EpiCompare: R package for the comparison and quality control of epigenomic peak files

Authors

Sera Choi[1,2], Brian M. Schilder[1,2], Leyla Abbasova[3], Alan E. Murphy[1,2], Nathan G. Skene[1,2*]

1. Department of Brain Sciences, Faculty of Medicine, Imperial College London, London, UK, W12 0BZ
2. UK Dementia Research Institute at Imperial College London, London, UK, W12 0BZ
3. Centre for Developmental Neurobiology, MRC Centre for Neurodevelopmental Disorders, King's College London, London, UK, SE1 1UL

* Corresponding author: Nathan G. Skene; n.skene@imperial.ac.uk

Abstract

Summary

EpiCompare combines a variety of downstream analysis tools to compare, quality control and benchmark different epigenomic datasets. The package requires minimal input from users, can be run with just one line of code and provides all results of the analysis in a single interactive HTML report. EpiCompare thus enables downstream analysis of multiple epigenomic datasets in a simple, effective and user-friendly manner.

Availability and Implementation

EpiCompare is available on Bioconductor (≥ v3.15):

https://bioconductor.org/packages/release/bioc/html/EpiCompare.html
All source code is publically available via GitHub:
https://github.com/neurogenomics/EpiCompare
Documentation website
https://neurogenomics.github.io/EpiCompare
EpiCompare DockerHub repository:
https://hub.docker.com/repository/docker/neurogenomicslab/epicompare

Introduction

Epigenetic processes are crucial regulators of gene expression and transcriptional activity (Allis & Jenuwein, 2016). There is an increasing interest towards understanding disease mechanisms with epigenetic factors, especially in cancer (Cheng et al., 2019), autoimmune diseases (Mazzone et al., 2019) and brain disorders (Hannon et al., 2019; Roussos et al., 2014). In response to this, a variety of novel epigenomic profiling technologies have emerged in recent years (Cazaly et al., 2019; Mehrmohamadi et al., 2021). Yet, how the performance of these new methods compare with traditional approaches and how they differ from one another is largely unknown. Therefore, we need a way of systematically comparing and analysing epigenomic data generated by different methods to contrast their strengths and weaknesses. This would help researchers to choose the appropriate method when designing epigenomic studies and to establish the optimal experimental protocols and data analysis workflows.

Typically, epigenomic data analysis consists of two parts: (1) data processing, where sequences are mapped and peaks are called; and (2) downstream analysis, where peaks are visualised and annotated. There have been movements to standardise and simplify data preprocessing steps through workflow-based pipelines (Ewels et al., 2020) such as *nf-core/chipseq* (Patel et al., 2021) and *nf-core/cutandrun (Cheshire et al., 2022)*, which require just one line of code to run. However for the downstream analysis, the tools are currently scattered in many different packages and platforms, with some requiring idiosyncratic input formats. This makes the latter part of the analysis challenging and time-consuming, especially for those with little or no computational experience.

To address these issues, we introduce EpiCompare, a Bioconductor (Huber et al., 2015) R package for the comparison and quality control of epigenomic data. EpiCompare is able to perform a variety of downstream analysis on multiple epigenomic datasets simultaneously, which can be executed with just one line of code. Some of the main functionalities include precision-recall and functional annotations, which help to assess the extent of overlapping peaks between files and to check if peaks annotate to the same genomic features. The package also generates a single report collating all results of the analysis into a single interactive report file, making it easy for users to view and interpret the results.

Implementation

EpiCompare was implemented using the R programming language (v4.2) (R Core Team, 2021) in accordance with all Bioconductor (Huber et al., 2015) coding and documentation standards. In addition, every time an update is pushed to EpiCompare, extensive checks and unit tests are automatically launched on three OS platforms (Unix, Mac, Window) via continuous integration workflows using both GitHub Actions and Bioconductor. The package EpiCompare can launch all analyses using a single master function (eponymously named *EpiCompare*). Users need only to supply peak files of interest as a named list of GenomicRanges objects

(Lawrence et al., 2013) or as paths to BED files to be automatically imported as GenomicRanges. It also includes several Boolean parameters, allowing users to decide which analyses to perform. When the function is executed, it parses the parameters into an R markdown file, which is ultimately rendered into an HTML document.

