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
*To whom correspondence should be addressed.

Abstract

Summary: Here we introduce fluff, a software package that allows for simple exploration, clustering and visualization of high-throughput sequencing experiments. 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

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, 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), command-line tools (Shen *et al.*, 2014), web tools (Ramírez *et al.*, 2014) and stand-alone applications (Ramírez *et al.*, 2014; Ye *et al.*, 2011). Here, we present fluff, a Python package for visual 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.

Program description and methods

General. The fluff module and command-line tools are implemented in Python. The package can be installed using *pip* or the *conda* package manager. Detailed documentation, including tutorials, is available at http://fluff.readthedocs.org. All fluff tools efficiently obtain reads from indexed BAM files and use read counts for visualization. Alternatively, the RPKM measure can be specified (Reads Per Kb per Million reads). Duplicate reads (SAM flag 1024) and reads that map to multiple locations (mapping quality 0) are removed by default. Where applicable, the scale of the Y-axis can be adjusted individually, or per group of tracks to enable comparison. Colors can be specified by name, HEX code or using ColorBrewer palettes. The format of the output figure is determined by the extension and all major formats (including PNG, JPEG, SVG and PDF) are supported.

Heatmaps and bandplots. The input for *fluff heatmap* consists of a BED file with genomic coordinates, e.g. peaks from a ChIP-seq experiment, and one or more indexed BAM files. The features in the BED file are extended up- and downstream (default 5kb) from the center and divided into bins (default 100bp). 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 BED 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. The data are plotted as heatmap where each row represents a feature. Besides the figure, fluff returns as output the genomic regions (including the cluster number if they were clustered) and the read counts, which can be used for further analysis.
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 Supplementary Figures 1 and 2.
As an alternative to a heatmap, *fluff bandplot* plots the average profiles in small multiples. The median enrichment is visualized as a black line with the 50th and 90th percentile as a dark and light colour respectively. The output from *fluff heatmap* can directly be used in *fluff bandplot*.

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 BAM files, together with annotation from a BED12-formatted file. By default, all reads will be extended to 200 base pairs before creating the profiles.

Example results

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 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 and four BAM files, for each of the monocytes and three types of macrophages (see Supplementary Data for details). Using k-means clustering (k = 5) with the Pearson correlation metric, the heatmap recapitulates the H3K27ac dynamics as described (Saeed *et al.*, 2014).

