Oct 22, 2025

ATAC-seq methods for consensus peak generation to investigate differences in chromatin accessibility per sample group

DOI
https://dx.doi.org/10.17504/protocols.io.36wgq326olk5/v1

Emeline Cherchame[1], Marieke Vromman[1], Justine Guégan[1,2], Marie Coutelier[1]

[1]Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, APHP, Hôpital de la Pitié Salpêtrière, Paris, France (RRID:SCR_026138);
[2]Klee Group - KCI, Centre d'Affaires La Boursidière, Le Plessis-Robinson, France

Emeline Cherchame: https://orcid.org/0000-0002-4140-180X;
Marieke Vromman: https://orcid.org/0000-0002-0163-0675
Justine Guégan: https://orcid.org/0000-0001-6087-5528
Marie Coutelier: https://orcid.org/0000-0002-0261-7210

CHERCHAME EC Emeline
ICM

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.36wgq326olk5/v1

External link: https://dac.institutducerveau.org

Protocol Citation: Emeline Cherchame, Marieke Vromman, Justine Guégan, Marie Coutelier 2025. ATAC-seq methods for consensus peak generation to investigate differences in chromatin accessibility per sample group. **protocols.io** **https://dx.doi.org/10.17504/protocols.io.36wgq326olk5/v1**

Manuscript citation:
Saline Jabre, Emeline Cherchame, Natalia Pinzón, Eline Lemerle, Marc Bitoun, Catherine Coirault, Lamin A/C protects chromatin accessibility during mechanical loading in human skeletal muscle. Cell Communication and Signaling, DOI: 10.1186/s12964-025-02437-z

Protocol status: Working
We use this protocol and it's working

Created: March 09, 2024

Last Modified: October 22, 2025

Protocol Integer ID: 96418

Keywords: ATAC-seq, consensus peaks, differential accessibility region, Peaks, chromatin accessibility, DARs, atac peaks per sample, seq methods for consensus peak generation, atac peak, atacseq pipeline, sample group atac, chromatin accessibility, differences in chromatin accessibility, accessible chromatin, standardized peak, merging peak, consensus peak generation, consensus peak, samples for peak calling, specific peak, atac, example on human atac, specialized analysis pipeline, human atac, seq data, loss of specific peak, peak, numerous peak, wider peaks with numerous peak, seq method, seq, peak calling

Disclaimer

This method has only been used on four projects with a limited number of sample groups. The protocol should therefore be used with caution. To confirm the signal, it is necessary to go back to unprocessed data using visualisation tools like IGV software or UCSC genome browser, working with both bigWig and NarrowPeaks files. We specifically recommend to be cautious when sample groups show a strong difference in general DNA accessibility.

Abstract

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) data require specialized analysis pipelines. The current pipelines deliver ATAC peaks per sample as output without a common reference. However, no gold-standard has been established for consensus peak calling within sample groups or for calling differential accessibility regions (DARs) between sample groups. Tools like HOMER suggest to pool all samples for peak calling, which could lead to the loss of specific peaks from one group. The same fallback exists for intersection stratgeies, while union approaches lead to higher false positive rates. Merging peaks using bedtools merge or tools developed for ChIP-seq, such as DiffBindR, on ATAC-seq data, leads to multiple issues regarding the length of the peaks, skewed towards wider peaks with numerous peaks longer than 1kb.
We therefore present an alternative protocol, with an example on human ATAC-seq data. Peak calling is first performed following the ENCODE guidelines, which are similar to the nf-core/atacseq pipeline. Then, the peaks are standardized and represented by a 500bp interval centered around their summit. For each sample group, the standardized peaks are then merged and filtered based on the replicates into a set of consensus peaks. Finally, the consensus peaks are compared between sample groups using gold-standard statistical methods to compute the DARs.

Guidelines

The following quality threshold are suggested, in accordance with the ENCODE recommendations.

- The total number of reads per sample should be > 50M (25M for paired-end sequencing).
- The number of filtered reads per sample (after removing reads mapping to chrM and to blacklisted regions, and duplicate reads) should be > 10M.
- The alignment of the reads to the reference genome should be > 80%.
- The ATAC-seq insert size profile should indicate the presence of the NFR (Nucleosome Free Region).
- The number of peaks per sample should be > 50,000.
- An enrichment of > 10% in the promoter-TSS regions (transcription start site) and > 20% in the intronic regions is expected.
- The FRiP scores should be > 0.3, though > 0.2 is acceptable.

Samples or groups that do not meet these thresholds should be excluded from the analysis.

Materials

Please find here our recommandations regarding computational GB memories and CPUs needed per rules :

```
mergePeaks:
    cpus: 4
    mem_gb: 24
    walltime: '12:00:00'

annoatePeaks:
    cpus: 6
    mem_gb: 48
    walltime: '12:00:00'

featureCounts:
    cpus: 6
    mem_gb: 48
    walltime: '12:00:00'

multiqc:
    cpus: 4
    mem_gb: 4
    walltime: '02:00:00'

__default__:
    cpus: 1
    mem_gb: 4
    walltime: '02:00:00'
```

The dataset used for the protocol is accessible through GEO Series accession number GSE288984 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE288984).

