Prenet: Predictive network from ATAC-SEQ data

Nazmus Salehin, Patrick Tam, Pierre Osteil

PRENET: PREDICTIVE NETWORK FROM ATAC-SEQ DATA

NAZMUS SALEHIN[1,2], PATRICK P.L. TAM[1,2], PIERRE OSTEIL[1,2]

[1]*Embryology Unit, Children's Medical Research Institute, The University of Sydney, Westmead, NSW 2145, Australia*

[2]*School of Medical Sciences, Sydney Medical School, University of Sydney, NSW 2006, Australia*

Assays for transposase-accessible chromatin sequencing (ATAC-seq) provides an innovative approach to study chromatin status in multiple cell types. Moreover, it is also possible to efficiently extract differentially accessible chromatin (DACs) regions by using state-of-the-art algorithms (e.g. DESeq2) to predict gene activity in specific samples. Furthermore, it has recently been shown that small dips in sequencing peaks can be attributed to the binding of transcription factors. These dips, also known as footprints, can be used to identify trans-regulating interactions leading to gene expression. Current protocols used to identify footprints (e.g. pyDNAse and HINT-ATAC) have shown limitations resulting in the discovery of many false positive footprints. We generated a novel approach to identify genuine footprints within any given ATAC-seq dataset.

Herein, we developed a new pipeline embedding DACs together with *bona fide* footprints resulting in the generation of a Predictive gene regulatory Network (PreNet) simply from ATAC-seq data. We further demonstrated that PreNet can be used to unveil meaningful molecular regulatory pathways in a given cell type.

Keywords: ATAC-seq, Footprints, Gene Regulatory Network

1. Introduction

Assays for transposase-accessible chromatin sequencing (ATAC-seq) is an efficient and robust technique requiring very little starting material (as few as 50,000 cells), making it a reliable tool for studying restricted biological systems (e.g. embryos).[1] Generally, ATAC-seq is used to determine accessible chromatin regions in cells. Recently, it has been shown to be a promising tool to detect transcription factor binding sites (TFBSs) by screening footprints within peaks called from ATAC-seq data.[2] Footprints are small dips (10-20 bp) observed in read pile-ups (or peaks) and are indicative of reduced transposase activity that may be associated with TF occupancy.

Combined with RNA sequencing (RNA-seq), these data allow scientists to build gene regulatory networks (GRN) at an unprecedented level of complexity. Nonetheless, RNA-seq data can be difficult to generate because of sample scarcity. This limitation prompted the development of a pipeline that can predict multiple levels of gene regulation only from ATAC-seq data to obtain a comprehensive predictive GRN without further sample preparation (Figure 1A).

Current ATAC-seq pipelines do not extract significantly differentially accessible chromatin (DACs) regions, which are important to determine remodelling of DNA tertiary structure between cell types. For instance, ATAC2GRN[3] only provides open chromatin regions based on MACS2 program (Figure 1B - blue area). DiffBind[4] extracts DACs but not specifically located to a promoter and TSS of a given gene (Figure1B – red area). Finally, DAStk[5] can be used to extract footprints using HINT-ATAC[6] program but not within a DACs (Figure 1B – yellow area). As we can see on Figure 1B, altogether, the three pipelines aforementioned do not lead to a comprehensive picture of the chromatin status within cells.

Thus, we decided to combine both the extraction of DACs specifically located to a promoter (PROM-DACs), predictive of a potentially active gene, together with their associated footprints, predictive of a TF bound to the promoter, to build a two-level Predictive Network (PreNet): TFs regulating genes expression at the promoter.

Moreover, PreNet comes with a solution to calculate Fold Changes (FC) and p.values associated to DACs between cell types. In addition, we describe that current footprinting tools, such as pyDNAse[7] or HINT-ATAC,[6] produce an important number of false positive footprints which can lead to misinterpretation, hence the devaluation of the data. To prevent this, we developed a False Discovery Rate (FDR) method to eliminate these artefact footprints. Combining both

approaches (DACs FC and p.values together with FPs FDR) into one pipeline allowed for the inferring of a more reliable predictive GRN from ATAC-seq data.

To illustrate the utility of PreNet, we used a well-known differentiation model from mouse Embryonic Stem Cells (mESCs) to the Definitive Endoderm (DE).[8] We further compared footprints analysis to ChIP-seq data in similar cell types and benchmark the new FDR filtering against ATAC2GRN by comparing Positive Predictive Value (PPV). Runtimes against the only available pipeline achieving a similar outcome (ATAC2GRN) further illustrate the increased power of the PreNet pipeline. PreNet can be used with any given set of ATAC-seq data and its respective reference genome. On top of providing a program requiring only ATAC-seq data, we improved the statistical analysis for genuine DACs and FPs leading to an increase of predictiveness as demonstrated here.

2. Methods

The PreNet pipeline is summarized in Figure 1B.

2.1. Availability of data and materials

PreNet can be downloaded from the project code repository https://github.com/ChildrensMedicalResearchInstitute/PreNet. PreNet is provided as a series of snakemake pipelines. The initial set up for the pipeline requires the tools found in Table 1. Set up for these tools must be completed in the *config.snakemake* file. The default configuration assumes the tools are present in the current PATH but can be modified to point directly to the tool location.

