Functional genomics of psoriasis

Alicia Lledo Lara
Hertford College
University of Oxford

A thesis submitted in partial
fulfilment of the requirements for the degree of
Doctor of Philosophy
Trinity Term, 2018

Abstract

Functional genomics of psoriasis

Alicia Lledo Lara, Hertford College, Trinity Term 2018
A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy of the University of Oxford

This is my abstract...

Acknowledgements

Thank you, thank you, thank you.

Declarations

I declare that unless otherwise stated, all work presented in this thesis is my own. Several aspects of each project relied upon collaboration where part of the work was conducted by others.

Submitted Abstracts

Title **Year**
Authors

Associated Publications

Title
Journal
Authors

Other Publications

Title
Journal
Authors

Contents

List of Figures

List of Tables

Abbreviations

Abbreviation	Definition
Ab	Antibody
ATAC-seq	
Atopic dermatitis	AD
ChIPm	
CLE	cutaneous lupus erythematosus
DMARDs	disease-modifying antirheumatic drugs
Fast-ATAC	
IDR	
GWAS	Genome-wide association studies
KC	Keratinocytes
NSAID	nonsteroidal antiinflammatory drug
Omni-ATAC	
PCA	
PI	Protein inhibitor
PsA	
QC	
qPCR	quantitative polymerase chain reaction
RA	Rheumatoid arthritis
ROS	Reactive oxygen species
SDS	Sodium dodecyl sulfate
SF	Synovial fluid

Chapter 1

Establishment of laboratory methods and analytical tools to assess genome-wide chromatin accessibility in clinical samples

1.1 Introduction

1.1.1 Principle of ATAC-seq and compatibility with clinical samples

Several techniques including DNase-seq, FAIRE-seq and MNase-seq have been used during the last few decades to map the accessible genome in different cell lines and some abundant sources of primary cells, as previously reviewed in Chapter **??**. All these techniques present, amongst other limitations, the large number of cells required as input material, making them unsuitable for the use in a clinical setting. Publication of ATAC-seq represented a revolution in the field to interrogate chromatin accessibility. ATAC-seq uses a hyperactive modification of the bacterial transposase Tn5 to perform simultaneous fragmentation and insertion of synthetic oligonucleotides (adapters) into native chromatin from 50,000 cells and also at the single-cell resolution (**Buenrostro2013**; Buenrostro et

al. 2015). The Tn5 reaction incorporates in a non-strand specific manner adapters containing the complementary sequences to the i5-R1 and i7-R2 elements, required for Illumina NGS. ATAC-seq represented a fast two-steps protocol, not requiring cross-linking, enzyme titration or sonication, that was able to yield information regarding nucleosome-free DNA (fragments ≤150bp) and DNA spanning nucleosomes (fragments >150bp). ATAC-seq data can also be used to identify TF foot-printing as well as nucleosome positioning in the genome. This new technique opened a new avenue to interrogate the chromatin landscape in clinical samples with limited input material as well as in rare cell populations with a shorter preparation time and results turn-over.

1.1.2 ATAC-seq limitations and advances in optimisation

Despite the advantages in terms of reduced input material and time processing, ATAC-seq revealed two major limitations involving high percentage of mitochrondrial (MT) DNA tagged by the Tn5 enzyme and insufficient sensitivity to detect all the accessible regions, partly due to high background noise (**Corces2016**; **Sos2016**). An optimised version of protocol particularly for hematopoietic cells, named Fast-ATAC, replaced the NP-40 detergent used in ATAC-seq by digitonin. This prevents solubilisation of the MT membrane, and performed lysis and transposition in a single step. This modifications efficiently reduced the percentage of MT reads down to approximately 10% and increased ATAC signal at annotated TSS, using only 5,000 cells as input material (**Corces2016**). Optimisation of the ATAC-seq protocols for KCs was also published by Bao and colleagues, where they performed the two ATAC-seq steps directly on the 96-well plate containing adherent NHEKs using an increased concentration of Tn5 in the transposition reaction (**Bao2015**).

The third generation of ATAC, known as Omni-ATAC, was released in 2017 and offered a generic version of the protocol optimised to yield high quality

data in any cell type and fresh or frozen tissue (**Corces2017**). The Omni-ATAC protocol consisted of lysis, a wash and transposition steps. In addition to the NP-40 and digitonin, used by the previous ATAC protocols, Omni-ATAC also included Tween-20 in the lysis buffer to improve cell permeabilisation. Comparison of Omni-ATAC with ATAC-seq and Fast-ATAC data demonstrated the higher variability in quality samples and sensitivity of the two later ones. Moreover, greater signal-to-noise ratios were also achieved by modification in the transposition buffer. Notably, this versatile protocol presented particular improvement in KCs with data demonstrating the inability of ATAC-seq and Fast-ATAC to yield good quality data in this cell type.

1.1.3 Challenges of ATAC data analysis

Although some guidelines for DHS data analysis were available, the release of the new ATAC methods also represented the need to adapt and develop additional tools and strategies for chromatin accessibility data analysis. In contrast to DNase-seq or FAIRE-seq data starting from high number of cells (minimum of $10x10^6$ cells), ATAC-seq and Fast-ATAC showed lower signal-to-noise ratios and higher variability across samples, as previously mentioned. This required appropriate implementation of quality control (QC) measurements in order to confidently identify good quality samples prior to downstream analysis. Regarding peak calling in ATAC, different algorithms have been applied, being MACS2 the preferred by the majority of them, including ENCODE (Table 1.1). The criteria to filter good quality peaks in ATAC represents another critical aspect, particularly for the libraries at the lower end of the quality spectrum. Using false discovery rate (FDR) has been the most widely applied criteria except for ENCODE data, where technical replicates are generated and the irreproducibility discovery rate (IDR) analysis has been used to identify robust peaks (Table 1.1).

Table 1.1: Summary table of ATAC-seq methodology analysis for peak calling, filtering and differential analysis. NA indicates the study did not perform or detail that aspect of the analysis.

Publication	Peak calling and filtering	Master list	Differential analysis
Corces *et al.*, 2016	MACS2 (-nomodel), peak summit extension +/−250bp, rank summits by pval	Maximally significant non-overlapping peaks.	Quantile normalisation and unsupervised hierarchical clustering.
ENCODE	MACS2 -nomodel, pairwise IDR analysis, filtering IDR<10%	Choosing longest pairwise IDR filtered list or only peaks present in the two samples pseudoreplicates.	NA
Turner *et al.*, 2018	MACS2 (-nomodel –q 0.01)	Merging all filtered called peaks from the different cell types.	*De novo*:DiffReps with fragment size 50bp.
Alasoo *et al.*, 2018	MACS2 (-nomodel -shift -25 -extsize 50 –q 0.01	Union of peaks from all conditions present in at least three samples of the same condition.	Peak based: TMM normalisation and lima voom (FDR<0.01).

Qu *et al.*, 2017	ZINBA PP>0.99.	Merging of filtered peaks from each individual sample.	Quantile normalisation and peak based in house Pearson correlation method.
Rendeiro *et al.* 2016	MACS2 (-nomodel -extsize 147)	Merge of peaks from all samples in an iterative process including permutations	Peak based: quantile normalisation and Fisher exact text (FDR<0.05).
Scharer *et al.* 2016	HOMER (-style dnase)	Merge of all overlapping peaks between all samples using HOMER mergePeaks	Peak based: TMM normalisation and edgeR package (FDR<0.05).

The feasibility to generate ATAC data from low number of cells and clinical samples represents an opportunity to perform differential chromatin accessibility analysis between conditions, cell types or groups of patients and healthy control samples. The most common approach is a peak based strategy, which requires building a non-redundant and non-overlapping list of high quality peaks, counting the reads mapping to those locations and perform normalisation across samples before conducting differential analysis with microarray or RNA-seq based-methods. A alternative is the known as *de novo* approach, used for ChIP data, which consists of using an sliding window to scan the genome and identify those regions showing read count differences between two groups of samples, avoiding peak calling bias (**Shen2013**).

1.1.4 The challenge of working with clinical samples

The opportunity to apply epigenetic assays to clinical samples has also highlighted a logistic problem regarding handling of clinical samples. Often sample recruitment takes place at distant geographical locations or out of normal working hours. This requires the application of preservation methods that prevents the alteration of the *in vivo* cellular characteristics and avoids confounders in the biological significance of the generated results. The main methods to preserve cell structure and DNA integrity involves cryopreservation or DNA-protein fixative compounds such as formaldehyde. Regarding preservation of pure cell populations for ATAC processing, a study in motor neurons demonstrated that slow-cooling using DMSO but not snap-freezing maintained intact cell nucli and chromatin organisation and overall yielded comparable ATAC-seq data to the one generated in fresh neurons (**Milani2016**). When working with mixed cell populations such as PBMCs, slow temperature cryopreservation with DMSO allows long term storage and also offers the flexibility of retrospective separation of distinct cell populations

by FACS following thawing and recovery. However, in a mixed population such as PBMCs some cell types are more sensitive to cryopreservation and that may lead to distinct alterations in the chromatin accessibility landscape and gene expression profile. In terms of fixatives, the Oxford Genomic Center at the WCHG had incorporated the use of dithio-bis(succinimidyl propionate) (DSP) to stabilise cell samples for single-cell transcriptomic applications demonstrating only moderate differences from fresh samples profiles (**Attar2018**). DSP is a reversible cross-linker of free amine groups that fixes proteins without damaging RNA and is compatible with microfluidics-based scRNA-seq systems, alike formaldehyde fixation. DSP preservation does not require sample freezing after fixation and samples can undergo successful immuno-staining as well FACS cell separation (**Espina2013**).

1.1.5 Aims

1.2 Results

1.2.1 Establishment of an ATAC-seq data analysis pipeline based on current knowledge

At the time of the first ATAC-seq publication (**Buenrostro2013**), well established protocols for the complete processing and data analysis were missing. Since then, several publications have implemented ATAC-seq and modifications of this protocol together with a wide range of data analysis strategies to answer different biological questions (Table 1.1). In the process of analysing ATAC-seq data, several limiting aspects are encountered, including QC assessment, peak callingfiltering and identification of differential chromatin accessibility regions between groups of samples. Using the current knowledge in the field as well as own analysis, the most appropriate criteria and parameters to implement in the

in-house pipeline were agreed. For this purpose different types of analysis were performed using ATAC-seq data generated with the (**Buenrostro2013**) protocol in paired $CD14^+$ monocytes and total $CD4^+$ T cells ($tCD4^+$) from three healthy individuals (ATAC-seq fresh samples generated for).

Sample QC

Regarding QC measurements, the variability in performance of ATAC-seq and Fast-ATAC, has required to identify appropriate parameters to determine the quality of the samples before proceeding with downstream differential analysis. This has been a dynamic process during the project that has benefited from the increase in the number of publications including ATAC data analysis as well as understanding the technical limitations from ATAC-seq and Fast-ATAC protocols. After continuous review of the different read-outs implemented across different publications, as well as the recently ENCODE updates, a comprehensive analysis was performed in order to identify the most informative QC measures as well as equivalence and correlation between them.