Safety warnings

- Improvement to be incorporated into the pipeline
  The reproducibility of the peaks should be evaluated by calculating the IDR (Irreproducible Discovery Rate) values.

Peak calling and QC

1 **Peak calling**

Peak calling from the ATAC-seq fastq files is performed following the ENCODE recommendations, with state-of-the-art methods (**https://www.encodeproject.org/atac-seq/**) (Benjamin C. Hitz, et. al, 2023). This is similar to the nf-core/atacseq pipeline (Ewels PA, et. al,, 2020).

Citation

Benjamin C. Hitz , Jin-Wook Lee , Otto Jolanki , Meenakshi S. Kagda, Keenan Graham , Paul Sud, Idan Gabdank , J. Seth Strattan , Cricket A. Sloan, Timothy Dreszer, Laurence D. Rowe , Nikhil R. Podduturi , Venkat S. Malladi, Esther T. Chan , Jean M. Davidson, Marcus Ho, Stuart Miyasato , Matt Simison , Forrest Tanaka , Yunhai Luo , Ian Whaling Eurie L. Hong , Brian T. Lee , Richard Sandstrom , Eric Rynes, Jemma Nelson , Andrew Nishida , Alyssa Ingersoll , Michael Buckley , Mark Frerker , Daniel S Kim , Nathan Boley , Diane Trout , Alex Dobin , Sorena Rahmanian , Dana Wyman , Gabriela Balderrama-Gutierrez , Fairlie Reese, Neva C. Durand, Olga Dudchenko , David Weisz , Suhas S. P. Rao, Alyssa Blackburn, Dimos Gkountaroulis, Mahdi Sadr , Moshe Olshansky , Yossi Eliaz , Dat Nguyen , Ivan Bochkov , Muhammad Saad Shamim, Ragini Mahajan, Erez Aiden, Tom Gingeras , Simon Heath, Martin Hirst, W. James Kent , Anshul Kundaje , Ali Mortazavi , Barbara Wold , and J. Michael Cherry
(2023). The ENCODE Uniform Analysis Pipelines. bioRxiv preprint.

https://doi.org/10.1101/2023.04.04.535623 LINK

Citation

Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S
(2020). The nf-core framework for community-curated bioinformatics pipelines.

https://doi.org/10.1038/s41587-020-0439-x LINK

1.1 The quality of the raw data is evaluated with FastQC (Andrews S., 2010). Adapters are trimmed and low-quality sequences (minimum length: 50bp) removed using fastp (Chen S., 2023) with default parameters.

Citation

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online.

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ LINK

Citation

Chen S (2023)
. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp.

https://doi.org/10.1002/imt2.107 LINK

1.2 Paired-end reads are mapped to the appropriate reference genome (in the example case, the human genome, build hg38) using Bowtie 2 (Langmead B. and Salzberg SL., 2012) with the following parameters:

- -X 2000 (maximum fragment length for valid paired-end alignments)

Reads mapping to mitochondrial DNA are excluded from the analysis. Duplicate reads and reads mapping to the ENCODE blacklisted regions (Amemiya HM. et al, 2019) are discarded using Picard (Broad Institute, 2019).

Citation

Langmead B, Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2.

https://doi.org/10.1038/nmeth.1923 LINK

Citation

Amemiya HM, Kundaje A, Boyle AP (2019)
. The ENCODE Blacklist: Identification of Problematic Regions of the Genome.

https://doi.org/10.1038/s41598-019-45839-z LINK

Citation

Broad Institute (2019). Picard Toolkit. GitHub Repository.

https://broadinstitute.github.io/picard/ LINK

1.3 Peaks are called using the MACS2 (Zhang Y. et al., 2008) callpeak function with the following parameters:

- -q (q-value threshold) 0.005
- --keep-dup FALSE
- --nolambda
- --nomodel (bypass the model building)
- --shift 100
- --extsize 200

Citation

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS
(2008). Model-based analysis of ChIP-Seq (MACS).

https://doi.org/10.1186/gb-2008-9-9-r137 LINK

1.4 Peaks are annotated with HOMER (Heiz et al., 2010) using the appropriate reference genome and default parameters.

Citation

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK
(2010)
. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.