For input, the snakemake pipeline requires the following (placed in *config.snakemake*):

1) The trimmed paired-end sequencing files for an ATAC-seq experiment in gzipped FASTQ format, within a "FASTQ" directory.
2) The bowtie2[9] index for the genome of interest. These may be downloaded from NCBI Genbank or manually created.
3) A list of chromosomes to keep for final analyses along with the size of each of the chromosomes (*xxxx.chrom.sizes*). The default list contains the non-mitochondrial chromosomes for the mouse genome (chr 1-20, chr X, chr Y). The *chrom.sizes* file can be found from ENCODE, UCSC or made custom.
4) A blacklist of regions to exclude in BED format. The mouse blacklist (ENCFF547MET.bed) can be found on ENCODE and a similar blacklist exists for the human genome.
5) The regions of interest in SAF format. In this case, promoter regions based on the mm10 reference genome are provided.
6) Transcription factor position-weighted matrix file in MEME format for FIMO[10] allocation. These can be downloaded from a transcription factor database of choice.

Table 1. Software requirements for PreNet

Tool	Version tested	Website
Samtools [11]	1.8	https://www.htslib.org/
Bowtie2 [9]	2.2.6	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Sambamba [12]	0.6.7	https://lomereiter.github.io/sambamba/
MACS2 [13]	2.1.1.20160309	https://taoliu.github.io/MACS/
assign_multimappers.py	Accessed April 3, 2018	https://github.com/kundajelab/atac_dnase_pipelines/commits/master/utils/assign_multimappers.py
featureCounts [14]	1.5.3	http://subread.sourceforge.net/
deeptools [15]	2.5.4	https://github.com/deeptools/deepTools
UCSC kentUtils [16]	4	http://genome.ucsc.edu/
R Statistical Software	3.5.1	https://www.R-project.org/
DESeq2 [17]	1.22.1	https://bioconductor.org/packages/release/bioc/html/DESeq2.html

Tidyverse libraries	1.2.1	https://cran.r-project.org/web/packages/tidyverse/index.html
HINT-ATAC [6]	0.21.1	http://www.regulatory-genomics.org/hint/
FIMO [10]	4.11.2	http://meme-suite.org/doc/fimo.html

2.2. Raw data extraction and alignment to reference genome

The pipeline can be started from either raw reads or aligned and filtered files. Within this paper, data from Simon and colleagues[8] (GSE94249) were downloaded from the Gene Expression Omnibus repository and analysed completely using the snakemake pipeline. Paired-end reads were aligned to the GRCm38/mm10 *M. musculus* reference genome using bowtie2[9] (-k 4 -X 2000 –local -mm), allowing for a maximum of 4 multimapping locations per read. Multimapping reads were then assigned randomly using assign_multimappers.py from the ENCODE project.[18] From the aligned data, unpaired and low-quality reads were excluded using samtools,[11] deduplicated using sambamba[12] and filtered to remove mitochondrial reads.

2.3. Differential Open Chromatin regions analysis of Promoter region

Aligned pairs were converted to pseudo-single ended reads and only the 5'-end was used for differential analysis. This represents the site of Tn5 transposase activity. Regions of open chromatin were identified using MACS2[13] (-q 0.05 --no-model --shift -100 --extsize 200) on pooled replicates. Tn5 transposase events within the promoter (-5 kb to TSS) and TSS (-1 kb to +l kb) regions (Figure 2A) were estimated using the number of 5'-ends within the region. This was achieved using featureCounts[14] (--read2pos 5) and raw counts were imported into R for analysis with DESeq2.[17] Differentially accessible regions were visualised using deeptools2 (Figure 2C) .[15]

2.4. Footprinting analysis

Aligned reads were combined to give one file per biological condition. *De novo* footprint analysis in promoter regions (-5kb to TSS) was performed using HINT-ATAC.[6] Sequences from identified footprints, extracted using bedtools, were assigned to a transcription factor using FIMO[10] from the MEME Suite, utilising the JASPAR 2018 vertebrate database[19] as a reference for transcription factor (TF) position weight matrices. Regions of predicted TF activity were assigned based on the initial promoter SAF file.

2.5. False discovery rate filtering for footprints

Footprints were called using HINT-ATAC (Figure 3A) on all promoter regions and subsequently segregated into those falling either within or outside the called peaks (Figure 3C). HINT-ATAC footprint scores were placed in descending order based on the footprint score. The footprints were filtered based on the proportion of footprints found in inaccessible regions of the genome, outside of ATAC-seq peaks (Figure 3B). For an FDR value of τ, the subset of top footprints is returned such that the proportion of footprints outside of ATAC-seq peaks are equal to τ; these are subsequently filtered for footprints that lie within peaks only. To achieve this, FPs lying outside peaks ("outP") were allocated as false positive footprints and a footprint score cut-off was generated by ranking all FPs, then selecting the lowest score that would provide a final list containing less than 1% of FPs located outside of a peak (in other words, less than 1% of false positive FPs). The goal here is to eliminate FPs in a peak that have a lower score than the 1% FPs detected from a region outside of a peak as they are not more significant, hence meaningful, than a FP found between two single isolated reads.

2.6. Benchmarking

Comparison against ATAC2GRN[3] was performed using version 2 of the optimised ATAC2GRN bash pipeline for paired-end experiments. The shell script was modified to run on the benchmarking system. Subsequent allocation of promoter footprints to a transcription factor was done using the PreNet pipeline as ATAC2GRN does not provide an allocation pipeline (only genomic location). The runtime benchmark was performed on a server running Centos 7, Linux kernel version 4.6 and 24 threads running at 2.90GHz. Where applicable, 64-bit versions of the programs have been used. Running times are given as user time as measured by the shell command *time.*

2.7. ChIP-seq validation

Where peak files are available for pipelines, no further analysis was carried out before comparison (GSM1288315, converted to GRCm38/mm10 *M. musculus* genome using liftOver[16]). Otherwise (GSM1782914), reads were aligned to GRCm38/mm10 genome using Burrow-Wheeler Aligner[20] and filtered to remove duplicates, unaligned and blacklist region reads. Peaks were called using MACS2 (-q 0.05). Peak files were then filtered to only include peaks within promoter regions (-5kb to TSS). Regions where transcription factors were assigned to footprints via PreNet were compared to ChIP-seq peaks to determine the Positive Predictive Value (PPV) of PreNet pre- and post-FDR filtering.