First, ATAC-seq fragment size distribution was analysed in each of the six libraries following down-sampling to 30 million reads after filtering, to facilitate the comparison across them. The fragment size distribution showed presence of nucleosome periodicity protecting the DNA during the transposition event (Figure 1.1 a), one of the indicators of chromatin integrity and thus good quality ATAC-seq libraries. All the samples presented appropriate nucleosome periodicity (every ~200bp) up to 600bp, clearly distinguishing chromatin organisation into mono-, di- and tri-nucleosomes. The relative intensity of nucleosome-free fragments (NFF,≤147pb, approximately) compared to nucleosome-bound DNA was greater for some of the samples (e.g CTL1 $CD4^+$ and $CD14^+$) and similar or lower for others (e.g CTL3 $CD4^+$ and $CD14^+$). NFF were clearly distinguished in all of the samples and is considered a compulsory

feature for ATAC-seq libraries to pass QC, according to the latest ENCODE recommendations (**ENCODE**).

Another QC measurement that was investigated and implemented was the enrichment of ATAC-seq signal over a random background of reads across all the TSS identified for Ensemble genes (Figure 1.1b). It is well established that nucleosome repositioning and an increase in chromatin accessibility take place at the genes TSS to allow TFs binding and initiation of transcription.

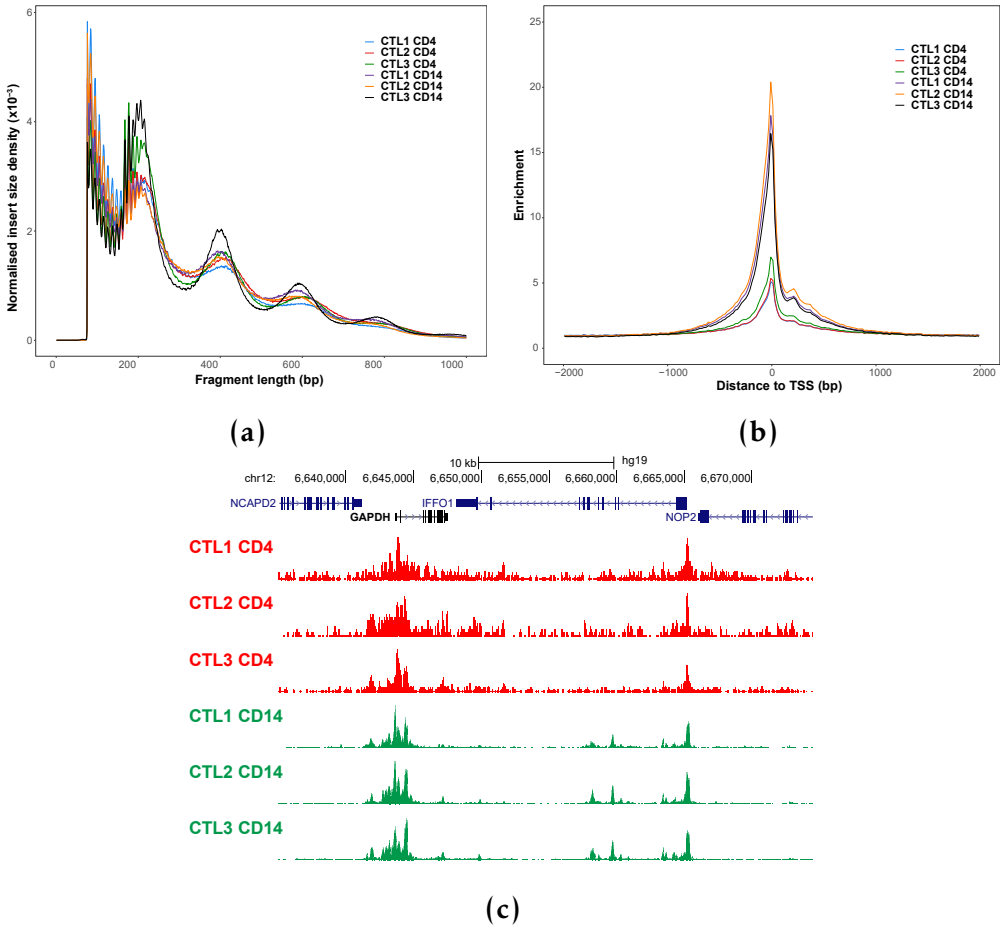


Figure 1.1: Measurements for QC assessment in ATAC-seq samples. For each of the CD14$^+$ monocytes and tCD4$^+$ samples used to establish ATAC analysis QC measures include a) density distribution of ATAC-seq fragment sizes, b) enrichment of ATAC-seq fragments across TSS of all the Ensembl genes and c) UCSC Genome Browser view illustrating the ATAC-seq normalised read density (y-axis) at the promoters of *GAPDH* and *NOP2* genes. In c) CD14$^+$ monocytes and tCD4$^+$ tracks are colour-coded in green and red, respectively.

Fold-enrichment signals over the TSS ranged between 5-7 for the tCD4$^+$ samples being much higher(between 17-20) in the CD14$^+$ samples. The lower sample quality of the tCD4$^+$ compared to CD14$^+$ samples indicated by the TSS enrichment values were recapitulated by the ATAC-seq reads pile up at the promoters of the glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) and the NOP2 Nucleolar Protein gene *NOP2*, presenting more background reads and less define signal for the tCD4$^+$ samples in red (Figure 1.1c).

As part of the QC assessment, the percentage of mitochondrial reads and the fraction of reads in peaks (FRiP) were also investigated (Table 1.2).

Sample	% MT reads	Fraction of reads in peaks
CTL1 CD4	14.9	9.8
CTL2 CD4	30.5	11.2
CTL3 CD4	28.8	11.6
CTL1 CD14	43.3	32.2
CTL2 CD14	36.8	57.0
CTL3 CD14	37.6	49.9

Table 1.2: ATAC-seq percentage of MT reads and fraction of reads in called peaks (FRiP). The percentage of MT reads was calculated over the total number of sequencing reads (before filtering). FRiP was calculated as the proportion ATAC-seq fragments overlapping significant peaks with standard filtering for all the samples (FDR<0.01).

FRiP score is an alternative way to TSS of assessing the background signal in different types of assays that are based on peak calling, including ChIP-seq. Positive correlation between the TSS fold-change enrichment and FRiP was observed (data not shown), being both appropriate inter-dependent QC measures to evaluate sample noise. The MT content ranged between 14.9-43.3% and, alike FRiP and TSS, it was higher in CD14$^+$ than in tCD4$^+$ and not directly related with any of the other QC measurements. Therefore, MT reads in this range did not appear to reflect in the samples quality and the main inconvenience related to the need of deeper sequencing to achieve the desired number of no MT reads for downstream analysis.

In summary, both TSS and FRiP appeared as appropriate signal-to-noise measures, with recommended threshold values by ENCODE and Alsoo *et al.*, 2018 of FRiP between 10-20% and TSS between 6-10. Importantly ENCODE has prioritised the use of TSS over FRiP as a more stable measure to determine the noise in the sample and will also be the chosen measure in this study. Overall, according to this analysis, all six samples passed showed appropriate ATAC-seq pattern of fragment size distribution, FRiP and also TSS, with exception of $tCD4^+$ CTL1 and CTL2, which were borderline for the 6 fold-enrichment threshold. Importantly, this differences in the enrichment around the TSS successfully recapitulated the differences observed in the ATAC-seq density signal of the UCSC Genome Browser tracks between the $CD14^+$ and $tCD4^+$ samples. This differences in ATAC-seq quality observed in these samples reflected the variability in performance of ATAC-seq and was a good scenario to determine the influence of borderline sample quality in the downstream analysis in order to choose the most robust strategy that allows maximising the use of precious clinical samples.

Peak calling and filtering

As part of the ATAC-seq pipeline implementation, peak calling and their their filtering criteria were another two aspects to determine. Although different peak callers have been used to analyse ATAC-seq data, MACS2 has been the preferred methodology by ENCODE and most of the publications (Table 1.1). MACS2 has initially been developed for ChIP data, but it has also been used for DHS and ATAC-seq disabling the model option and manually setting the shift (–shift) and extension size (–extsize), which refer to the number of bp and direction for the reads to be shifted and the number of bp for them to be extended, respectively. Since the –extsize should correspond to the average fragment size, it was set to 200bp which was the average fragment size calculated for the ATAC-

seq libraries in this project. The –shift was set to -100, as it is recommended to be -1/2 of the fragment size when analysing chromatin accessibility data, such as DHS or ATAC-seq. Later publications based on clinical samples have also used similar parameter for ATAC peak calling (**Wang2018**).

Although the optimisation of the previous parameter escaped from the aim of this thesis, a systematic analysis of the effect of sequencing depth and the sample quality on the peak calling was conducted to better control the effect of both variables in the downstream analysis. For each of the six samples, random sub-sampling of reads was performed every 5 million, ranging from 5 to 30 total million reads, and followed by peak calling with arbitrary filtering for false discovery rate (FDR)<0.01. The number of called peaks passing filtering showed an steady increase over the read depth reaching a *plateau* at approximately 25 million reads (Figure 1.2 a). This was consistent with the decay in the increments of called peaks over read depth showing small variation from 20 million reads onwards (Figure 1.2b). Moreover, lower number of peaks were detected in $tCD4^+$ samples compared to $CD14^+$ when using standard FDR<0.01 filtering, highlighting the influence of sample quality on the total number of significant called peaks. Interestingly, sample quality measured by FRiP (which relies on peak calling) reflected very low changes over read depth and was stable from 15 million reads onwards for all six samples (Figure 1.2 c), similarly to TSS (data not shown). Overall, this confirmed that measurement of sample quality using FRiP or TSS was not affected by the sequencing depth.