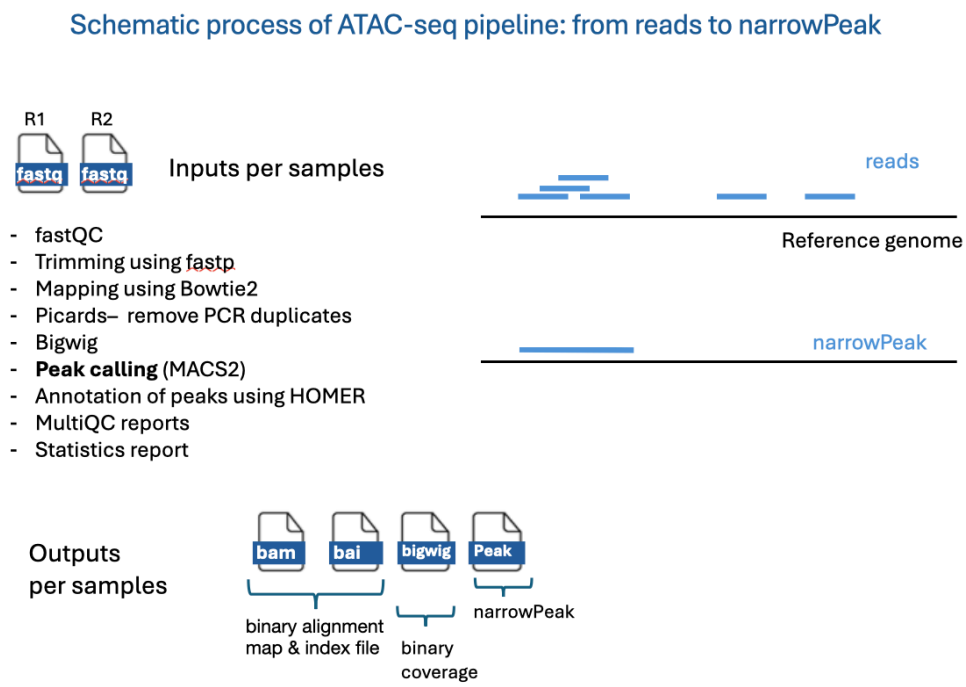
https://doi.org/10.1016/j.molcel.2010.05.004 LINK

2 **Quality control**

The quality reports from the different steps (FastQC, fastp, Picard, Bowtie 2, MACS2) are summarized with multiQC. Several metrics are evaluated using the ENCODE guidelines. In summary:

- The number of total reads per sample should be > 50M (25M for paired-end).
- The number of filtered reads per sample (after removing reads mapping to chrM and to blacklisted regions, and duplicate reads) should be > 10M.
- The alignment of the reads to the reference genome should be > 80%.
- The ATAC-seq insert size profile should indicate the presence of the NFR (Nucleosome Free Region).
- The number of peaks per sample should be > 50,000.

3

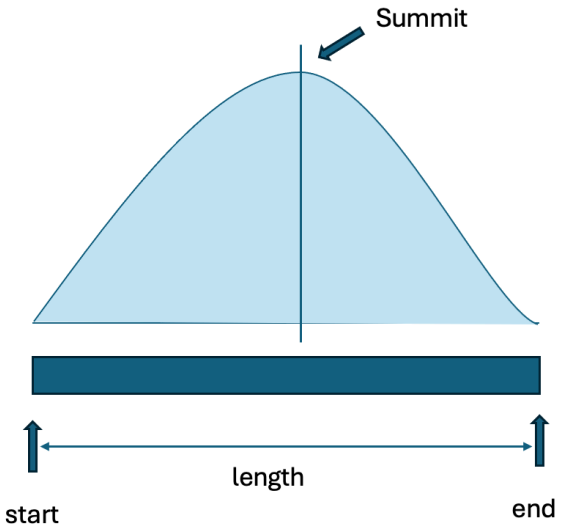


Generating consensus peak sets

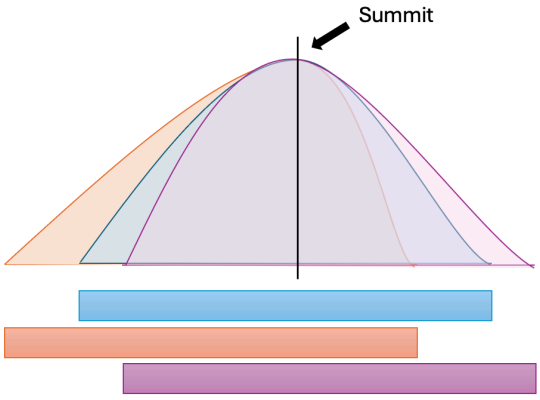
4 **Standardization of peaks around their summits**

According to common practices, the peak coordinates are standardized and represented by a 500bp-interval centered around their summit. We call this set "extended peaks".

4.1 An ATAC-seq peak can be described by its start, end, and summit coordinates. Because similar peaks can have variable start and end positions, the peak summit is used to standardize the peaks.

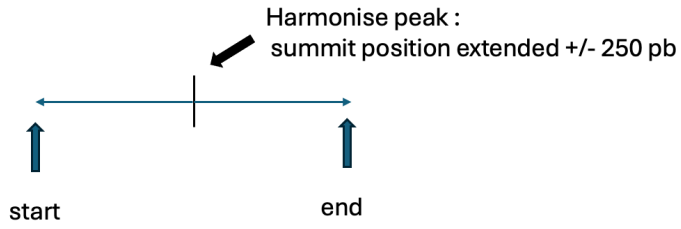


peak coordinates



In 3 different samples, the same peak can show density variability and slightly different coordinates.

For example, these three peaks have different start and end coordinates but the same summit position. The summits have the same coordinate with a high coverage (sequencing depth), but a different start and end coordinate with lower coverage. Standardization of the peaks considers them identical.



Peaks are harmonized into a 500bp interval centered around their summit.