First, *EpiCompare* runs three quality control and standardisation checks on all input peak files. The first step involves removal of blacklisted genomic regions containing well-known irregular or anomalous signals (Amemiya, Kundaje & Boyle, 2019). Filtering out these peaks is recommended for quality measures and thus, *EpiCompare* requires that users specify a 'blacklist' peak file. The second control step uses BRGenomics (v1.1.3) *tidyChromosomes* (DeBerardine, 2022) feature to remove peaks that are found in non-standard or mitochondrial chromosomes. Lastly, the final check ensures that all input peak files are based on the same reference genome build. Users must specify the genome build used to generate the peak files and if needed, *EpiCompare* uses rtracklayer (v1.56.0) *liftOver* (Lawrence, Gentleman & Carey, 2009) function to translate the genomic coordinate of peak files across builds.

Usage

EpiCompare can be installed on any Unix, Mac, or Windows OS using BiocManager. Alternatively, EpiCompare can be installed using its dedicated Docker or Singularity container (hosted on DockerHub), which greatly alleviates common challenges with installation and reproducibility. Once EpiCompare is installed, the package can be used with a single line of code or one function call (*EpiCompare*). The function requires two inputs: a list of peak files of interest and a 'blacklist' peak file containing genomic regions of irregular signals. All peak data can be specified as GenomicRanges objects or as paths to BED files. To ensure that genome builds of peak files agree, users must also state the genome build that was used to generate the peak, reference and blacklist files, which can be supplied as a single genome build (e.g. *genome_build="hg19"*) or a named list of mixed genome builds (e.g. *genome_build=list(peakfiles="hg19", reference="hg38", blacklist="hg38")*). In addition to human genome builds (hg19, GRCh38), *EpiCompare* can ingest and/or output files aligned mouse (mm9, mm10) genome builds using interspecies chain files.

In addition, *EpiCompare* offers a suite of analysis tools and plot options to choose from, allowing users to tailor their downstream analysis of epigenomic data. The two optional inputs include reference peak file and duplicate summary outputs from Picard (Broad Institute, 2019). The plot options are summarised in **Table 1**. Once all analyses are complete, the rendered HTML file can be automatically launched in any web browser, or within Rstudio (RStudio Team, 2022). All data and plots produced by the analyses are also stored in a subfolder called "EpiCompare_files".

EpiCompare can also call consensus peaks from groups of peak files via the function *compute_consensus_peaks*. Multiple methods for calling consensus peaks are offered, including a fast but simplistic overlap strategy (*method=“granges”*), and a slower but more accurate strategy that incorporates modelling of peak distributions (*method=”consensusseeker”*) (Samb et al., 2015). This can be helpful as a pre-step for reducing the number of samples being input to *EpiCompare*, and making files more comparable to “replicated peaks” files in databases like ENCODE.

All of the core functions used internally by the main function *EpiCompare* are exported so that they can be used in custom workflows as they may be more generally useful to the bioinformatics community (e.g. *compute_consensus_peaks, gather_files, plot_precision_recall, compute_corr, rebin_peaks, overlap_heatmap, plot_enrichment liftover_grlist*). See here for documentation on all exported functions: https://neurogenomics.github.io/EpiCompare/reference