Regarding peak calling filtering, most of the ATAC-seq publications using MACS2 have arbitrarily used an FDR<0.01 (Table 1.1). However, this arbitrary way of filtering may not remove low quality peaks equally successfully in samples at the lower side of the acceptable quality spectrum. Moreover, filtering based on MACS2 FDR<0.01 does not take into account the reproducibility of the called peaks. In collaboration with Dr. Gabriele Migliorini and following ENCODE

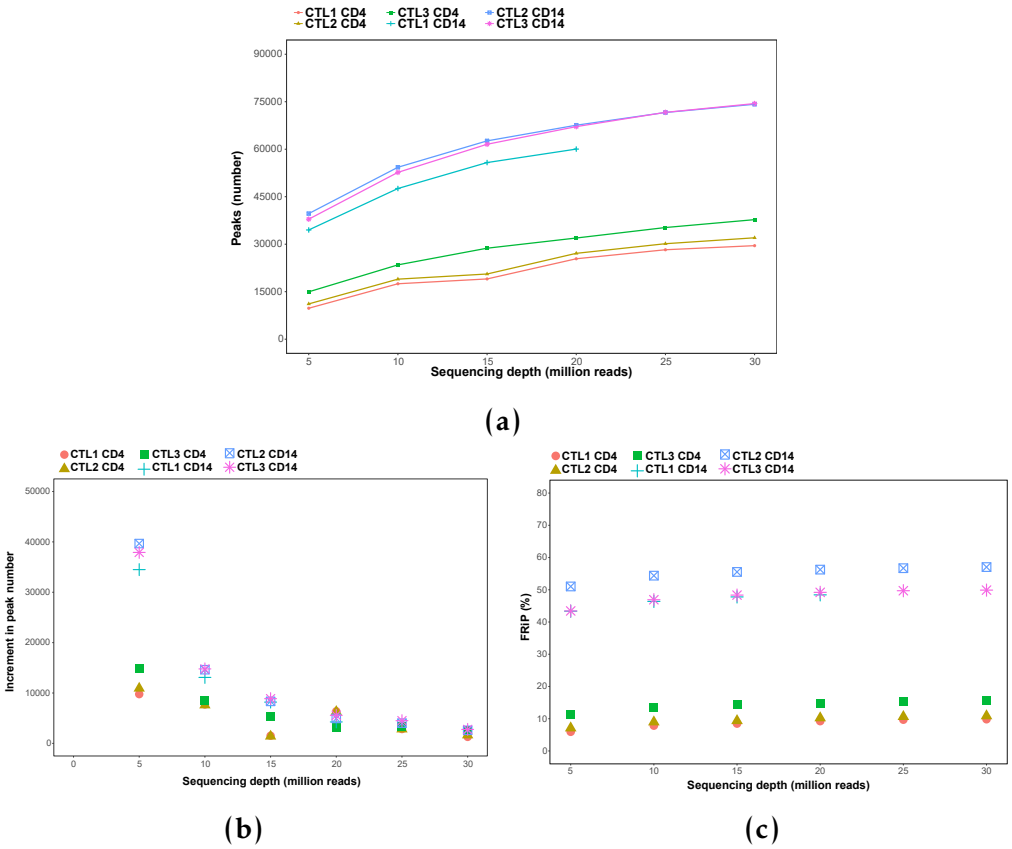


Figure 1.2: FRiP and peak calling at different sequencing depths in ATAC-seq libraries. For a serie of sequencing depths (from 5 to 30 million of reads after filtering) representation of a) number of called peaks (standard filtering using FDR<0.01, b) the increment on the number of called peaks and c) FRiP as a function of the sequencing coverage in the six samples included for the analysis in this section.

pipeline, IDR was used to experimentally identify the most appropriate pval to filter the called peaks in each individual sample. Filtered reads from each sample partitioned in half to create two pseudoreplicates, peaks were called in each pseudoreplicate and the percentage of peaks sharing IDR rank position when filtering the called peaks at number of decreasing pvals were calculated (Figure 1.3 a and b). This strategy was performed at a range of number of reads for each samples (as above) to determine the effect of sequencing depth of the suitability of this peak calling filtering approach. Variation in the optimal pval presenting the largest percentage of IDR shared peaks between the two pseudoreplicates varied when the sequencing depth was lower than 10 million reads (Figure 1.3 a and b), not being appropriate to perform this analysis when . This variation was more pronounced pronounced and extended in the tCD4$^+$, which presented the lower TSS values, when compared to the CD14$^+$ monocytes ATAC-seq libraries (representative examples of CTL2 CD4$^+$ and CD14$^+$ monocytes in Figure 1.3 a and b). The shape of the curves were also influenced by the sample quality. For appropriate sequencing depth (15-20 million reads), the CD14$^+$ monocytes (TSS enrichment $\geq$10) samples presented a profile reaching a single maximum of shared IDR peaks for a particular filtering pval (Figure 1.3 b), which was -log10 pval 8 in all the three samples (data not shwn). In contrast, the same analysis revealed two pvals for which the percentage of IDR shared peaks reached two local maximums in tCD4 (Figure 1.3 a).

Filtering the CD4$^+$ peaks at the pval of the first local maximum (10^-11) for peaks sharing IDR rank reduced the percentage of peaks overlapping noise (e.g heterochromatin, repetitive sequences and repressed regions) when compared to the same annotation using the list of significant peaks filtered based on FDR<0.01 (Figure 1.3 c). In summary, this IDR analysis appeared as systematic method to identify an optimum pval to perform sample-specific filtering of the technically reproducible peaks when the sequencing depth was over 10 million reads. Those

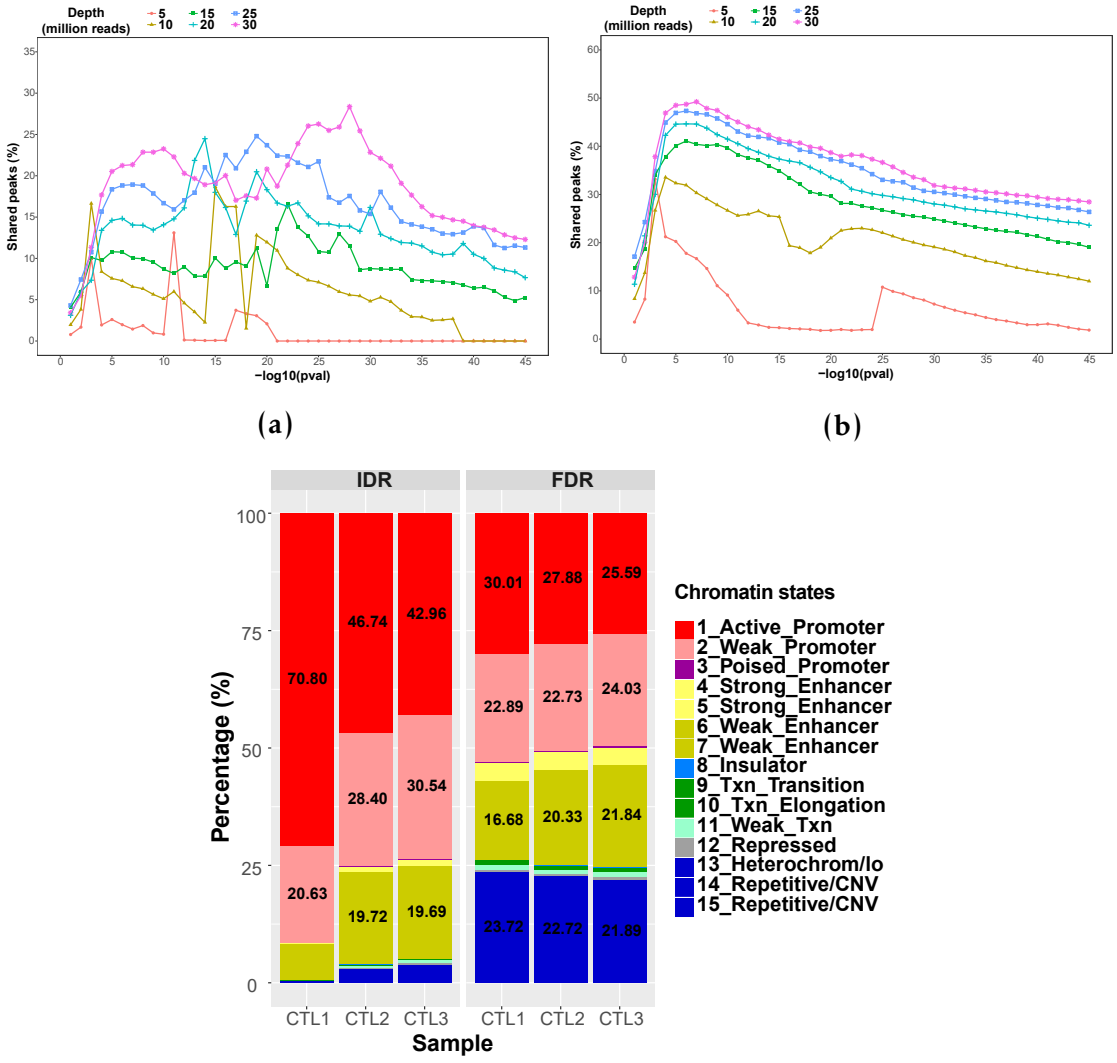


Figure 1.3: Peak calling filtering using IDR analysis in ATAC-seq samples. For each of the sequencing depths tested (from 5 to 30 million of reads after filtering), illustration of the percentage of peaks sharing IDR rank between the two pseuroreplicates when using different pval filtering thresholds in CTL2 a) $CD14^+$ monocytes and b) $tCD4^+$ (used as representative samples differing in quality for this analysis). c) Annotation (in percentage) of the CTL1, CTL2 and CTL3 $tCD4^+$ ATAC-seq peaks filtered for FDR<0.01 or optimal pval from the IDR analysis (pval=10^-11) with the corresponding cell type specific Epigenome Roadmap chromatin segmentation map.

filtered peaks will be used downstream to build the master list of regions across all the samples and perform differential chromatin accessibility analysis.

Differential chromatin accessibility analysis

In this project, a peak-based approach using the number of read counts overlapping the peaks included in the consensus master list (ML_all) was implemented to perform differential analysis. One of the main limitations of the ATAC-seq and Fast-ATAC protocols (discussed in the next section) is the background signal. Therefore, in collaboration with Dr Hai Fang, a method to calculate an empirical cut-off was conducted to minimise the impact of background read counts on the peaks included for the differential analysis (**Xinmin2005**; Jonker et al. 2014). Moreover, due to lack of consensus in terms of normalisation and differential analysis methods in ATAC-seq (Table 1.1), two strategies were tested.