2.8. Cytoscape GRN generation

Final output from the snakemake pipeline, interactions between FPs filtered by FDR < 0.01 and target DACs, were imported into Cytoscape[21] compatible file format for gene network visualisation. Each GRN comprises of FPs located within a DAC with an open TSS. Nodes with no interactions were excluded from the GRN. Gene ontology search on developmental pathway was conducted using Panther v14.[22]

2.9. Scalability of PreNet

The snakemake pipeline allow for the allocation of threads for the respective processes. As defaults, threaded processes are allocated 4 threads; therefore, for the processing of 6 files, as in the paper, the pipeline can utilise up to 24 cores. This can be scaled up or down as required. The pipeline allows for the comparison of any number of biological replicates for 2 conditions.

3. Results

3.1. DACs at promoters are predictive of active genes.

We analysed bulk ATAC-seq dataset from Simon and colleagues[8] on mESC (n=3) differentiation toward DE (n=3). Firstly, we evaluated whether the promoter regions (Figure 2A and 2B) were differentially opened or closed between the two cell states for both strands of DNA (annotated "+" or "-") (Figure 2C); these regions are annotated as PROM-DACs. We restricted our analysis to promoter regions as there is still strong debate on how to properly annotate enhancer targets (e.g. HiC, methylation…) and how they regulate gene expression. We uncovered 1041 differential opening and 1489 differential closing of PROM-DACs during the mESC to DE differentiation (Figure 2B). This method leads to the detection of false positive regions, such as DACs between two samples that do not present proper enrichment - peaks as defined by MACS2[13] -, or, in other words, DACs between two closed chromatin regions. To segregate positive regions, we filtered out PROM-DACs on the criterion that at least one of the cell types should display an open chromatin region (i.e., a "peak") based on the threshold applied by MACS2 algorithm. From this analysis, we excluded 28% (422/1489) and 46% (475/1041) of regions in ESC and DE samples respectively (Figure 2B – grey circle), indicating that a large proportion of the DACs was firstly selected from solely closed chromatin regions in both samples. Genuine PROM-DACs were plotted into a heatmap (Figure 2C). As we can see, selected regions display differential accessibility between both samples after applying our filtering method.

To enable PreNet to select genuine DACs, we then analysed regions surrounding the TSS (+/- 1kb) and selected for those with an open TSS, hence with the ability of being transcribed, using MACS2 peak calling software. We assumed that an active transcription site requires an accessible TSS (TSS-DACs, Figure 2A). This new parameter allowed us to further eliminate 61 (4%) and 182 targets (17%) in the ESC and DE samples, respectively (Figure 2B – light green circles). Finally, we calculated the TSS-DAC log2 fold changes (log2FC) from raw reads between the two cell types and plotted it against the PROM-DAC log2FC (Figure 2D). As expected, the correlation between TSS and PROM region log2FC in the two cell states was higher than random ($r = 0.65$) and significant ($p < 2.2\times10^{-16}$) validating that accessibility of the promoters is linked to that of TSSs. This result is also visible in Figure 2C with most PROM-DACs having an open TSS (yellow regions surrounding the TSS). Conversely, very few TSSs were inaccessible when the promoter region was open (5 in the ESC and 4 in the DE cells were further removed from the final list) (Figure 2 – blue circles). PreNet filtering led to the selection of 1001 DACs in ESCs and 855 in DE (Figure 2B and 2D).

After annotating the DACs regions to genes, we discovered that pluripotency genes were more accessible in mESCs than in DE cells (e.g. *Nr0b1, Dppa2, Dppa4, Zfp42*, *Prdm5* and *Klf5*) whilst DE cells displayed a reverse pattern for mesendodermal genes (e.g. *Smad3, Fgf8, Pdx1,* and *Mesp1*) (Figure 2D) which are expected to be expressed in both

cell types. Example of DACs selected by PreNet are shown in gene tracks for *Nr0b1, Dppa2* and *Smad3* (Figure 2E). This result confirms that PreNet selection criteria retain genuine cell-type specific chromatin accessible regions.

3.2. PreNet Pipeline Improve the Quality of Footprints Prediction.

To unveil trans-regulatory elements binding on open chromatin regions within the promoter, we extracted FPs from promoter regions using HINT-ATAC (see Methods). Predicted protein-DNA interactions were extracted (2,941,546 for the ESC and 2,978,511 for the DE) (Figure 3A). After investigation we discovered that the majority of these FPs were false positive for two reasons: 1) HINT-ATAC (used in DAStk and ATAC2GRN) does not make the distinction between a proper footprint found in a DACs and dips observed at the edge of a peak or two single-reads separated by a small region (Figure 3B); 2) HINT-ATAC scores FPs based on the flanking reads. That scoring method comes with flaws as FPs could be assigned to closed chromatin regions. More importantly, there is no method allowing to set up a threshold to select genuine FPs within a peak. Indeed, FP are scored but there is no probability associated to that score to predict the presence of these FPs on a promoter in a given cell type.