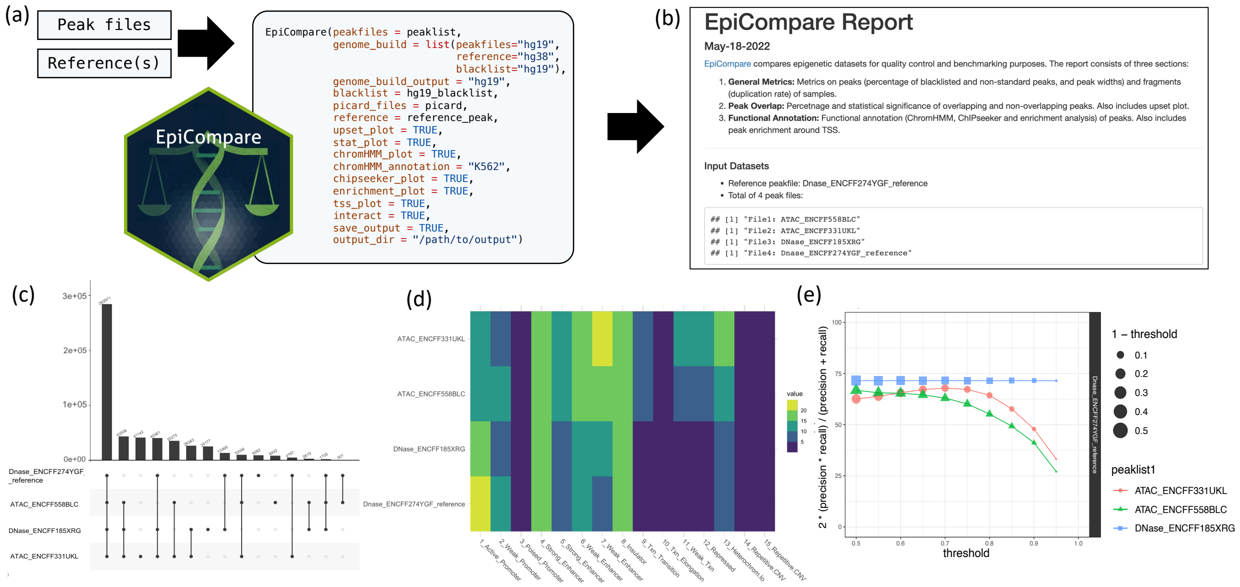
Table 1. Summary of plot options in *EpiCompare*

Plot	**Description**
Upset Plot	Upset plot of the number of overlapping peaks between files.
Stat Plot	Box plot showing the distribution of statistical significance (q-values) of sample peaks that are overlapping and non-overlapping with the reference peak file.
Precision-recall plot	Computes precision-recall curves across different peak strength thresholds. Metadata columns to be used for thresholding can be automatically inferred based on known relevant columns generated by SEACR, MACS2/3 and HOMER.
Correlation plot	Standardises all peak files by rebinning them into tiles of a user-defined width across the genome, and then compute pairwise correlation statistics between all peak files.
ChromHMM Plot	Heatmap of ChromHMM(6) annotation of peaks.
Chipseeker Plot	Bar chart of ChIPseeker(7) annotation of peaks.
Enrichment Plot	Dot plot of KEGG pathway and GO enrichment analysis of peaks.
TSS Plot	Peak frequency around (+/-3000bp) transcriptional start site.

Output

EpiCompare generates an interactive HTML report containing all results of the analysis. The exact code used to generate each section is embedded within the report (collapsed by default). The report is organised into three parts: General Metrics, Peak Overlap and Functional Annotation. All sections are easily navigated by an interactive table of contents. The General Metrics section presents information on individual peak files, including the number of peaks, percentage of peaks in blacklisted regions and non-standard chromosomes, distribution of peak width, and duplication rate of mapped fragments. The Peak Overlap section provides the frequency, percentage, statistical significance, precision-recall and correlation of overlapping and non-overlapping peaks between the sample and reference peak files. Finally, the Functional Annotation section contains the functional annotation of peaks. These are ChromHMM (Ernst & Kellis, 2017), ChIPseeker (Yu, Wang & He, 2015), enrichment analysis (KEGG pathway and GO) and the frequency of peaks around the transcriptional start site.