4.2 The header of the MACS2 output (summits.bed), is removed and the summit coordinates are extended by +/- 250 bp.

```
input: "{sample}_summits.bed"
output: "{sample}_summits_extended.bed"
log: "step_summits_extend_{sample}.log"
```

Command

Extend peak summits

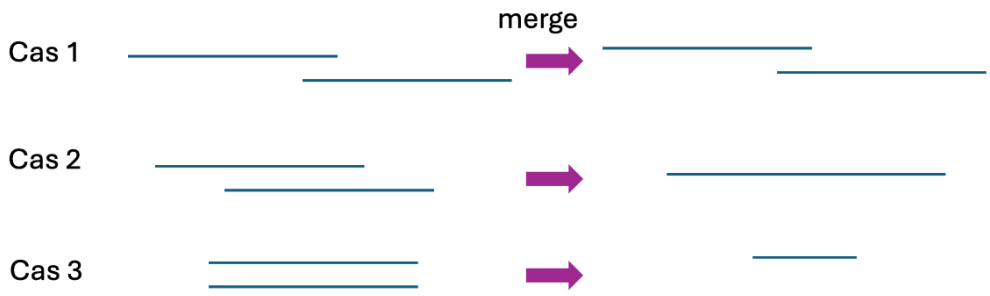
```
awk NR > 1 { OFS=\t; print $1, $2-250, $3 251, $4, $5 } {input} >
{output} 2> {log}
```

5 **Generation of consensus peak sets per sample group**

The extended peak coordinates are merged per sample group in order to obtain a peak set that represents the accessible chromatin for that group.

The mergePeaks script from HOMER is used. The merging of overlapping peaks is determined by the "-d" parameter: when the distance between the peak centers is smaller than 250 bp, overlapping peaks are merged. If it's a strict overlap (same start-end coordinates) the peak is reduced to 250pb, as coded by HOMER. If it's an extended

overlap (different start-end coordinates) then the merge is performed using the minimal start and maximal end as coordinates.



schematic visualisation of merged processed by HOMER

In this example, the two sample groups are called Control and Treated.

5.1 Merge standardized peaks per sample group

Command

Merge peaks for the control group

```
mergePeaks PATH/Control2_summits_extended.bed \
PATH/Control3_summits_extended.bed \
PATH/Control4_summits_extended.bed \
-d 250 > Control.bed 2> merge_Control.log
```

Command

Merge peaks for the treated group

```
mergePeaks PATH/Treated2_summits_extended.bed \
PATH/Treated3_summits_extended.bed \
PATH/Treated4_summits_extended.bed \
-d 250 > Treated.bed 2> merge_Treated.log
```

5.2 The output BED files are formatted to include an unique peak identifier: "chr_start_end".

```
input: Control.bed
output: Control_format.bed
log: step_Control_FormatBed.log
```

Command

Format bed file

```
awk 'NR > 1 { OFS="\t"; print $2, $3, $4, ".", $7, $8, $9, $10, $1,
$2"_"$3"_"$4 }' {input} | sort -T '/tmp' -k1,1 -k2,2n > {output} 2>
{log}
```

5.3 Peaks are filtered to be present in at least two replicates of that sample group. These are considered consensus peaks for that sample group.

```
input:  Control_format.bed
output: Control_filtered.bed
log: step_Control_Filtered.log
```

Command

Filter peaks

```
awk 'BEGIN {OFS="\t"} { if ($6 >= 2) print $1, $2, $3, $4, $5, $6,
$7, $8, $9 }' {input} > {output} 2> {log}
```

5.4 The bed files are formatted to keep only chromosome, start, end, and sample group.

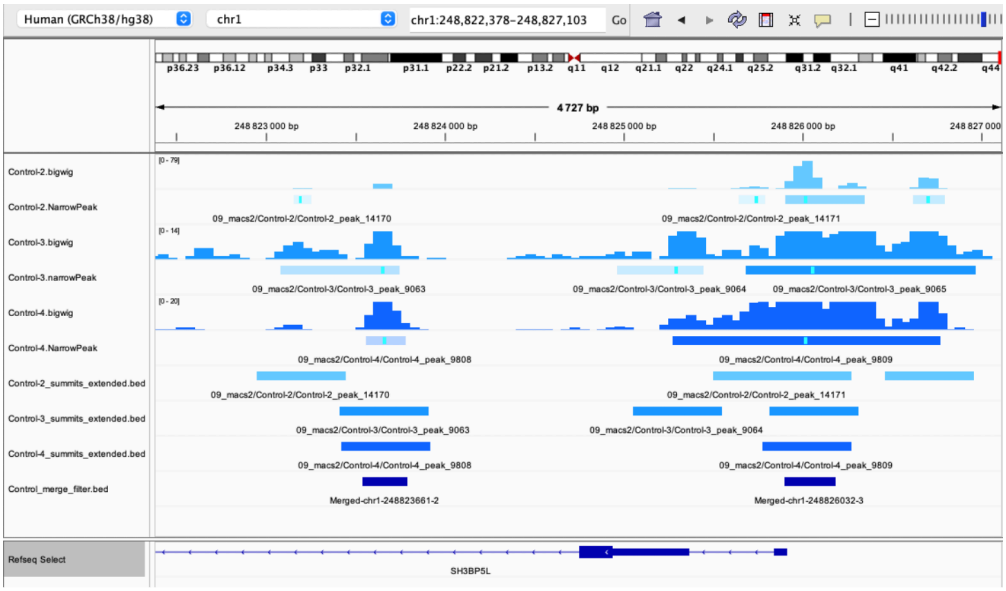
```
input: Control_filtered.bed
output: Control_final.bed
log: step_Control_format_Filtered.log
```