To overcome these issues, a more stringent method to select FPs was designed based on two criteria:

1) A true positive FP is, by definition, a feature observable within an accessible region (Figure 3B). To visualise the impact of our selection criteria on the FPs list, incremental top ranking FPs against scores were plotted (2,941,546 for the ESC - ESC-FPs are shown as an example on Figure 3A). Then FPs in peaks ("inP") were represented on a similar plot (Figure 3C). Only 482,979 FPs, corresponding to 16% of all FPs previously found in mESC (in DE, 714,562 FPs corresponding to 24% of total FPs (Figure 3G)) are detected within a peak. Surprisingly, some of the false positive targets display either a very high score or a score close to 0 (Figure 3D). Both these results raise some concerns regarding the reliability of using HINT-ATAC scoring as the sole method for FPs analysis (Strategy used in ATAC2GRN).
2) Applying a False Discovery Rate (FDR) calculations of FPs using inP versus outP as true positive versus false positive qualification, in order to select in a robust manner more genuine FPs. To image our method, FDR was plotted against cumulative selection of FP (first step = 100 top score FPs, second step = 200 top score FPs, etc…) (Figure 3E) (see Methods). According to our observation in Figure 3D, we can see that in the 100 top score FPs 80% are false positive, similarly to when all FPs are selected (step containing all FPs, n= 2,941,547 in Figure 3E). We selected a step containing a maximum of FPs while retaining less than 1% of FPs. By doing so, 213,271 FPs were shortlisted corresponding to 7% of total ESC-FPs originally found using HINT-ATAC (and 8% for DE-FPs) (Figure 3F and G).
3) Finally, we filtered out the outP's FPs from the FDR corrected list (Figure 3F).

The PreNet tool is believed to generate a more genuine list of FPs from a particular ATAC-seq dataset that remains to be validated.

3.3. Validation of PreNet and Genuine GRN Construction

Targets of footprints allocated to SMAD2/3 in DE from PreNet were compared to those found from chromatin immunoprecipitation assay followed by deep sequencing (ChIP-seq) from two independent groups: Wang 2017 (GSM1782914)[23] and Yoon 2015 (GSM1288315)[24] , also in the DE. We compared the putative FP targets associated to SMAD2/3 before and after applying FDR correction to the list of targets from the ChIP-seq experiments giving a Positive Predictive Value (PPV) that is the proportion of SMAD2/3 allocated footprints that are also found in at least one of the ChIP-seq datasets. Prior to FDR correction, 747 of the 2,794 SMAD2/3 allocated footprints overlap with the ChIP-seq dataset, so a PPV of 26.7%. After applying FDR correction, the PPV climbs to 41.7% (584 out of 1,401 footprints) (Figure 3H). According to our expectation, PreNet leads to a better prediction of promoter targets physically bound by SMAD2/3 than using only HINT-ATAC scoring methods.

To further validate the predictiveness of our tool, a GRN was built linking PROM-DACs to FPs across the two cell states (Figure 2C). Only FPs binding PROM-DACs were plotted. We found that the ESC GRN retained 502 genes, which were enriched for "LIF stimulation response", gene set responsible for mESC pluripotency maintenance (Figure 3E). On the other hand, DE cells showed a more restricted network (n = 112 genes) with an enriched gene set linked to "liver development", an endoderm derivative (Figure 3F).

These validation steps (increased PPV and enrichment for meaningful GOs) confirmed that the PROM-DACs selection tools combined with an FP FDR correction embedded in PreNet led to the generation of a more genuine GRN.

3.4. Benchmarking PreNet Against Available Pipelines Shows Increased Efficiency.

To increase the impact of PreNet, we performed a run time and PPV comparison as the main criteria for assessing the increased power of PreNet over existing tools using the same datasets. We decided to compare its efficacy against the only available pipeline, ATAC2GRN.[3]

The user runtime for PreNet was 80 CPU-hours. This process takes 6 FASTQ files as input and provides both differentially accessible regions as well as footprints within the promoter. In comparison, ATAC2GRN took 38.8 and 47.5 CPU-hours for each experiment. Allocating the footprints from the whole genome to a transcription factor using the single-threaded FIMO was stopped at 48h real time. Subset footprints found within the promoter region by ATAC2GRN were allocated using FIMO; this process took 53.22 min and 53.87 min for each condition in addition to the ATAC2GRN pipeline running time. In summary, ATAC2GRN is significantly slower than PreNet.

The predictive value of PreNet was benchmarked using SMAD2/3 ChIP-seq data. ATAC2GRN combined with FIMO allocation resulted in a PPV of 33.3%, wherein 688 out of a total of 2,063 footprints allocated to SMAD2/3 by FIMO overlapped with at least one of the ChIP-seq datasets (Figure 3K). This result is higher than the unfiltered PPV for PreNet (26.7%) but less efficient than FDR correction (41.7%).

Overall, the PreNet pipeline is more efficient at predicting *bona fide* FPs from ATAC-seq data.

4. Discussion

Current ATAC-seq analysis pipelines include strong biases leading to annotations of false positive hits for accessible chromatin and potential transcription factors binding on promoter regions. We attempted to solve both issues by firstly restricting the analysis to the promoter regions and intersecting accessible regions with the accessibility of TSS. This analysis pinpointed a close correlation between the TSS and promoter accessibility. We used this filtering to correct for regions that do not have an accessible starting site for transcription, which corresponded to a large proportion of the dataset.

Secondly, we developed a novel approach to improve the specificity of footprints analyses by removing false positive hits using scoring method combined with False Discovery Rate filtering. Then, both gene sets were combined to generate a GRN for each cell type. Genes that could be linked to form a GRN display enrichment for gene ontologies generally associated with each cell type. Finally, we demonstrated that our approach leads to a more robust positive predictive score when compared to the only available tool that performs a similar analysis: ATAC2GRN.