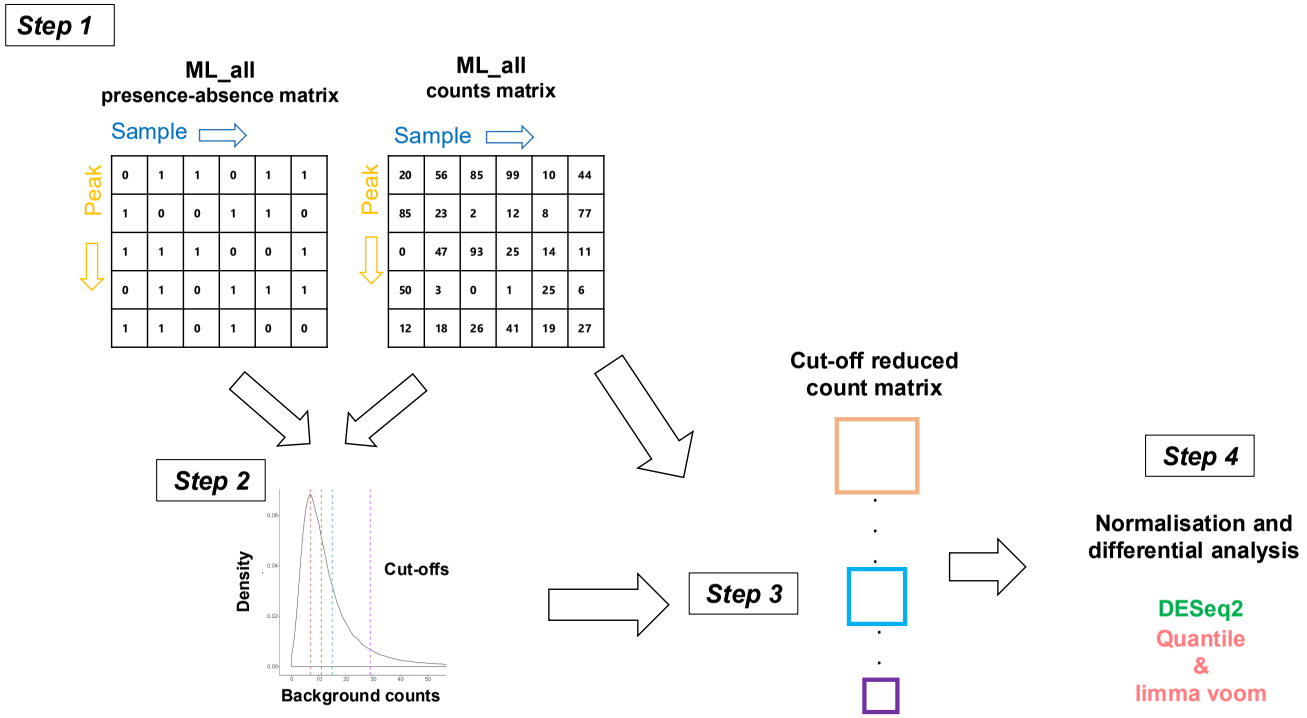


Figure 1.4: Work flow illustrating the strategy to account for ATAC background noise prior to differential analysis. A combined consensus peak master list (ML_all) was built as explained in Chapter **??**, including all the significant peaks from each sample that were shared by at least 30% of the total number of samples included in the analysis. The peaks were further transformed to obtain non-overlapping 500bp homogenous entities. The data regarding the ML_all can be represented by two matrix (Step 1). The first one is the significance matrix with each entry indicating the significance of a peak (in rows) in each sample (in columns) as in presence (1, significant) or absence (0, non-significant). The second matrix is the count matrix storing the number of reads mapped to the peak (in rows) for the each sample (in columns). A density distribution plot was generated with the read counts from the absent peaks (0) in each sample and used to define a sequence of twenty cut-offs illustrating the number of counts showed by a particular percentage of the total absent peaks (background counts) (Step 2). The defined cut-off were used to filter out peaks from the ML_all and generate a serie of a reduced matrix (Step 3) that were tested for normalisation and differential chromatin accessibility analysis by two methods (quantile&limma voom or DESeq2) (Step 4).

From the count matrix for the six samples defined above, the read counts from those peaks that were absent in each sample (since the the ML_all includes peaks being present in at least 30% of the total sample number)(Figure 1.4 Step 1) were used to generate a density distribution plot (Figure 1.4 Step 2). From this plot, a sequence of twenty cut-offs were defined, with each representing the number of counts showed by a particular percentage of the total absent peaks (Figure A.1). Each cut-off was used to filter out those peaks from the ML_all raw count matrix, whose values in more than three samples were lower than the background counts (Figure A.1 Step 3). The three samples were chosen, as it corresponds to the smallest group of replicates in this particular experimental design and ensures peaks absent in one condition were retained. As a result, each cut-off generates a reduced matrix of low-noise peaks (that is, the raw counts ML_all matrix only for remaining peaks) that was normalised using quantile or DESeq2 (library size and variance stabilisation (**Love2014**)) used to conduct differential analysis with limma voom or DESeq2, respectively. Both normalisation methods performed appropriately for all the reduced master lists across the six samples, with slightly greater consistency for the quantile normalisation across the two groups (Figure **??** a and b). Differential chromatin accessibility analysis using quantile&limma voom showed greater number of significant (FDR<0.01 and absolute fold change (abs(FC))>1.5) differential chromatin accessible regions (DARs) compared to DESeq2 across all the cut-off reduced counts matrix (Figure **??** c). The two approaches presented progressive decrease in the number of DARs from the 75% cut-off onward, indicating a progressive reduction in the number of false positive hits reported for peaks with read counts close to the background noise cut-off. Further increases in the cut-off value however are expected to also remove true positives, so an intermediate value is chosen here. Depending on the noise inherent to an experiment this threshold may vary.

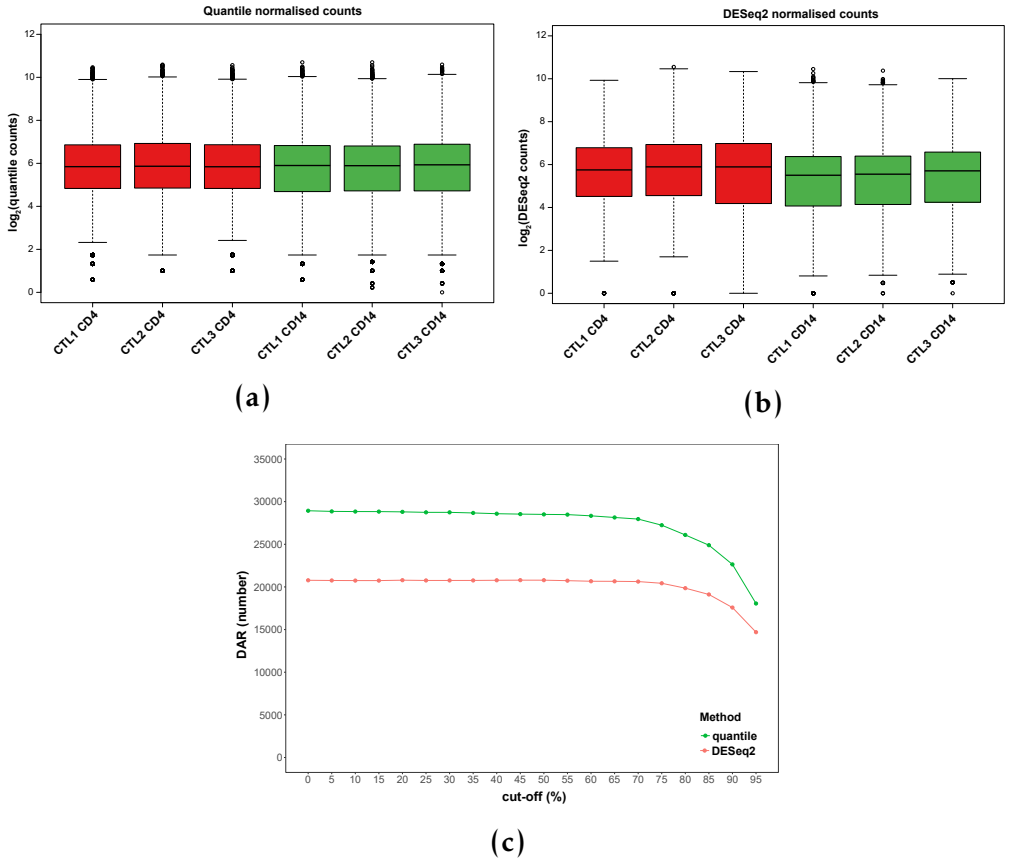


Figure 1.5: Normalisation and differential chromatin accessibility analysis for different cut-offs using quantile normalisation&limma voom and DESeq2. Boxplots representing the log_2FC of the read counts for each of the peaks from the unfiltered master list normalised by a) quantile or b) DESeq2 in the three $CD14^+$ monocytes and total $CD4^+$ healthy control paired samples. c and d) Representation of the number of significant DARs (FDR<0.01 and abs(FC)>1.5) detected in the differential analysis by limma voom (using quantile normalisation) or DESeq2 when using a sequence of empirical background noise cut-offs to filter the peak master list. In d) the upper x-axis represents the total number of peaks for filtered master list at each of empirical background noise cut-offs.

From this analysis, 80% was chosen as a conservative filtering cut-off, with the majority of the 19,855 DESeq2 significant DARs the more conservative method, the most conservative method (DESeq2) recapitulated by limma voom at the same significance threshold (FDR<0.01 and abs(FC)>1.5) (Figure 1.6 a). FDR rank revealed that out of the first 19,855 limma voom DARs 18,768 were the same as the retrieved by DESeq2. Moreover, very significant positive correlation was found between FCs for all the regions included in the 80% cut-off reduced matrix reported by the two methods (R=0.999, p-val=2.2^{-}16) (Figure **??** b). Overall, this suggested that the differences in the number of the significant DARs reported by limma voom and DESeq2 could partly be driven by differences in the multiple testing correction based on FDR.

Lastly, the significant (FDR<0.01 and abs(FC)>1.5) DARs identified by DESeq2 were divided in those more accessible in CD14^{+} monocytes (open in monocytes) or tCD4^{+} (open in CD4^{+}). Enrichment analysis for cell type-specific epigenetic features, including FANTOM5 eRNAs,histone marks and DHSs, was conducted in each of the two groups of DARs. The DARs open in CD14^{+} monocytes included as the top hits for each of the categories T cell eRNAs, CD4^{+} H3Kme1 and H3K27ac and CD14^{+} DHSs (Figure 1.7). Conversely, the top enriched features for open in monocytes DARs included eRNAs, H3K27ac and DHSs in monocytes. Overall, this enrichment analysis confirmed the ability of this this differential analysis method to identify significant and robust DARs that highlight cell-type specific regulatory regions for CD14^{+} monocytes and tCD4^{+} cells.