Command

Format merged peaks file

```
awk BEGIN {OFS=\t } { print $1, $2, $3, $9, - } {input} > {output} 2>
{log}
```

5.5 Consensus peaks can be visualised using IGV (Robinson JT., et. al, 2011) (one file per sample group).



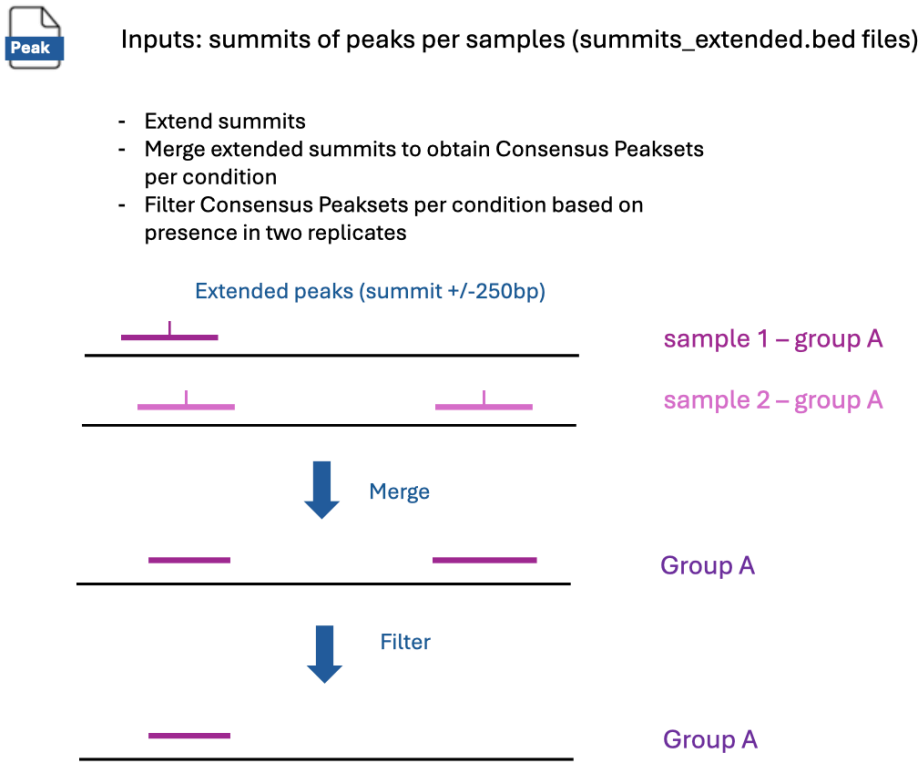
Merge of extended peak summits per group: sequencing depth and peak calling output are shown for three samples, followed by extended peaks bed files, then filtered consensus peaks bed file.

Citation

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011). Integrative genomics viewer.

https://doi.org/10.1038/nbt.1754 LINK

Schematic process: Generate Consensus Peaksets per SampleGroup



5.6 All consensus peak files (from each group) are concatenated and sorted.

```
input: Control_final.bed, Treated_final.bed
output: All_Consensus_peaks.bed
log: step_gather_All_Consensus_peaks.log
```

Command

Gather and sort all consensus peaks

```
cat {input} | sort -k1,1 -k2,2n > {output} 2> {log}
```

5.7 The consensus peaks are merged using HOMER with the same parameters as in step 4.1, generating one file containing all the peaks from all the sample groups. Because the peaks were concatenated in the previous step, there is only one input file, and the peaks with exact overlaps are not reduced to 250bp but remain their original size.

input: All_Consensus_peaks.bed
output: Merge_All_Consensus_peaks.bed
log: step_merge_All_Consensus_peaks.log

Command

Merge consensus peaksets

```
mergePeaks -d 250 {input} > {output} 2> {log}
```

5.8 The consensus peak file is formatted and sorted to allow IGV visualisation.

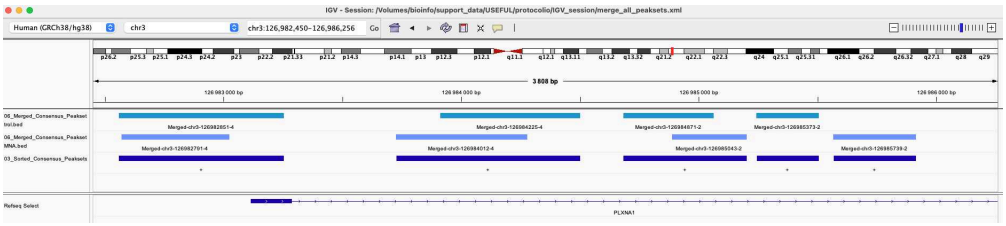
```
input: Merge_All_Consensus_peaks.bed
output: Final_All_Consensus_peaks.bed
log: step_Sort_All_Consensus_peaks.log
```