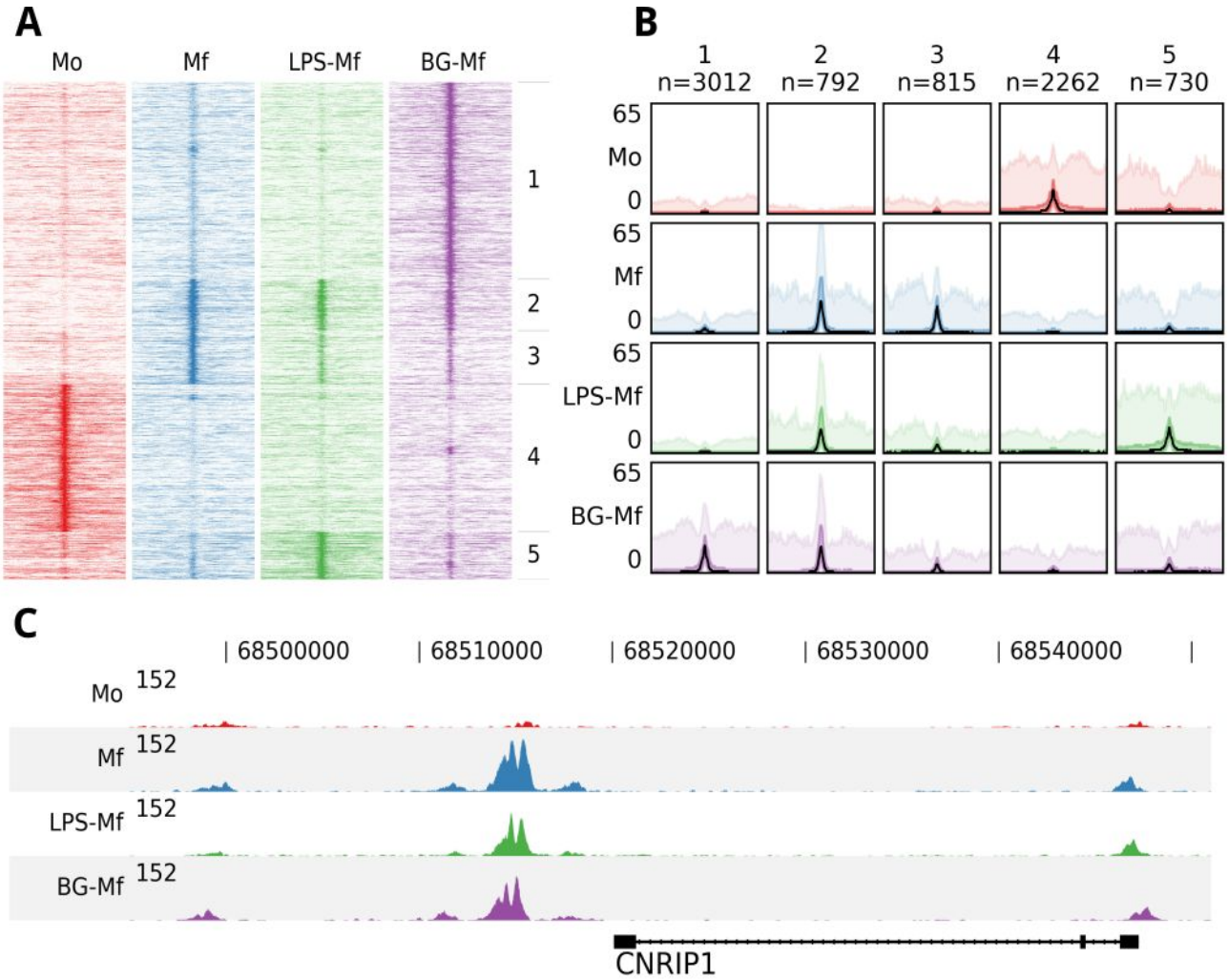

Figure 1. An example of the fluff output. All panels were generated by the fluff command-line tools and were not post-processed or edited. (A) Heatmap showing the results of k-means clustering (k=5, metric=Pearson) of dynamic H3K27ac regions in monocytes (Mo), naïve macrophages (Mf), tolerized (LPS-Mf) 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.

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 visualizes one 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.

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 straightforward 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.

Funding

Funding: This work was supported by the Netherlands Organisation for Scientific Research (NWO-ALW) [863.12.002 to S.J.v.H.]. G.G. was supported by the US National Institutes of Health (NICHD) [R01HD069344].
Conflict of Interest: none declared.

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. Funding for the project was provided by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 282510 BLUEPRINT.

References

van den Boom,V. *et al.* (2016) Non-canonical PRC1.1 Targets Active Genes Independent of H3K27me3 and Is Essential for Leukemogenesis. *Cell Rep.*, **14**, 332–346.

Buenrostro,J.D. *et al.* (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.

Dale,R.K. *et al.* (2014) metaseq: a Python package for integrative genome-wide analysis reveals relationships between chromatin insulators and associated nuclear mRNA. *Nucleic Acids Res.*, **42**, 9158–9170.

Hesselberth,J.R. *et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.

Hontelez,S. *et al.* (2015) Embryonic transcription is controlled by maternally defined chromatin state. *Nat. Commun.*, **6**, 10148.

Huber,W. *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, **12**, 115–121.

Kouwenhoven,E.N. *et al.* (2015) Transcription factor p63 bookmarks and regulates dynamic enhancers during epidermal differentiation. *EMBO Rep.*, **16**, 863–878.

Lister,R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

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

Menafra,R. *et al.* (2014) Genome-wide binding of MBD2 reveals strong preference for highly methylated loci. *PLoS One*, **9**, e99603.

Ramírez,F. *et al.* (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, **42**, W187–91.

Roadmap Epigenomics Consortium *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

Robertson,G. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.

Saeed,S. *et al.* (2014) Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science*, **345**, 1251086.

Shen,L. *et al.* (2014) ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*, **15**, 284.

Ye,T. *et al.* (2011) seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.*, **39**, e35.