1.2.2 Assessment of ATAC-seq transposition times and comparison with Fast-ATAC protocol in relevant cell types

In addition to establishing the appropriate pipeline to analyse ATAC-seq data (NGS data processing, QC and differential analysis), the effect in duration

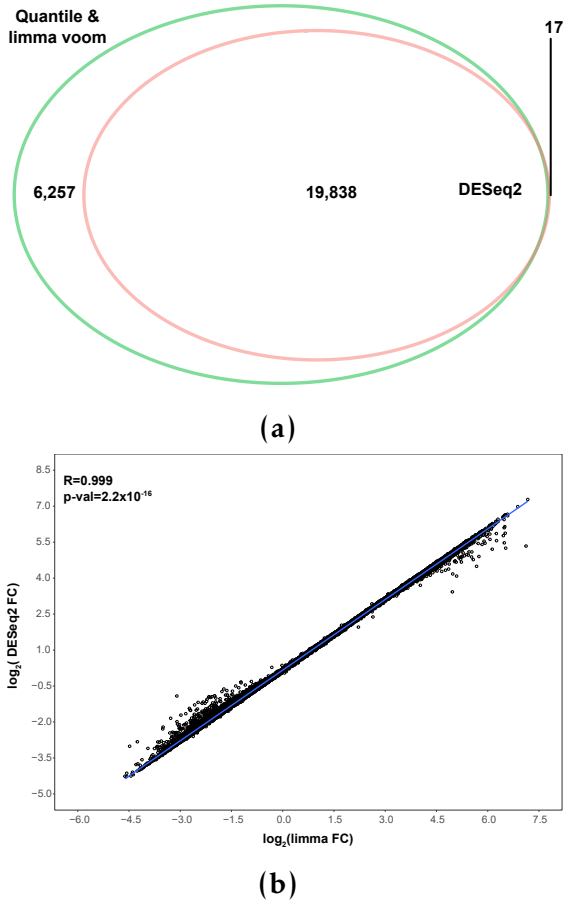


Figure 1.6: Comparison of the DARs identified by differential analysis using limma voom or DESeq for the master list filtered at the optimal cut-off 80%. a) Venn diagram illustrating the common and distinct significant (FDR<0.01 and abs(FC)>1.5) DARs identified by differential analysis in the filtered master list for the 80% optimal cut-off using limma voom or DESeq2. b) Representation of the correlation between limma voom and DESeq2 $\log_2$FCs (no FDR filtered) in each the peaks from the filtered master list at the 80% optimal cut-off. Pearson correlation coefficient (R) and significance (pval) are indicated. Limma voom was applied to quantile normalised count data.

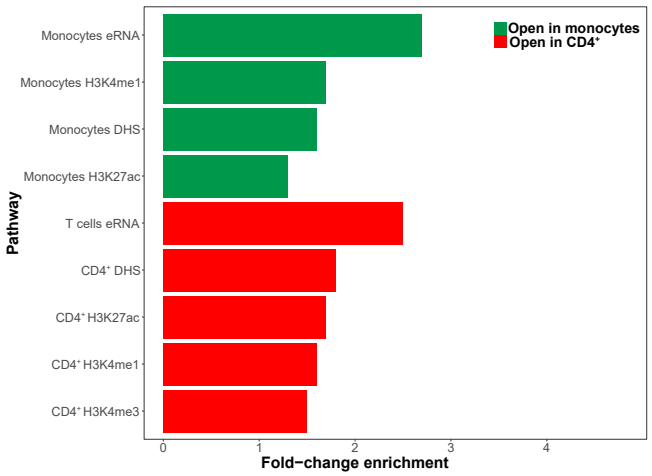


Figure 1.7: Enrichment analysis for the identified DARs by DESeq2 between CD14$^+$ monocytes and tCD4$^+$ cells. Barplot representing the FC for the top significantly enriched (FDR<0.01) FANTOM5 eRNAs and, histone marks and DHSs from Blueprint. Enrichment analysis was performed separately for the significant (FDR<0.01 and abs(FC)>1.5) DARs more accessible in CD14$^+$ monocytes (open in monocytes) or tCD4$^+$ (open in CD4$^+$).

of the transposition reaction was evaluated in the main immune cell types of interest for this project. ATAC-seq was performed for three transposition times (20, 30 and 40 min) in CD14$^+$ monocytes, tCD4$^+$, tCD8$^+$ and CD19$^+$ cells with the impact on varied ATAC measurements explored. ATAC-seq data generated by the three transpostion times produced appropriate fragment size distribution which recapitulated the nucleosome periodicity pattern (Figure 1.8 a). The duration of the transposition times is known to have an effect on the proportion of nucleosome-free and nucleosome-bound (mono-nucleosomes and beyond) regions tagged by the adapters. Ideally, the transposition reaction should maximise NFF (≤150bp), where TF and other proteins bind. In order to explore this effect in the different cell types, the ratio between NRF and nucleosome bound fragments (NBF>150bp) was calculated (Figure 1.8 b). The change in patterns of this ratio across time was heterogenous between cell types. For example, tCD8$^+$ presented the greatest proportion of NFR for 20 min of transposition whereas the NFR/NBR reached the maximum at 40 min for the

$CD14^+$ monocytes. Despite this heterogeneity, the increment in this ratio across transposition times was very moderate in all cell types but $tCD8^+$ cells.

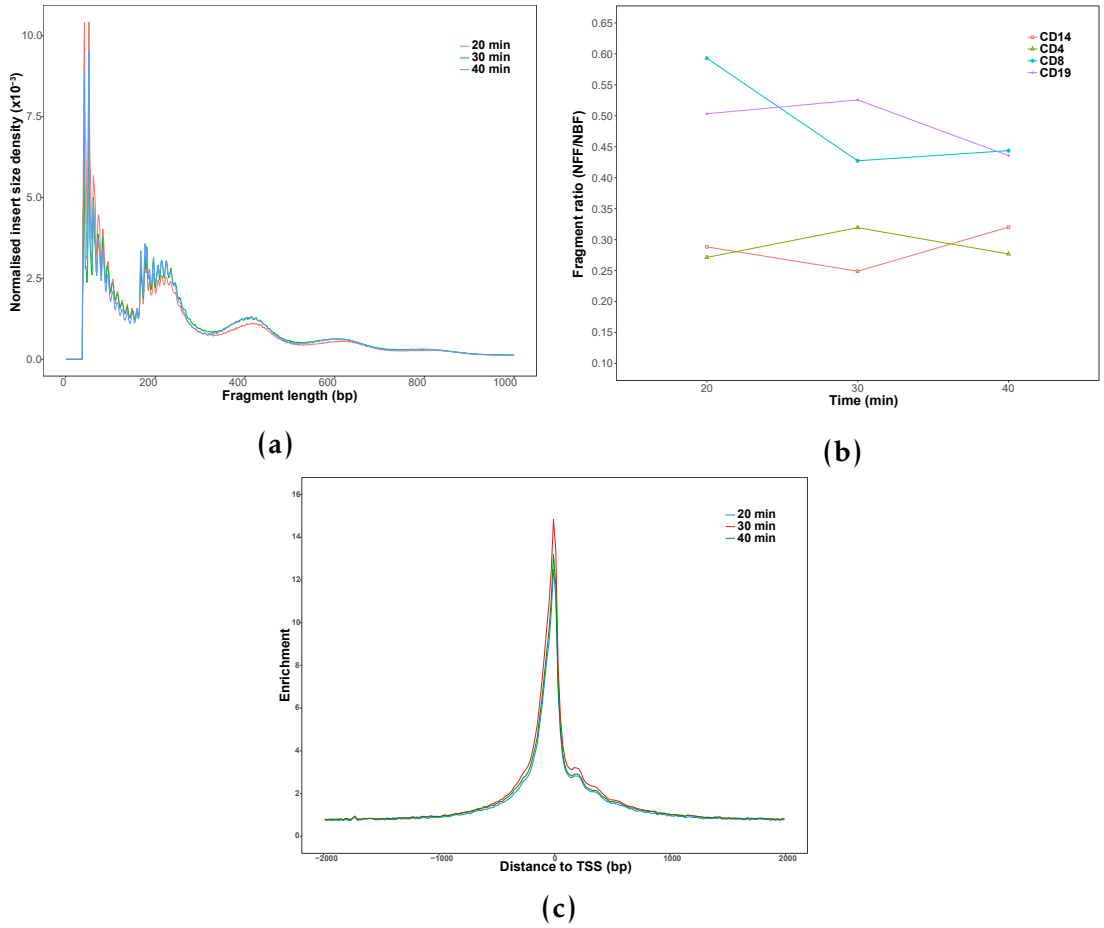


Figure 1.8: Assessment of the effect of transposition times on the ATAC-seq measurements a) Representative plot of the ATAC-seq fragment sizes density distribution following 20, 30 and 40 min of transposition in healthy total $CD8^+$ cells. b) Changes in the ratio between nucleosome-free fragments (NFF) (fragments ≤150bp) and long (>151bp) ATAC-seq nucleosome-bound fragments (NBF) across different transposition times in $CD14^+$ monocytes, $tCD4^+$, $tCD8^+$ and $CD19^+$ cells.

Transposition times did not significantly affect the signal-to-noise ratio measured as fold-enrichment at the TSS and the largest enrichment values showing the best enrichment at different times across the four cell types (Figure 1.8 c). For all cell types, 30 and 40 min yielded the ATAC-seq libraries with the largest enrichment, with the differences between the two being very moderate (not more than 4 units difference) across all cell types (data not shown). Before performing this formal comparison for the transposition times, some

sample recruitment had already been conducted using standard ATAC-seq with transposition for 40 min, as it showed to be the most appropriate condition based on the tapestation profiles from library QC (here not shown). Although in this analysis 30 min showed to be the best condition, the differences between 30 min and 40 min were minor and 40 min was kept for the rest of patients samples recruitment using ATAC-seq.

In addition to testing different transposition times, comparison between the standard ATAC-seq and the improved Fast-ATAC protocol (**Corces2016**) was also conducted. The aim was to reproduce the two main improvements (reduction of MT reads and signal enhancement) from the Fast-ATAC publication before implementing the replacement of the ATAC-seq protocol, used at the time for patients' samples processing. Similarly to the Corces and colleagues data, the percentage of MT reads in the Fast-ATAC libraries was lower for all the cell types compared to ATAC-seq (Figure 1.9 a). Importantly, the MT percentage in $tCD4^+$ and $tCD8^+$ cells showed a reduction from 50% to less than 20%. Regarding the reduction in background signal, opposing trends were observed across cell types (Figure 1.9 b). ATAC-seq TSS enrichment was improved by the Fast-ATAC protocol in $CD14^+$ monocytes (6.9 to 23.2) and $tCD4^+$ cells (14.8 to 17.9). The improvement was particularly relevant in $CD14^+$ monocytes, where the TSS fold enrichment in ATAC-seq was just above the threshold cut-off of 6. In contrast, Fast-ATAC $tCD8^+$ and $CD19^+$ libraries presented a reduction when compared to the ATAC-seq libraries. Since only one repeat was performed and TSS enrichment has been observed to vary from sample to sample using the ATAc-seq, no final conclusion could be drawn from these results. Conversely, the significant the decrease of MT reads together with the reduced duration of the protocol supported the replacement of ATAC-seq by Fast-ATAC for future patient recruitments (see Chapter **??**).