Command

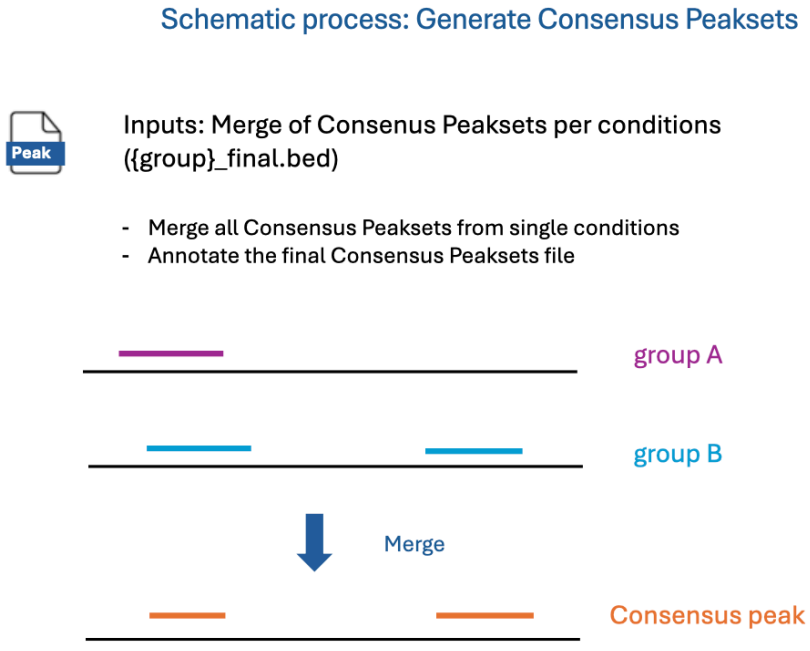
Reshape and sort bed file

```
awk 'NR > 1 { OFS="\t"; print $2, $3, $4, $5, $7, $8, $9, $10, $1,
$2\"_\"$3\"_\"$4 }' {input} | sort -k1,1 -k2,2n > {output} 2> {log}
```

5.9 Final consensus peaks can be visualised using IGV (Robinson JT., et. al, 2011).



The consensus peaks bed files per group are shown, followed by the final consensus peaks bed file.



6 Consensus peak annotation

6.1 The consensus peak file is annotated using HOMER (Heinz et al., 2010) using the appropriate reference genome database.

Citation

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK
(2010)
. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.

https://doi.org/10.1016/j.molcel.2010.05.004 LINK

```
input: Final_All_Consensus_peaks.bed
output:
            stats=Annotated_All_Consensus_peaks.stats,
            bed=Annotated_All_Consensus_peaks.annoHomer
params: reference=hg38
benchmark: "benchmarks/macs2/Homer.json"
log: step_Annotate_All_Consensus_peaks.log
```

Command

annotatePeaks

```
annotatePeaks.pl {input} {params.reference} \
-annStats {output.stat} > {output.bed} 2> {log}
```

6.2 A GTF file is generated based on the annotated consensus peak file.

```
input: Final_All_Consensus_peaks.bed
output: All_Consensus_peaks.gtf
log: step_makeGTF_All_Consensus_peaks.log
```

Command

Convert bed to GTF format

```
awk BEGIN { OFS=\t} { print $1, ATACflow, Peaksets, $2, $3, ., $4, .,
gene_id= $7 } {input} > {output} 2> {log}
```

7 **Consensus peak example metrics**

Here is a summary of the consensus peak metrics we generated across several projects.

	A	B
	Mean length (bp)	450
	Min length (bp)	249
	Max length (bp)	5000
	N with length < 500 bp	90 000
	N with length < 2000 bp	150 000
	Total number of peaks	155 000

Average of consensus peak metrics generated in several ATAC-seq projects.

8 **Quality control of the consensus peaks**

Quality metrics from the GEO GSE288984 dataset are shown below as an example.