To demonstrate the functionalities, we used EpiCompare to contrast the profiling of open chromatin regions of human K562 cells using ATAC-seq and DNase-seq. Several of the figures included in the report can be seen in Figure 1 (see https://neurogenomics.github.io/EpiCompare/inst/report/EpiCompare_example.html for the full report). The two ATAC-seq (ENCFF558BLC and ENCFF333TAT) and DNase-seq (ENCFF274YGF and ENCFF185XRG) datasets were obtained from ENCODE (ENCODE Project Consortium, 2012). Using EpiCompare, we can see a difference between the two methods, especially in ChromHMM annotations and precision-recall plot (Figure 1d&1e).



```
EpiCompare(peakfiles = peaklist,
           genome_build = list(peakfiles="hg19",
                               reference="hg38",
                               blacklist="hg19"),
           genome_build_output = "hg19",
           blacklist = hg19_blacklist,
           picard_files = picard,
           reference = reference_peak,
           upset_plot = TRUE,
           stat_plot = TRUE,
           chromHMM_plot = TRUE,
           chromHMM_annotation = "K562",
           chipseeker_plot = TRUE,
           enrichment_plot = TRUE,
           tss_plot = TRUE,
           interact = TRUE,
           save_output = TRUE,
           output_dir = "/path/to/output")
```

EpiCompare Report

May-18-2022

EpiCompare compares epigenetic datasets for quality control and benchmarking purposes. The report consists of three sections:

1. **General Metrics:** Metrics on peaks (percentage of blacklisted and non-standard peaks, and peak widths) and fragments (duplication rate) of samples.
2. **Peak Overlap:** Percetnage and statistical significance of overlapping and non-overlapping peaks. Also includes upset plot.
3. **Functional Annotation:** Functional annotation (ChromHMM, ChIPseeker and enrichment analysis) of peaks. Also includes peak enrichment around TSS.

Input Datasets

- Reference peakfile: Dnase_ENCFF274YGF_reference
- Total of 4 peak files:

```
## [1] "File1: ATAC_ENCFF558BLC"
## [1] "File2: ATAC_ENCFF331UKL"
## [1] "File3: DNase_ENCFF185XRG"
## [1] "File4: Dnase_ENCFF274YGF_reference"
```

Figure 1. Flowchart demonstrating the use of *EpiCompare*. This example compares the open chromosome regions of human K562 cells profiled using ATAC-seq and DNase-seq. (a) Peak files are input into the master function (*EpiCompare*). (b) The function outputs an HTML report

containing all results of the analysis. (c) Upset plot showing the number of overlapping peaks between peak files. (d) ChromHMM annotation of peak files. (e) Plot showing the precision-recall score across the peak calling stringency thresholds.

Conclusion

Here, we presented EpiCompare, a Bioconductor R package for the comparison and quality control of epigenomic data. The package offers a selection of downstream analysis tools, enables processing of multiple epigenomic datasets in parallel and allows users to tailor their analyses. All of this can be executed with just one R function with minimal input from users, making the usage less demanding for those with little computational experience. Lastly, it generates a single report containing all results of the analysis, providing a simple, efficient and user-friendly way of comparing epigenomic datasets. EpiCompare will continue to be optimised and enhanced over time, with new features such as API access to thousands of peak files stored on public databases already underway.

Acknowledgements

We like to thank Sarah Marzi, Alexi Nott and Di Hu for contributing to the discussions of early results and for suggesting new functionalities of EpiCompare throughout the development process.

Funding

This work was supported by the UK Dementia Research Institute which receives its funding from UK DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK.

References

Allis, C.D. & Jenuwein, T. (2016) The molecular hallmarks of epigenetic control. *Nature reviews. Genetics*. 17 (8), 487–500.

Amemiya, H.M., Kundaje, A. & Boyle, A.P. (2019) The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific reports*. 9 (1), 9354.

Broad Institute (2019) *Picard toolkit*. https://broadinstitute.github.io/picard/.

Cazaly, E., Saad, J., Wang, W., Heckman, C., Ollikainen, M. & Tang, J. (2019) Making Sense of the Epigenome Using Data Integration Approaches. *Frontiers in pharmacology*. 10, 126.

Cheng, Y., He, C., Wang, M., Ma, X., Mo, F., Yang, S., Han, J. & Wei, X. (2019) Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. *Signal Transduction and Targeted Therapy*. 4 (1), 1–39.

Cheshire, C., charlotte-west, Bot, N.-C., Ladd, D., Fields, C., Patel, H., Deu-Pons, J., Ewels, P. & Menden, K. (2022) *nf-core/cutandrun: nf-core/cutandrun v2.0 Copper Cobra*. doi:10.5281/zenodo.6624266.