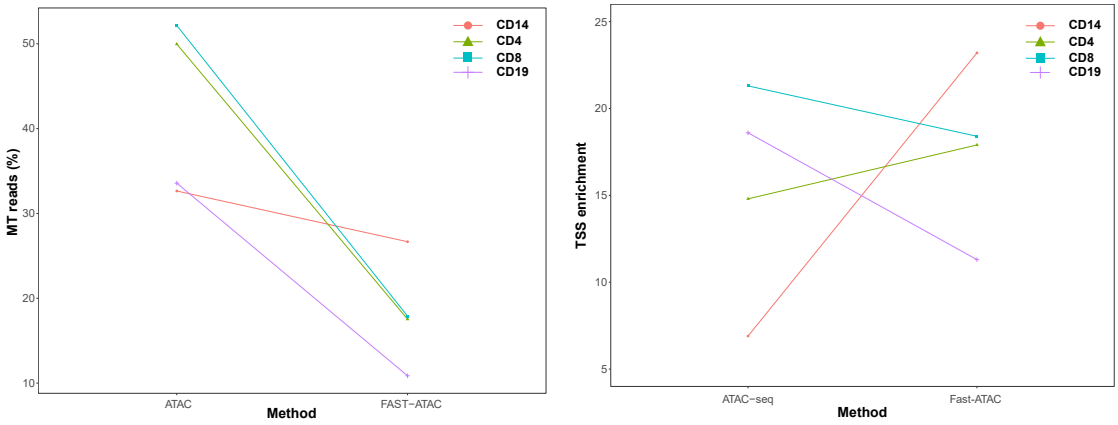


Figure 1.9: Differences in MT DNA abundance and and TSS enrichment between ATAC-seq and Fast-ATAC protocols. Representation of changes in a) percentage of MT reads and b) TSS fold-enrichment between ATAC-seq and Fast-ATAC libraries for $CD14^+$ monocytes, $tCD4^+$, $tCD8^+$ and $CD19^+$ cells.

1.2.3 Limitations of ATAC-seq and FAST-ATAC to assess chromatin accessibility in KC

This project also aimed to characterise the regulatory landscape in KCs, one of the most relevant cell types in psoriasis pathophysiology. In order to determine the performance of the standard ATAC-seq protocol from Buenrostro *et al.*, 2013 (referred to ATAC 1 in this subsection), a cell suspension from a psoriatic lesional skin biopsy was generated and ATAC 1 was performed in 50,000 cells at two different transposition times (30 and 40 min). Since biopsy handling and lesional epidermal KCs are particularly challenging, this was considered the best system to test the performance of the standard protocol in the clinical setting of interest for the study. Library QC based on tapestation profiles for the two samples revealed expected DNA fragment sizes that recapitulated the characteristic nucleosome pattern every ~200bp generated by transposition of nucleosome-free and nucleosome-bound DNA (Figure 1.11 a). This was consistent with the fragment size distribution from the NGS data, presenting NFF and NBF (mono-and di-nuclosomes only) for both transposition times (Figure 1.11 b).

However the relative abundance of the mono-nucleosome fragments appeared to have the same density as the NFF, which is not observed in higher quality libraries. Regarding the signal-to-noise ratio, both of the transposition times libraries presented TSS fold enrichment below the acceptable cut-off threshold of 6, showing slightly better signal (3.5 fold-enrichment) the 30 min library (Figure 1.11 c). Cell suspension obtained from skin biopsies using trypsinisation of the epidermal layer are enriched in KC(∼90%). However, they also contain significant amounts of dead cells and free-DNA releases by apoptotic cells which could contribute to the increased background noise observed above.

Following Buenrostro and colleagues ATAC-seq protocol, a modified version for KCs by Bao and colleagues (named ATAC-seq 2 in this subsection) and the Fast-ATAC protocols was released (**Bao2015**; **Corces2016**). Interestingly, Bao's protocol was applied directly on the cell culture plate containing adherent NHEKs, avoiding a trypsinisation step that could increase cell death. In line with Bao's paper, to prevent background noise due to presence of dead in the assessment of the different protocols, two systems were implemented. First, all the cells obtained from the epidermis of control skin biopsies were cultured for 3h in a 96-well plate and washed afterward to minimise the presence of apoptotic cells. This procedure known as adherent assay allows to isolate viable undifferentiated KCs. Second, cultured NHEKs (50,000 cells) were also incorporated as a control to test the performance of the different ATAC protocols. The three ATAC protocols (ATAC 1, ATAC 2 and Fast-ATAC using C1 conditions, Table 1.3) were performed directly on the plate adherent cells with no cell detachment step.

For the three protocols, the library size distribution of sequenced fragments showed presence of the NFR and a poorly defined nucleosome pattern, particularly for the ATAC 2 protocol in both NHEKs and adherent KCs from skin biopsies (Figure **??** a). Similarly to the results in cell suspensions from

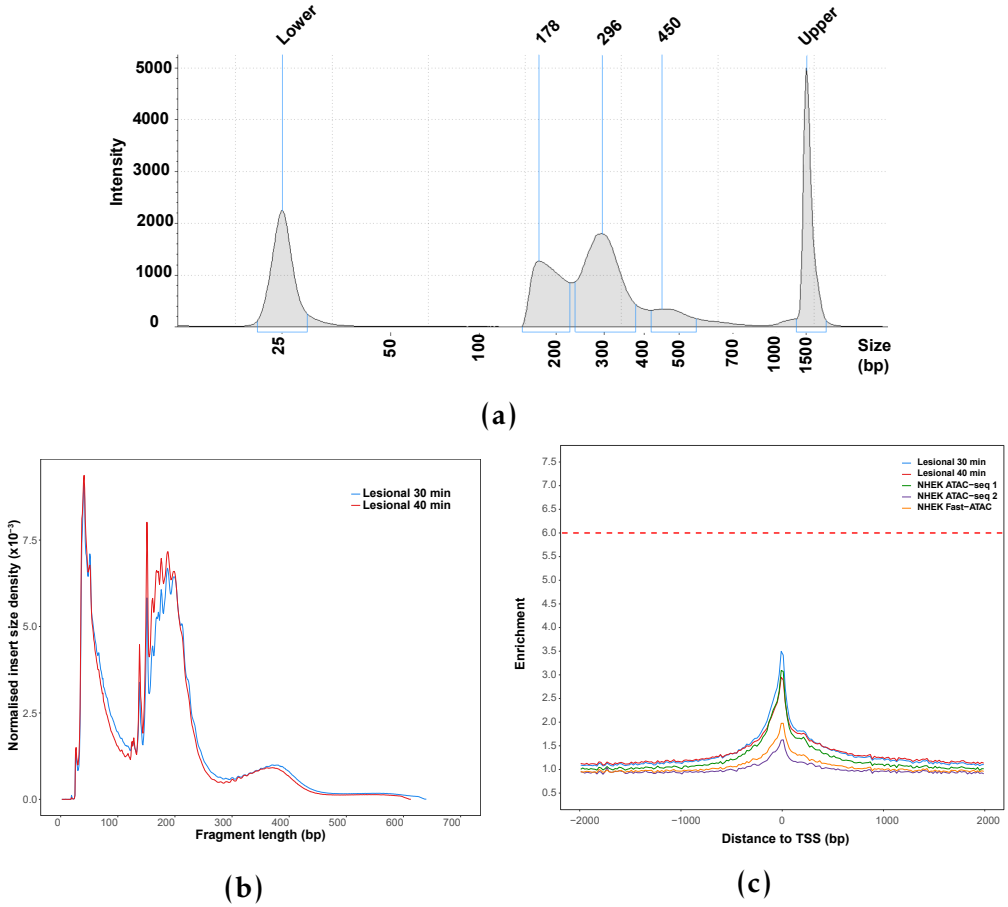


Figure 1.10: QC assessment of different ATAC protocols in psoriasis KCs and NHEKs. a) Pre-sequencing quantification of DNA fragment sizes in the ATAC libraries (tapestation profile) and b) the density distribution of sequenced fragments for ATAC 1 libraries generated in 50,000 KCs in suspension isolated from one psoriasis patient lesional skin. Two transposition times (30 and 40 min) were tested in the same sample KCs suspension and only the tapestation profile for the 30 min transposition has been included as the representative one. c) Fold-enrichment of ATAC fragments across the Ensembl annotated TSS from the ATAC 1 psoriasis lesional KCs libraries (previously mentioned in a and b) and NHEK libraries generated with ATAC 1, ATAC 2 and Fast-ATAC protocols performed directly on the 96-well plate adherent cells.

Protocol	Lysis and transposition	Key parameters
ATAC 1 (**Buenrostro2013**)	Two steps	0.1% NP-40 and 2.5µL Tn5
ATAC 2 (**Bao2015**)	Two steps	0.05% NP-40 and 5µL Tn5
Fast-ATAC (**Corces2016**)	One step	C1*: 0.01% digitonin, 2.5µL Tn5 C2: 0.01% digitonin, 0.5µL Tn5 C3: 0.025% digitonin, 0.5µL Tn5 C4: 0.025% digitonin, 2.5 µL Tn5

Table 1.3: Description of the most relevant parameter from the ATAC-seq and FAST-ATAC protocols assayed in NHEK and skin biopsies. Transposition time for all the different protocols was 30 min.(*) corresponds to the original Fast-ATAC conditions from Corces *et al.*, 2016.

psoriatic lesional epidermis, TSS enrichments were very low, particularly for the ATAC 2 protocol, and in all the instances under the acceptable cut-off in the two cell systems (Figure 1.11 c and Figure A.2).

Further optimisation of Fast-ATAC was performed modifying the original concentration of the NP-40 detergent and the Tn5 enzyme (Table 1.3). Library QC using tapestation profiles to assess the DNA fragment size distributions failed to show the nucleosome pattern profile expected in ATAC, thus not proceeding for NGS (Figure A.3).