- Consensus peak size

	A	B
	Mean length (bp)	377
	Min length (bp)	249
	Max length (bp)	1,769

Consensus peak size metrics

- The number of consensus peak

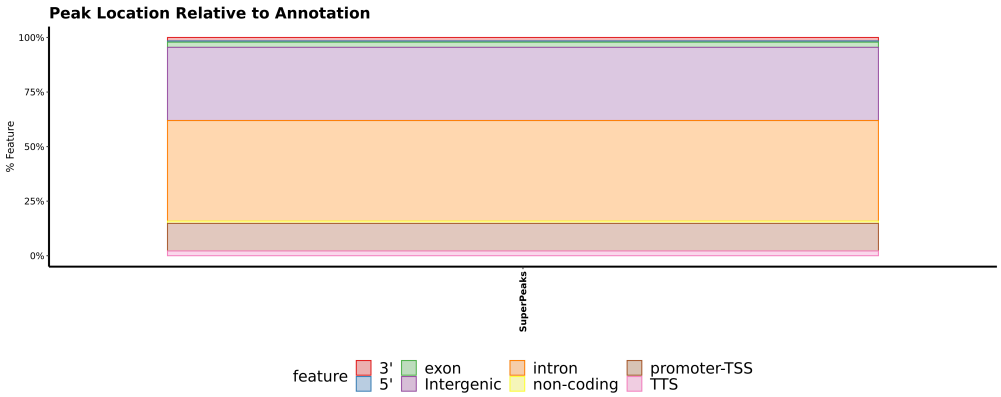
	A	B
	N with length < 500 bp	100,393

	A	B
	N with length < 1000 bp	113,937
	Total number of peaks	114,122

Number of consensus peak respecting size metrics

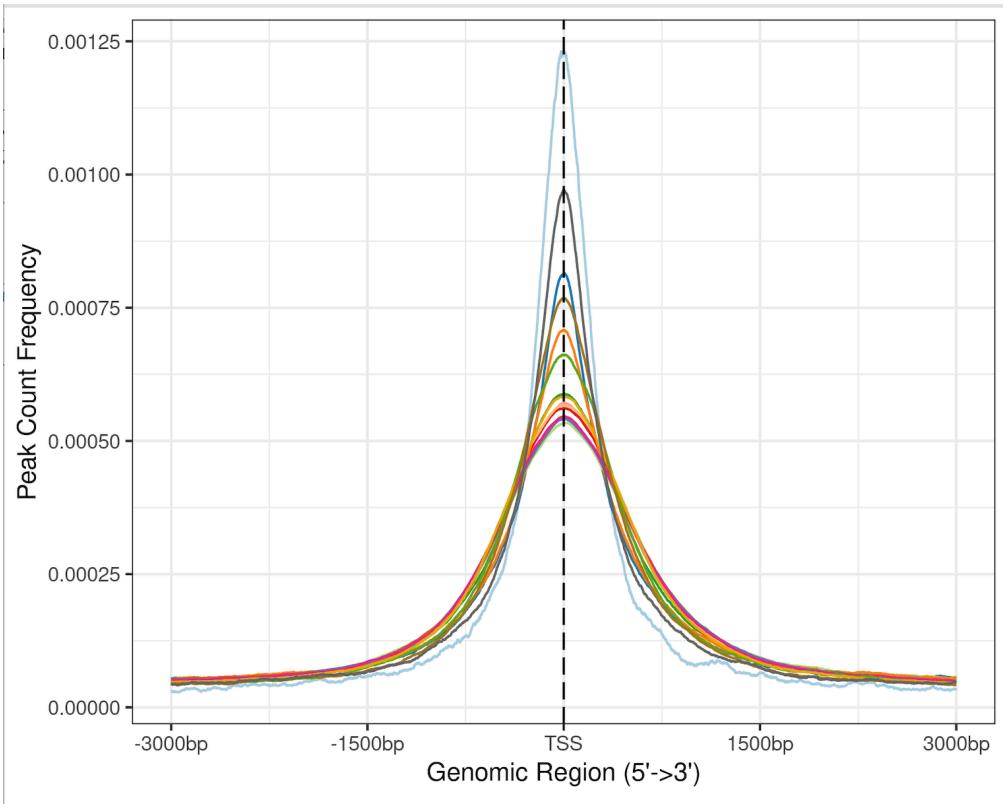
- Consensus peak annotation

An enrichment of > 10% in the promoter-TSS regions (transcription start site) and > 20% in the intronic regions is expected.



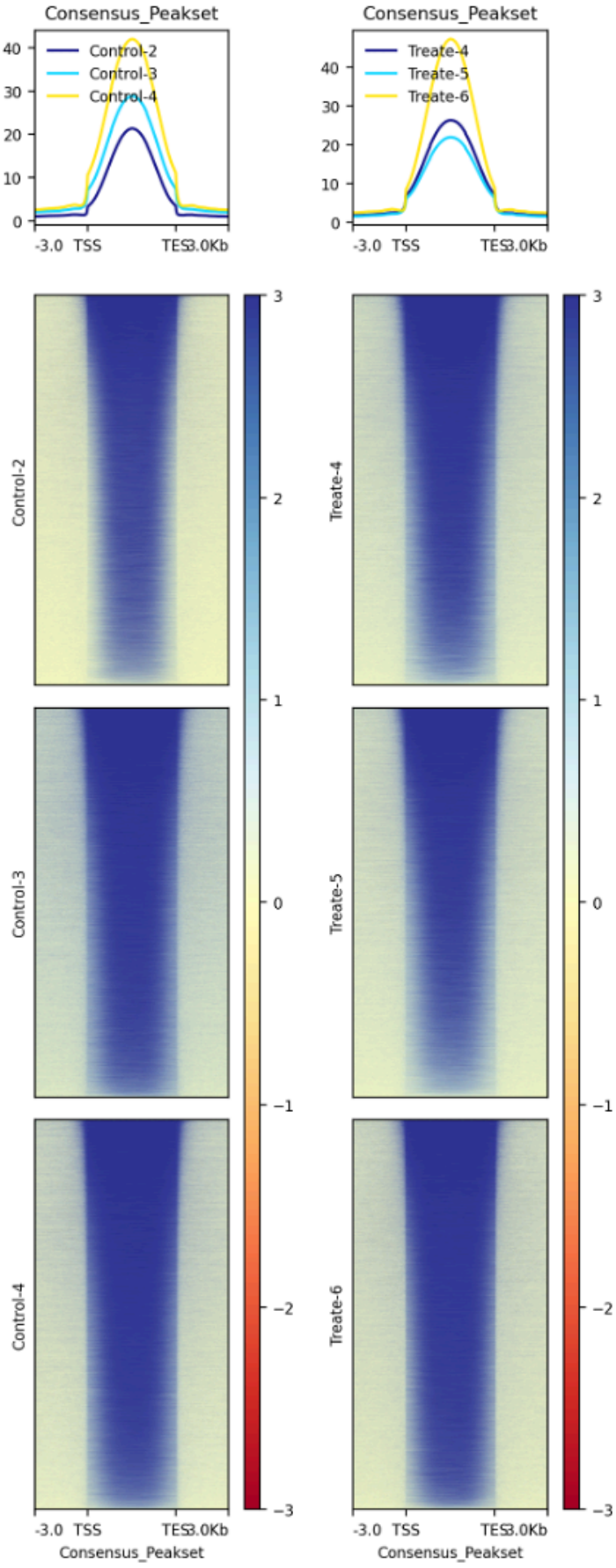
Consensus peak annotation
TSS = Transcription Start Site (from -1kb to +100bp)
TTS = Transcription Termination Site (from -100 pb to +1kb)