Although, only predictive GRN can be inferred from our pipeline, further validation through wet lab techniques are required, but we believe this approach will help in extracting important information from ATAC-seq datasets by refining putative targets to further validate.

We herein described an innovative method to analyse ATAC-seq data comprehensively while refining the obtained gene sets. This allowed us to extract *trans*-regulation of gene expression using only a single sequencing method. PreNet could be applied to any ATAC-seq dataset and potentially to single-cell data (not tested here) providing they are compared between clusters of cells. PreNet is believed to enhance ATAC-seq analysis power.

Acknowledgments

The authors acknowledge the University of Sydney HPC service at The University of Sydney for providing high performance computing resources that have contributed to the research results reported within this paper.

Funding bodies: Our work was supported by the Australian Research Council (DP160103651), NS is supported by the Australian Postgraduate Award from University of Sydney and CMRI Scholarship, PO is funded by the Sir Norman Greg fellowship and PPLT is a NHMRC Senior Principal Research Fellow (Grant ID 1110751).

Declaration of interest

The authors declare no competing interests

References

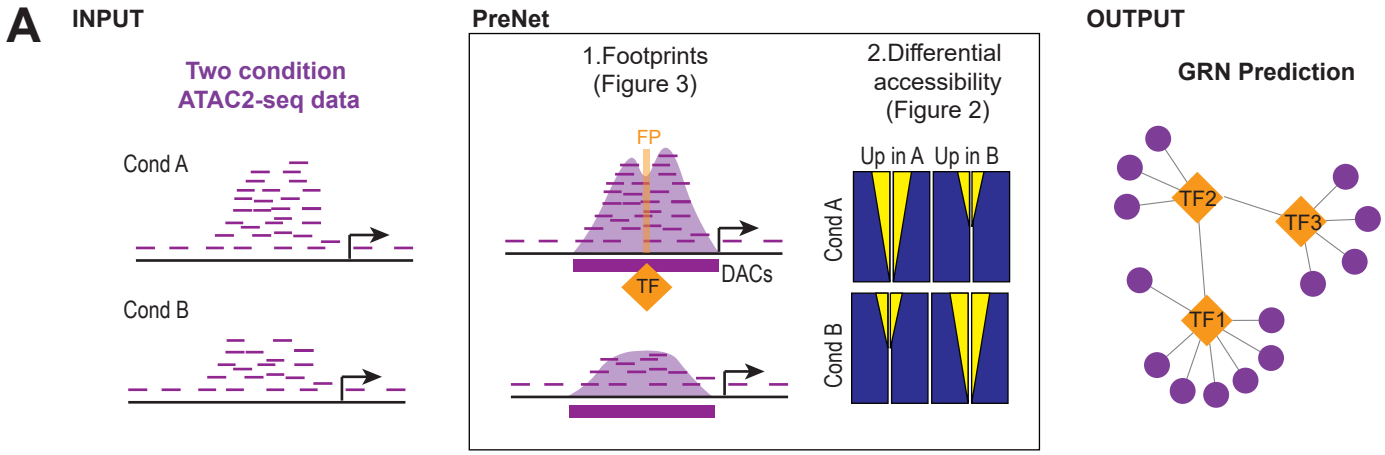
1. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–8 (2013).
2. Sung, M. H., Baek, S. & Hager, G. L. Genome-wide footprinting: Ready for prime time? *Nat. Methods* **13**, 222–228 (2016).
3. Pranzatelli, T. J. F., Michael, D. G. & Chiorini, J. A. ATAC2GRN : optimized ATAC-seq and DNase1-seq pipelines for rapid and accurate genome regulatory network inference. *BMC Genomics* **19:563**, 1–13 (2018).
4. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
5. Tripodi, I. J., Allen, M. A. & Dowell, R. D. Detecting differential transcription factor activity from ATAC-Seq data. *Molecules* **23**, 1–11 (2018).
6. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* **20**, 1–21 (2019).
7. Piper, J. *et al.* Wellington-bootstrap : differential DNase-seq footprinting identifies cell-type determining transcription factors. *BMC Genomics* **16:1000**, 1–8 (2015).
8. Simon, C. S. *et al.* Functional characterisation of cis -regulatory elements governing dynamic Eomes expression in the early mouse embryo. **144**, 1249–1260 (2017).
9. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–9 (2012).
10. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
11. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
12. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Genome analysis Sambamba : fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2017).
13. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
14. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
15. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
16. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
17. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15:550**, 1–21 (2014).
18. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
19. Khan, A. *et al.* JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).
20. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
21. Shannon, P. *et al.* Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
22. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).
23. Wang, Q. *et al.* The p53 Family Coordinates Wnt and Nodal Inputs in Mesendodermal Differentiation of Embryonic Stem Cells. *Cell Stem Cell* **20**, 70–86 (2017).
24. Yoon, S. J., Foley, J. W. & Baker, J. C. HEB associates with PRC2 and SMAD2/3 to regulate developmental fates. *Nat. Commun.* **6**, 1–12 (2015).

