoS One*, 9(6):e99603.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., and Manke, T. (2014). deeptools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, 42(Web Server issue):W187–91.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4(8):651–657.

Saeed, S., Quintin, J., Kerstens, H. H. D., Rao, N. A., Aghajanirefah, A., Matarese, F., Cheng, S.-C., Ratter, J., Berentsen, K., van der Ent, M. A., Sharifi, N., Janssen-Megens, E. M., Ter Huurne, M., Mandoli, A., van Schaik, T., Ng, A., Burden, F., Downes, K., Frontini, M., Kumar, V., Giamarellos-Bourboulis, E. J., Ouwehand, W. H., van der Meer, J. W. M., Joosten, L. A. B., Wijmenga, C., Martens, J. H. A., Xavier, R. J., Logie, C., Netea, M. G., and Stunnenberg, H. G. (2014). Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science*, 345(6204):1251086.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*, 15:284.

Shoresh, N. and Wong, B. (2012). Points of view: Data exploration. *Nat. Methods*, 9(1):5.

Statham, A. L., Strbenac, D., Coolen, M. W., Stirzaker, C., Clark, S. J., and Robinson, M. D. (2010). Repitools: an R package for the analysis of enrichment-based epigenomic data. *Bioinformatics*, 26(13):1662–1663.

van den Boom, V., Maat, H., Geugien, M., Rodríguez López, A., Sotoca, A. M., Jaques, J., Brouwers-Vos, A. Z., Fusetti, F., Groen, R. W. J., Yuan, H., Martens, A. C. M., Stunnenberg, H. G., Vellenga, E., Martens, J. H. A., and Schuringa, J. J. (2016). Non-canonical PRC1.1 targets active genes independent of H3K27me3 and is essential for leukemogenesis. *Cell Rep.*, 14(2):332–346.

Xie, W., Schultz, M. D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J. W., Tian, S., Hawkins, R. D., Leung, D., Yang, H., Wang, T., Lee, A. Y., Swanson, S. A., Zhang, J., Zhu, Y., Kim, A., Nery, J. R., Urich, M. A., Kuan, S., Yen, C.-A., Klugman, S., Yu, P., Suknuntha, K., Propson, N. E.,

Chen, H., Edsall, L. E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W.-Y., Chi, N. C., Antosiewicz-Bourget, J. E., Slukvin, I., Stewart, R., Zhang, M. Q., Wang, W., Thomson, J. A., Ecker, J. R., and Ren, B. (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, 153(5):1134–1148.

Ye, T., Krebs, A. R., Choukrallah, M.-A., Keime, C., Plewniak, F., Davidson, I., and Tora, L. (2011). seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.*, 39(6):e35.

Zhang, Y., Yong, Z., Tao, L., Meyer, C. A., Jérôme, E., Johnson, D. S., Bernstein, B. E., Chad, N., Myers, R. M., Myles, B., Wei, L., and Shirley Liu, X. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137.