DeBerardine, M. (2022) *BRGenomics: Tools for the Efficient Analysis of High-Resolution Genomics Data*. https://mdeber.github.io.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*. 489 (7414), 57–74.

Ernst, J. & Kellis, M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nature protocols*. 12 (12), 2478–2492.

Ewels, P.A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M.U., Di Tommaso, P. & Nahnsen, S. (2020) The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology*. 38 (3), 276–278.

Hannon, E., Marzi, S.J., Schalkwyk, L.S. & Mill, J. (2019) Genetic risk variants for brain disorders are enriched in cortical H3K27ac domains. *Molecular brain*. 12 (1), 7.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., et al. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*. 12 (2), 115–121.

Lawrence, M., Gentleman, R. & Carey, V. (2009) rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* . 25 (14), 1841–1842.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. & Carey, V.J. (2013) Software for computing and annotating genomic ranges. *PLoS computational biology*. 9 (8), e1003118.

Mazzone, R., Zwergel, C., Artico, M., Taurone, S., Ralli, M., Greco, A. & Mai, A. (2019) The emerging role of epigenetics in human autoimmune disorders. *Clinical epigenetics*. 11 (1), 34.

Mehrmohamadi, M., Sepehri, M.H., Nazer, N. & Norouzi, M.R. (2021) A Comparative Overview of Epigenomic Profiling Methods. *Frontiers in cell and developmental biology*. 9, 714687.

Patel, H., Wang, C., Ewels, P., Silva, T.C., Peltzer, A., Behrens, D., Garcia, M., mashehu, Rotholandus, Haglund, S. & Kretzschmar, W. (2021) *nf-core/chipseq: nf-core/chipseq v1.2.2 - Rusty Mole*. doi:10.5281/zenodo.4711243.

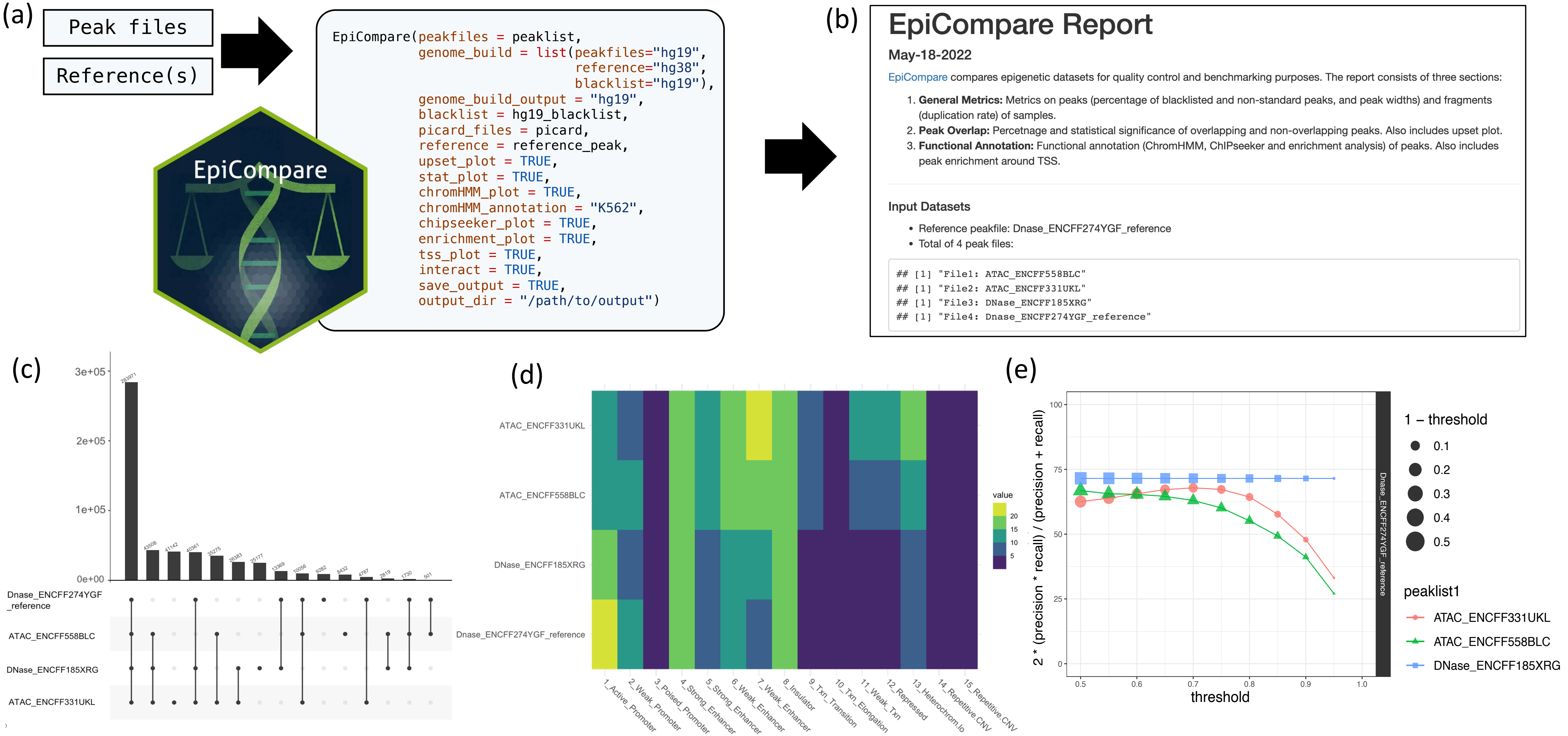
R Core Team (2021) *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing. https://www.R-project.org/.