Figure Legend

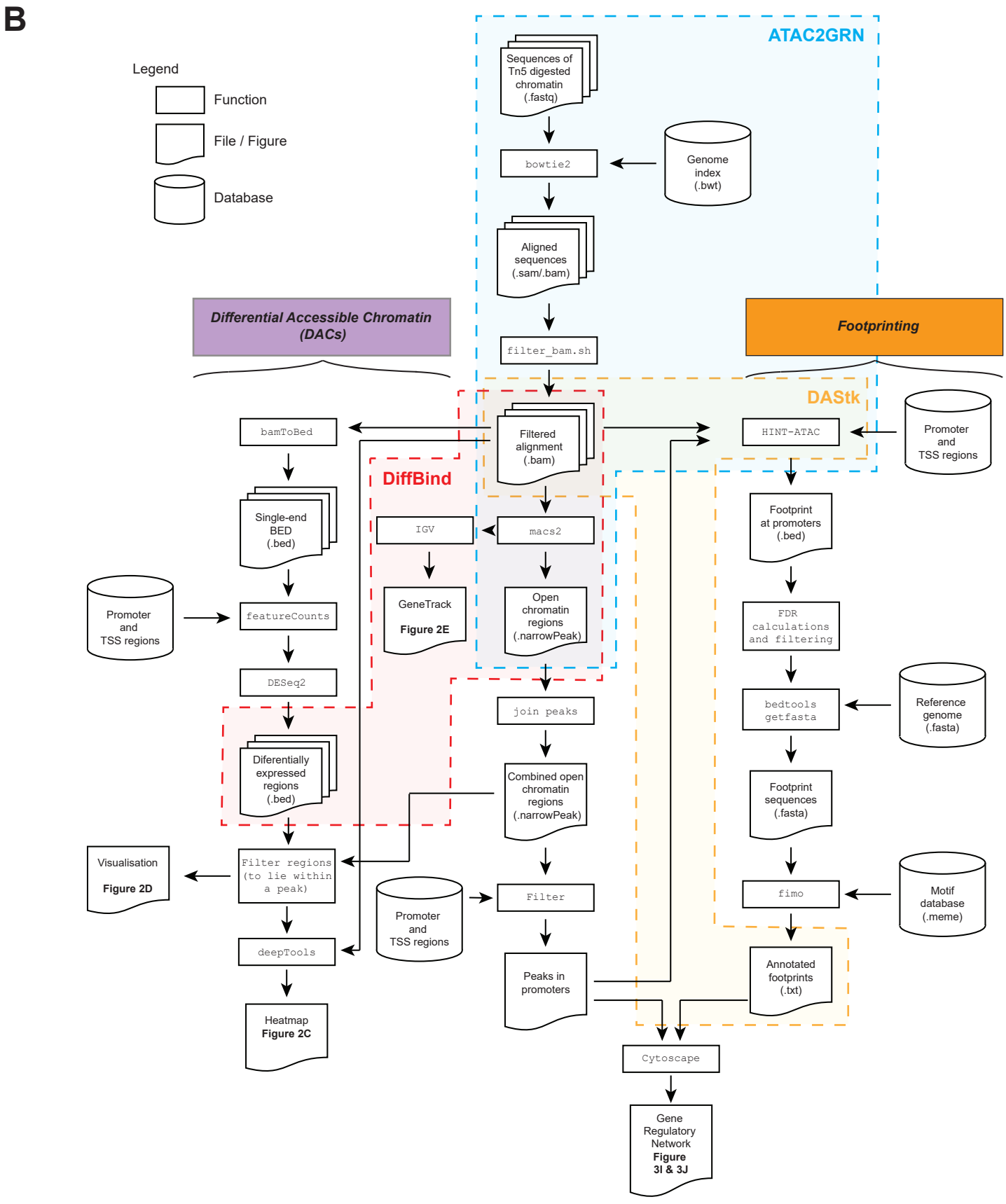
Figure 1:
A) Schematic diagram of the PreNet pipeline. **B)** Flowchart of the PreNet pipeline. Dashed area represents the part of PreNet pipeline used in other available software (ATAC2GRN (blue), DiffBind (red) and DAStk (yellow)). Only, ATAC2GRN was used to benchmark PreNet. DAC: Differentially Accessible Chromatin, FP: Footprint, TSS: Transcription Start Site