Towards the end of experimental work for this project, a new protocol known as Omni-ATAC was published (**Corces2017**). Omni-ATAC was a protocol suitable for every cell type, in contrast to ATAC 1 and Fast-ATAC optimised for hematopoietic cells (**Buenrostro2013**; **Corces2016**). Performance of this protocol in 50,000 viable NHEKs in suspension yielded the expected fragment size distribution for sequenced fragments, with the greatest abundance for NFR followed by mono and di-nucleosome fragments (Figure 1.11 b). Moreover, high TSS enrichment values (approximately 20 fold) were observed for the two replicates (Figure 1.11 c). When performing overlap between the Omni-ATAC

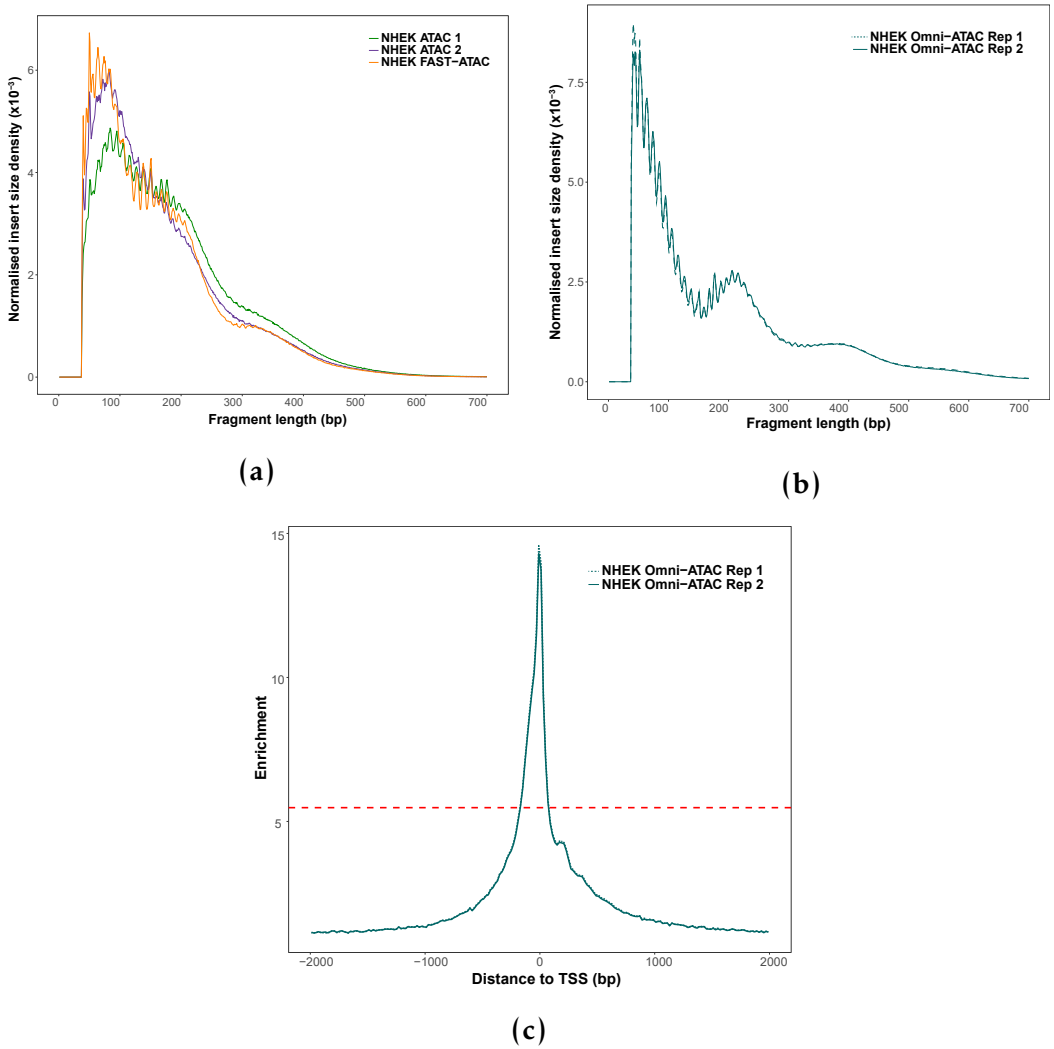


Figure 1.11: QC assessment of Fast-ATAC and Omni-ATAC in cultured NHEK. Representation of the fragment sizes density distribution in NHEKs libraries generated using a) ATAC1, ATAC2 and Fast-ATAC or b) Omni-ATAC protocols. b) Fold-enrichment of ATAC fragments across the Ensembl annotated TSS from two Omni-ATAC technical replicates.

sample peaks filtered for a low stringency pval (pval=0.01) and the ENCODE DHSs from 125 cell types, the highest percentage (approximately 55%) was reported for NHEKs (Figure 1.12 a).

In contrast, the same analysis for the ATAC 1, ATAC 2 and Fast-ATAC peaks only showed 20% or less overlap with ENCODE NHEKs, supporting the higher quality and specificity of Omni-ATAC accessible regions in KCs, even at a very low pval filtering. The differences in the quality of ATAC signal between Omni-ATAC and the prior generation of protocols was clearly observed at the chr17 locus harbouring a number of keratin genes, which are the main components of KCs' cytoskeleton (Figure 1.12 b). Omni-ATAC clearly presented the lowest background noise, the highest signal intensity and the greatest number of high quality peaks across the different keratin (KRT) genes when compared to the samples generated with the other ATAC protocols. Overall, this data was consistent with Corces *et al.*, 2017, where consistent successful results in NHEKs were shown, and it encourages future testing of Omni-ATAC in KCs from psoriasis patients biopsies processed through adherent assay to minimise the presence of dead cells.

1.2.4 Effect of cryopreservation and fixation in the chromatin landscape of immune primary cells

Experimental design and sample description

As previously introduced, research using clinical samples represents a logistical challenge. In the context of this thesis two different approaches were of interest and a collaborative project was established with High-Throughput Genomics at the WCHG. On one side, the cryopreservation of PBMCs in liquid nitrogen using DMSO followed by thawing, recovery and FACS isolation of the cell population of interest (Figure 1.13). On the other side, the performance of an optimised protocol developed by High-Throughput Genomics using DSP in

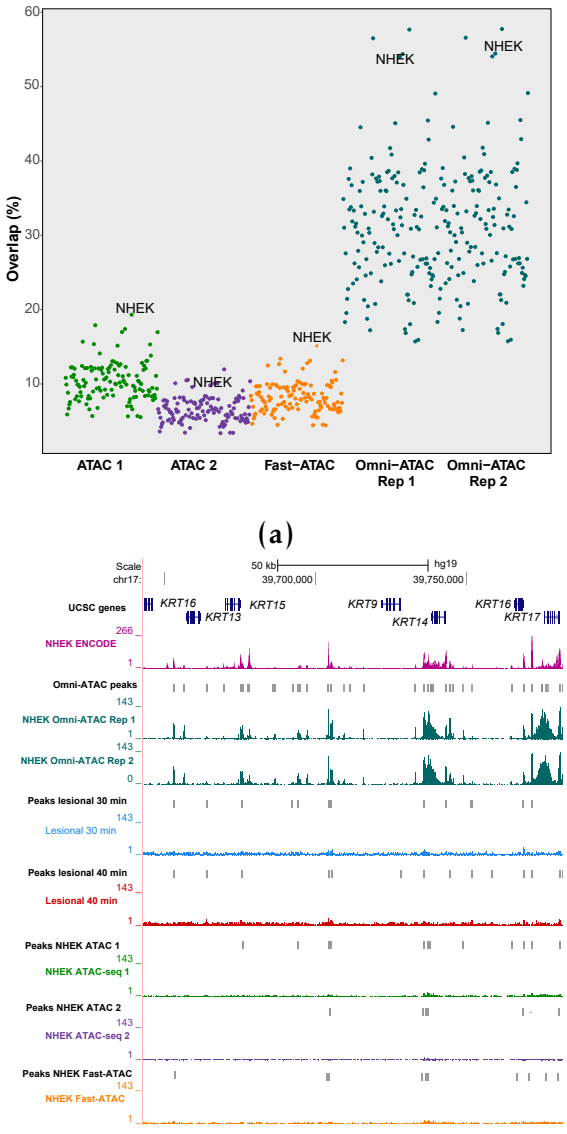


Figure 1.12: Comparison of the ENCODE DHSs overlap and signal density at the chr17 keratin family gene locus across different ATAC protocols. For the psoriasis lesional KCs and the NHEK libraries generated with the different ATAC protocols a) represents the percentage of overlap between significant peaks from each sample using low stringent pval and open DHS chromatin regions in 125 ENCODE cell types; and b) UCSC Genome Browser view illustrating the normalised read density (y-axis) at the chr17 locus (x-axis) containing several genes from the keratin (KRT) family.

scRNA-seq (**Attar2018**) was investigated as a short term preservation method for FACS-isolated relevant cell types .

In order to investigate the performance of these two strategies, three healthy volunteers matched for sex and age were processed on different days to simulate the experimental design when using patients samples (Figure 1.13). PBMCs were isolated from blood and a fraction was stained with the appropriate panel of Abs to isolate $CD14^+$ monocytes and $tCD4^+$ T cells, as detailed in Chapter **??**. ATAC-seq was performed on 50,000 $CD14^+$ and $tCD4^+$ freshly isolated cells (Figure 1.13 Day 0, ATAC-seq fresh). A fraction of the FACS-sorted $CD14^+$ and $tCD4^+$ cell were appropriately fixed with DPS, stored at 4°C for 24h and processed for ATAC-seq (Figure 1.13 Day 1, ATAC-seq fixed). Lastly, an aliquot of the PBMCs were cryopreserved on the day of collection and stored in liquid nitrogen followed by thawing and recovery in culture for approximately 30 min (as detailed in Chapter **??**). After recovery, PBMCs were stained with the appropriate Ab panel and $CD14^+$ and $tCD4^+$ cells were isolated to perform ATAC-seq (Figure 1.13 Day 7, ATAC-seq frozen). For each control sample, three matched ATAC-seq libraries were generated in the two cell populations: ATAC-seq fresh, ATAC-seq fixed and ATAC-seq frozen.

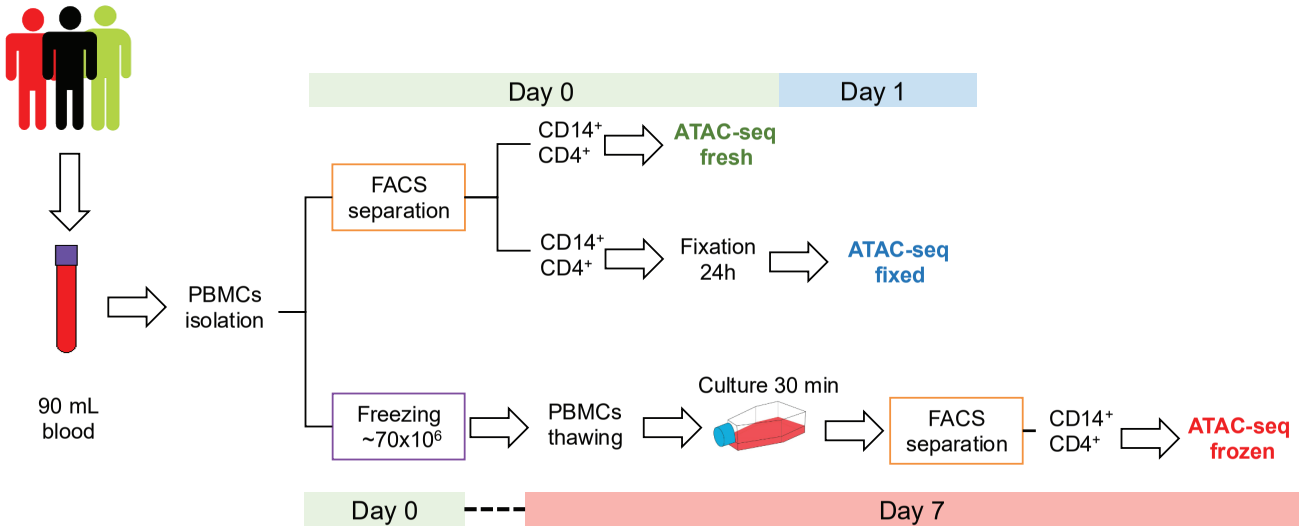


Figure 1.13: Experimetal design to assess the impact of cryopreservation and fixation in the chromatin accessibility of immune primary cells. Three healthy control individuals were recruited in independent days and PBMCs were isolated from 90mL of blood (Day 0). In Day 0, a fraction of PBMCs were used for FACS staining and isolation of 50,000 $CD14^+$ and $tCD4^+$, which were directly processed for ATAC-seq (ATAC-seq fresh). Also in Day 0, a 50,000 FACS-sorted $CD14^+$ and $tCD4^+$ cells were fixed with DPS, stored at 4°C for 24h and processed for ATAC-seq in Day 1 (ATAC-seq fixed). Lastly, in Day 0 a fraction of the PBMCs ($70x10^6$ million cells) were cryopreserved in DMSO and slow-cooling. At day 7 of storage in liquid nitrogen, PBMCs were thaw, recovered in culture for 30 min and stained with FACS Abs to isolate 50,000 $CD14^+$ and $tCD4^+$cell to perform ATAC-seq (ATAC-seq frozen).