Roussos, P., Mitchell, A.C., Voloudakis, G., Fullard, J.F., Pothula, V.M., et al. (2014) A role for noncoding variation in schizophrenia. *Cell reports*. 9 (4), 1417–1429.

RStudio Team (2022) RStudio: integrated development environment for R. *Boston, MA*. http://www.rstudio.com/.

Samb, R., Khadraoui, K., Belleau, P., Deschênes, A., Lakhal-Chaieb, L. & Droit, A. (2015) Using informative Multinomial-Dirichlet prior in a t-mixture with reversible jump estimation of nucleosome positions for genome-wide profiling. *Statistical applications in genetics and molecular biology*. 14 (6), 517–532.

Yu, G., Wang, L.-G. & He, Q.-Y. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* . 31 (14), 2382–2383.



```
EpiCompare(peakfiles = peaklist,
           genome_build = list(peakfiles="hg19",
                               reference="hg38",
                               blacklist="hg19"),
           genome_build_output = "hg19",
           blacklist = hg19_blacklist,
           picard_files = picard,
           reference = reference_peak,
           upset_plot = TRUE,
           stat_plot = TRUE,
           chromHMM_plot = TRUE,
           chromHMM_annotation = "K562",
           chipseeker_plot = TRUE,
           enrichment_plot = TRUE,
           tss_plot = TRUE,
           interact = TRUE,
           save_output = TRUE,
           output_dir = "/path/to/output")
```

EpiCompare Report

May-18-2022

EpiCompare compares epigenetic datasets for quality control and benchmarking purposes. The report consists of three sections:

1. **General Metrics:** Metrics on peaks (percentage of blacklisted and non-standard peaks, and peak widths) and fragments (duplication rate) of samples.
2. **Peak Overlap:** Percetnage and statistical significance of overlapping and non-overlapping peaks. Also includes upset plot.
3. **Functional Annotation:** Functional annotation (ChromHMM, ChIPseeker and enrichment analysis) of peaks. Also includes peak enrichment around TSS.

Input Datasets

- Reference peakfile: Dnase_ENCFF274YGF_reference
- Total of 4 peak files:

```
## [1] "File1: ATAC_ENCFF558BLC"
## [1] "File2: ATAC_ENCFF331UKL"
## [1] "File3: DNase_ENCFF185XRG"
## [1] "File4: Dnase_ENCFF274YGF_reference"
```