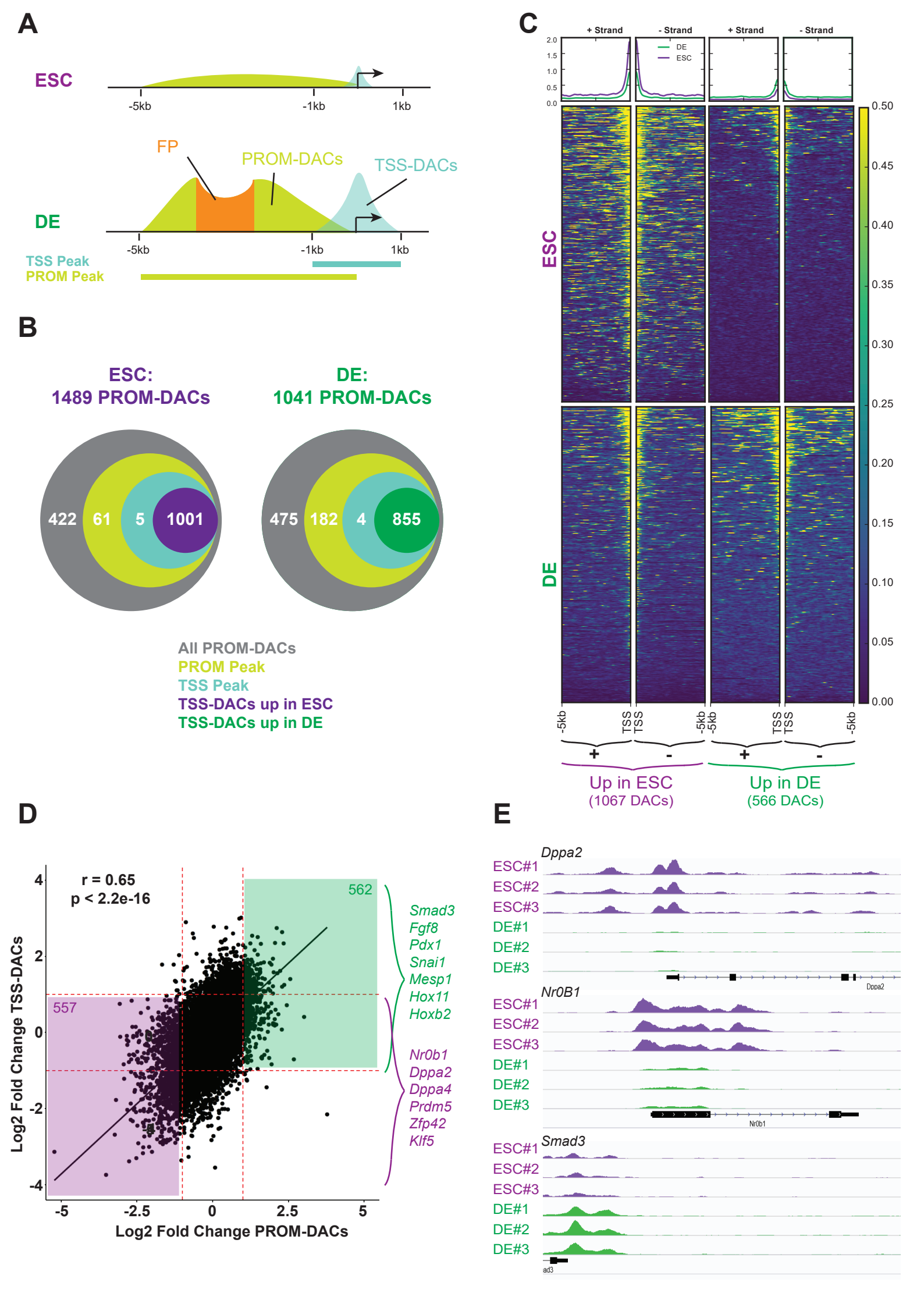
Figure 2:
A) Diagram showing annotations of the different chromatin regions analysed using this pipeline. A region opening during differentiation was drawn as an example. **B)** Venn Diagram representing incremental selection process. Grey: all PROM-DACs; light-green: PROM-DACs within a peak; light-blue: PROM-DACs having an accessible TSS; purple: PROM-DACs not having an accessible TSS in ESC; green: PROM-DACs not having an accessible TSS in DE; **C)** DACs plots showing open regions in ESC compared to DE cells (left panel) and in DE compared to ESC (right panel). **D)** Correlation plot between Log2 fold change in PROM-DACs and TSS-DACs in both cell types. **E)** Gene track plots of two DAC regions in ESC *Dppa2* and *Nr0b1* (*Dax1*) and one in DE cells, *Smad3*. DAC: Differentially Accessible Chromatin, DE: Definitive Endoderm, ESC: Embryonic Stem Cells, FP: footprints, PROM: Promoter, TSS: Transcription Start Site.