- TSS enrichment



Peak count frequency at each genomic position relative to the TSS.

- Heatmaps of the ATAC-seq signal in consensus peak regions.


Consensus_Peakset
Control-2
Control-3
Control-4
Consensus_Peakset
Treate-4
Treate-5
Treate-6
40
30
20
10
0
-3.0 TSS
TES3.0Kb
Control-2
Control-3
Control-4
Treate-4
Treate-5
Treate-6
3
2
1
0
−1
−2
−3
Consensus_Peakset

Plots generated using Deeptools computeMatrix and PlotHeatmap function (v.3.5.6) on consensus peak coordinates. Each line is a specific genomic interval, the blue signal represents high coverage. We can verify that the TSS region is enriched for ATAC-seq signal in all the samples.

Generating the ATAC-seq counts matrix

9 A peak count matrix is generated using featureCounts (Yang et al., 2014) based on the consensus peaks GTF file. FeatureCounts counts the number of reads in each region of the reference GTF file to generate a count matrix. Furthermore, it reports the percentage of "Successfully assigned alignments", corresponding to the fraction of reads mapping to called peak regions (FRiP score). According to the ENCODE recommendations, the FRiP scores should be > 0.3, though > 0.2 is acceptable.

Parameters :

- -p (option for paired-end reads / to remove for single-end)
- -t Peaksets
- -g gene_id

Citation

Liao Y, Smyth GK, Shi W (2014)
. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.

https://doi.org/10.1093/bioinformatics/btt656 LINK

Generating Differential Accessibility Regions (DARs)

10 To identify Differential Accessibility Regions (DARS), DESeq2 or EdgeR/limma can be used. It's recommended to use DESeq2 when high specificity is required for large sample sizes and to use EdgeR (within limma package) when high sensitivity is required or for smaller sample sizes (Gontarz et al., 2020).

Citation

Gontarz P, Fu S, Xing X, Liu S, Miao B, Bazylianska V, Sharma A, Madden P, Cates K, Yoo A, Moszczynska A, Wang T, Zhang B
(2020). Comparison of differential accessibility analysis strategies for ATAC-seq data. Scientific reports.

https://doi.org/10.1038/s41598-020-66998-4 LINK

11 **DAR analysis is performed**

Counts are normalized with the edgeR package (v3.28.0) (Robinson et al., 2010) from Bioconductor. Regions are filtered based on a cpm > 1 in at least 3 samples. DARs are identified with limma (v3.50.3) (Ritchie et al., 2015). Multiple hypothesis adjusted p-values are calculated with the Benjamini-Hochberg procedure to control FDR. DARs are filtered for FDR < 0.05, then annotated with the previous HOMER annotation.

Citation

Robinson MD, McCarthy DJ, Smyth GK (2010)
. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.

https://doi.org/10.1093/bioinformatics/btp616 LINK

Citation

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015)
. limma powers differential expression analyses for RNA-sequencing and microarray studies.

https://doi.org/10.1093/nar/gkv007 LINK

Citations

Step 1.1

Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp.
https://doi.org/10.1002/imt2.107

Step 10

Gontarz P, Fu S, Xing X, Liu S, Miao B, Bazylianska V, Sharma A, Madden P, Cates K, Yoo A, Moszczynska A, Wang T, Zhang B. Comparison of differential accessibility analysis strategies for ATAC-seq data.
https://doi.org/10.1038/s41598-020-66998-4

Acknowledgements

We gratefully acknowledge Catherine Coirault for allowing us to use her dataset to present our protocol.

Author contributions :
Conceptualization and methodology: EC, JG. Software: EC, MV. Supervision: EC, MC. Writing - original draft: EC. Review & editing: MC, MV. Project administration: MC.