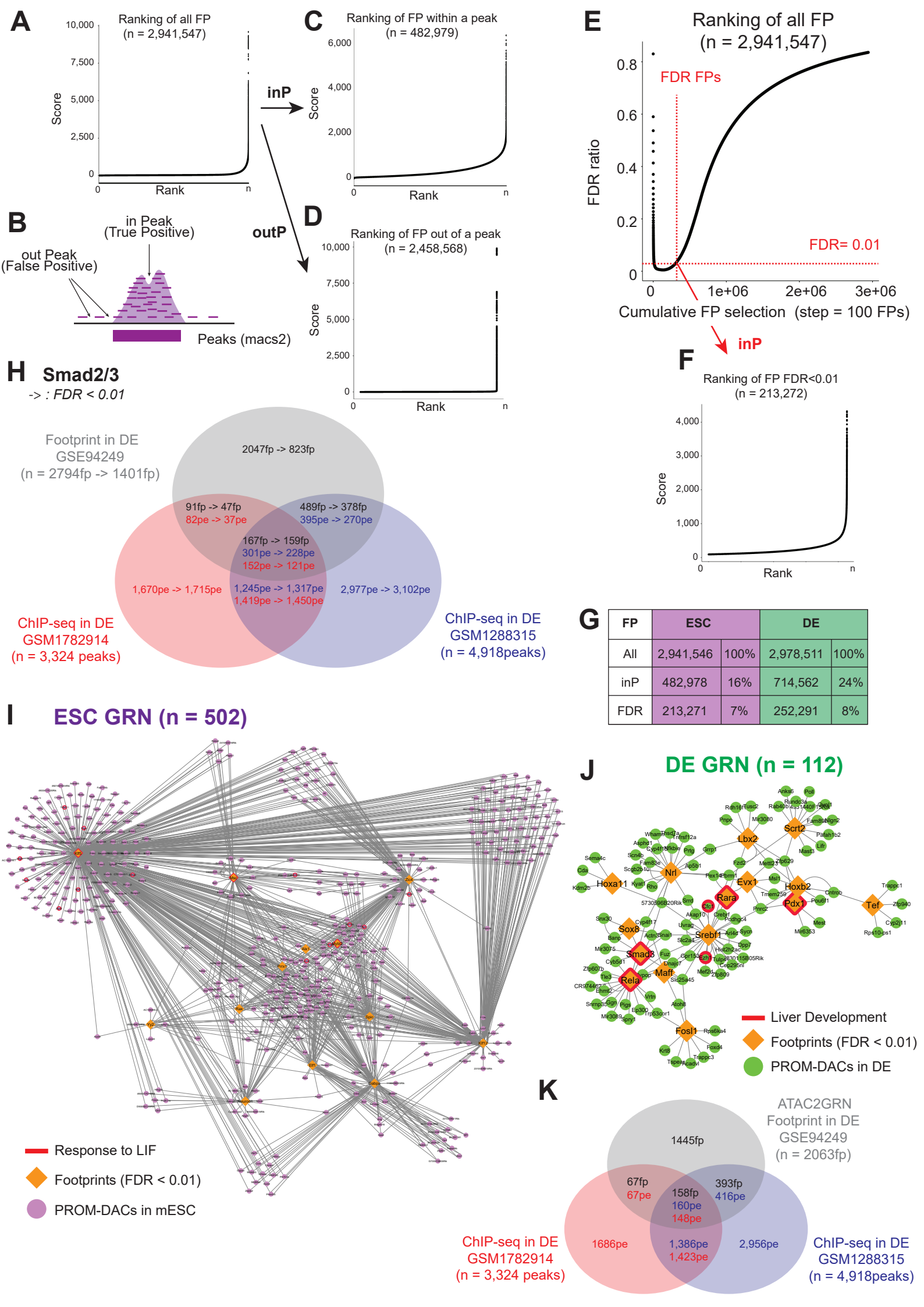
Figure 3:
A) All footprints from HINT-ATAC algorithms ranked based on their score in ESC as an example. **B)** Diagram illustrating annotation of footprints using in the PreNet pipeline. True positives are the footprints found within a peak. False positives, as described herein, are found within closed regions (non-peak-called region) or at the edge of a peak. **C)** Footprints falling within a peak (top panel) and **D)** outside a peak (bottom panel) with a maximum score around 10,000. **E)** False Discovery Rate (%) in cumulative selection of Footprints (Step = 100 FPs). Red dashed lines indicate the threshold used in our study where we selected the highest cumulative step containing a maximum of footprints with less than 1% FPR. **F)** Footprints after FDR discovery showing a maximum score just above 4,000. **G)** Summary table of footprints after applying peak allocation and FDR corrections. **H)** Venn diagram showing intersection of targets of Smad2/3 compared to two ChIP-seq datasets before and after filtering (arrow-head). Black colour indicates FDR correction in footprints selection. Red colour FDR correction overlap with Smad2/3 ChIP-seq data GSM1288315. Blue colour FDR correction overlap with Smad2/3 ChIP-seq data GSM1782914. **I)** GRN generated from ESC data. Red outline highlights genes related to “Response to LIF”. **J)** GRN generated from DE data. Red outline highlights genes related to “Liver development”. **K)** Venn diagram of ATAC2GRN footprints for comparison to that of PreNet (H). DE: Definitive Endoderm, ESC: Embryonic Stem Cells, FDR: False Discovery Rate, fp: footprints, GRN: Gene Regulatory Network, inP: in a peak, outP: oustide a peak, pe: peaks


A
INPUT
Two condition
ATAC2-seq data
Cond A
Cond B
PreNet
1.Footprints
(Figure 3)
FP
TF
DACs
2.Differential
accessibility
(Figure 2)
Up in A
Up in B
Cond A
Cond B
OUTPUT
GRN Prediction
TF2
TF3
TF1

(Figure 1B)


B
Legend
Function
File / Figure
Database
ATAC2GRN
Sequences of Tn5 digested chromatin (.fastq)
bowtie2
Genome index (.bwt)
Aligned sequences (.sam/.bam)
filter_bam.sh
Differential Accessible Chromatin (DACs)
Footprinting
DAStk
DiffBind
bamToBed
Filtered alignment (.bam)
HINT-ATAC
Promoter and TSS regions
Single-end BED (.bed)
IGV
macs2
Footprint at promoters (.bed)
GeneTrack
Figure 2E
Open chromatin regions (.narrowPeak)
FDR calculations and filtering
Promoter and TSS regions
featureCounts
DESeq2
join peaks
bedtools getfasta
Reference genome (.fasta)
Diferentially expressed regions (.bed)
Combined open chromatin regions (.narrowPeak)
Footprint sequences (.fasta)
Visualisation
Figure 2D
Filter regions (to lie within a peak)
Filter
fimo
Motif database (.meme)
Promoter and TSS regions
deepTools
Peaks in promoters
Annotated footprints (.txt)
Heatmap
Figure 2C
Cytoscape
Gene Regulatory Network
Figure
3I & 3J


A
ESC
-5kb
-1kb
1kb
FP
PROM-DACs
TSS-DACs
DE
TSS Peak
PROM Peak
B
ESC:
1489 PROM-DACs
DE:
1041 PROM-DACs
422
61
5
1001
475
182
4
855
All PROM-DACs
PROM Peak
TSS Peak
TSS-DACs up in ESC
TSS-DACs up in DE
C
+ Strand
- Strand
+ Strand
- Strand
DE
ESC
ESC
DE
-5kb
TSS
TSS
-5kb
Up in ESC
(1067 DACs)
Up in DE
(566 DACs)
D
r = 0.65
p < 2.2e-16
562
557
Log2 Fold Change TSS-DACs
Log2 Fold Change PROM-DACs
Smad3
Fgf8
Pdx1
Snai1
Mesp1
Hox11
Hoxb2
Nr0b1
Dppa2
Dppa4
Prdm5
Zfp42
Klf5
E
Dppa2
Nr0B1
Smad3
ESC#1
ESC#2
ESC#3
DE#1
DE#2
DE#3



FP	ESC		DE	
All	2,941,546	100%	2,978,511	100%
inP	482,978	16%	714,562	24%
FDR	213,271	7%	252,291	8%