Chromatin structure characterisation in the different conditions

All samples from each of the two cell types presented more than 15 million reads, which have previously been shown as the minimum to proceed with appropriate ATAC-seq analysis and peak calling (Figure 1.14). For $CD14^+$ monocytes the median of reads across the fresh, frozen and fixed were more similar (58.6, 64.2 and 39.6 million reads, respectively) (Figure 1.14 a) than in the $tCD4^+$ samples, where the frozen and fixed presented lower median of total million reads compared to the controls (43.8, 32.9 and 28.8 million reads respectively)(Figure 1.14 b).

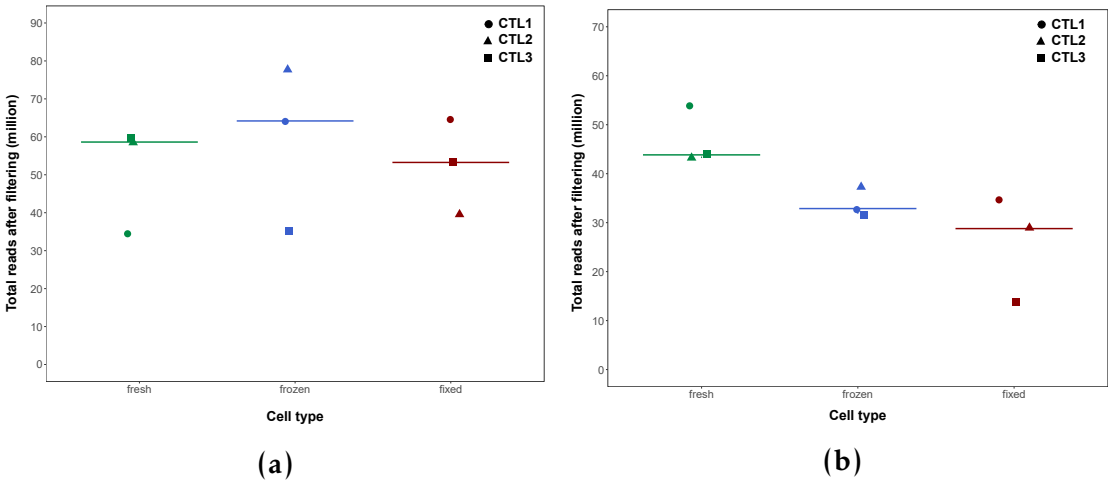


Figure 1.14: Total number of ATAC-seq reads for the fresh, frozen and fixed $CD14^+$ monocytes and total $CD4^+$ samples. Representation of the million of reads after filtering for the fresh, fixed and frozen ATAC-seq libraries in a) $CD14^+$ monocytes and b)$tCD4^+$ cells.

The ATAC-seq signal-to-noise ratios across the TSS presented a similar median for the fresh and fixed $CD14^+$ monocytes libraries (17.4 and 16.5 fold-enrichment, respectively) and higher for the frozen samples(26.3 fold-enrichment)(Table A.1). The TSS enrichment in the frozen and fixed $tCD4^+$ samples were considerably higher (16.1 and 14.3 fold-enrichment, respectively) than the median of the fresh samples (5.6), borderline for the ENCODE recommended threshold. Interestingly, the CTL1 fixed samples of both cell

types presented considerably lower TSS enrichment (2.5 and 7.9, respectively) compared to the other fixed samples (Table A.1).

In terms of the fragment size distribution, the profiles of all the samples, except fixed CD14$^+$ monocytes and tCD4$^+$ from CTL1, were similar, showing NFR <150bp and fragments corresponding to mono-, di-, tri- and tetra-nucleosomes (Figure 1.15 a and b). The fixed samples presented lower density of NFF when compared to fresh and frozen, which appeared to show very similar distributions in both cell types as expected. Particularly, fixed CD14$^+$ monocytes and tCD4$^+$ in CTL1 had extremely low abundance of NFF (Figure 1.15 red-dashed line in a and b), consistent with the very low TSS enrichment for CTL1 CD14 ATAC-seq fixed, previously highlighted.

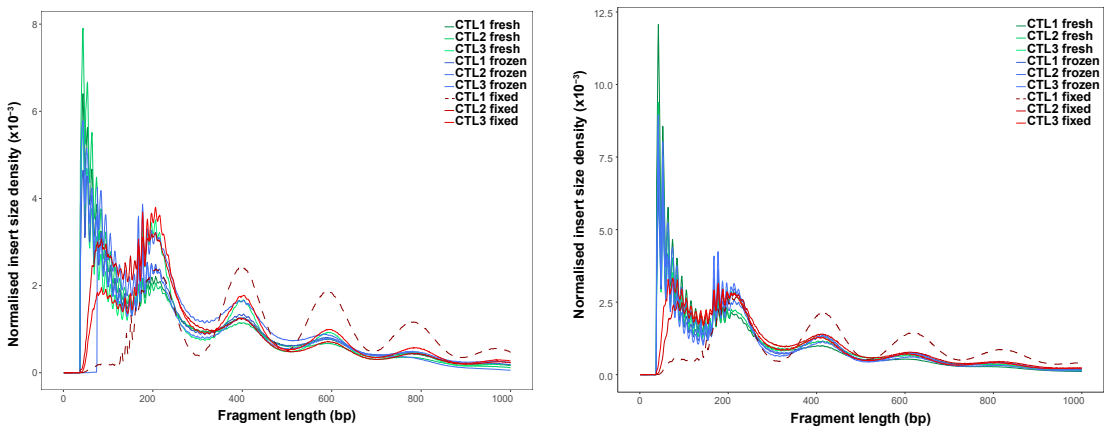


Figure 1.15: Fragment size density distribution for ATAC-seq fresh, fixed and frozen in CD14$^+$ monocytes and tCD4$^+$ cells. The distribution of ATAC-seq fragment's length are illustrated for a) CD14$^+$ monocytes and b) tCD4$^+$ cells and colour-coded by condition (fresh=green, frozen=blue and fixed=red).

Study of maintenance of chromatin structure across and within the TSS was conducted as described in (**Scharer2016**). The nucleosome-free fragments (<150bp) from all the samples showed a single peak of enrichment at the nucleosome-depleted TSS position, with tCD4$^+$ fresh samples presenting the lowest enrichment (Figure 1.16 a and c). The pattern of enrichment of di-nucleosome fragments (ranging between 260 and 340bp) demonstrated in the

majority of the samples a characteristic periodicity in the TSS surroundings, with two peaks of enrichment mapping at the up-stream and down-stream positioned nucleosomes (Figure 1.16 b and d).

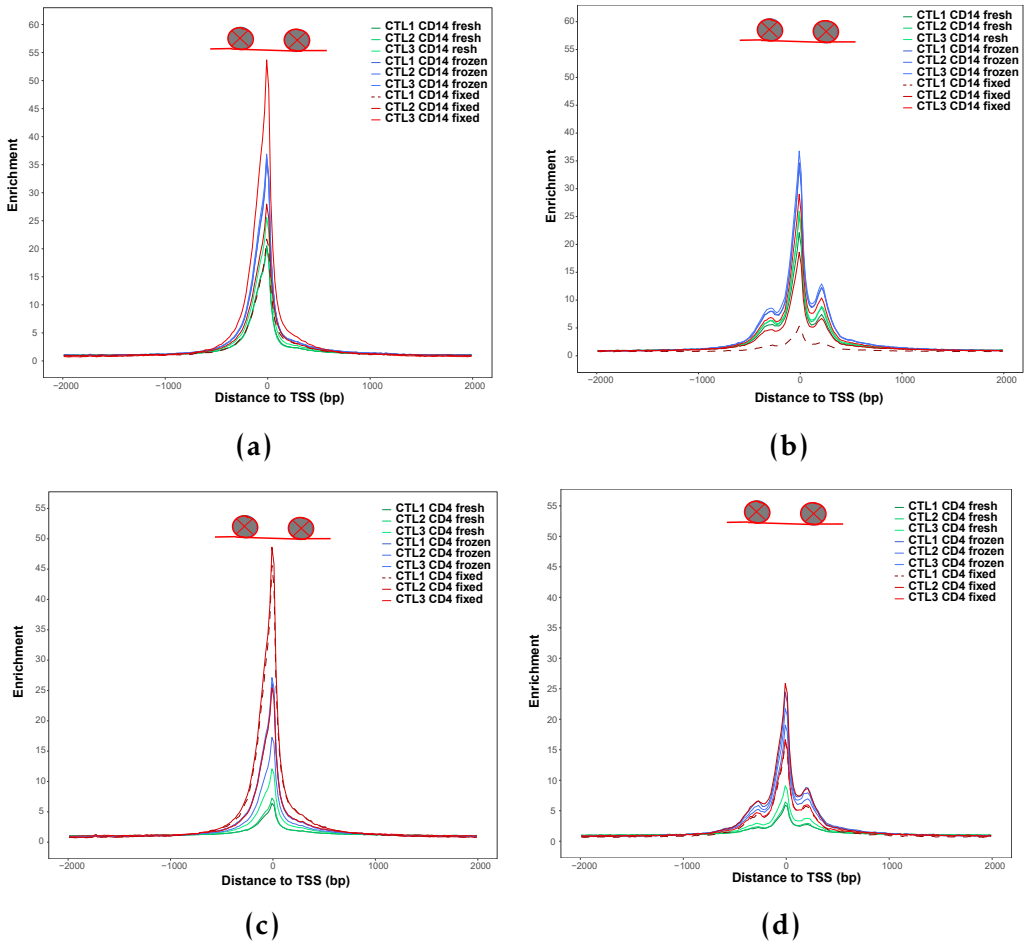
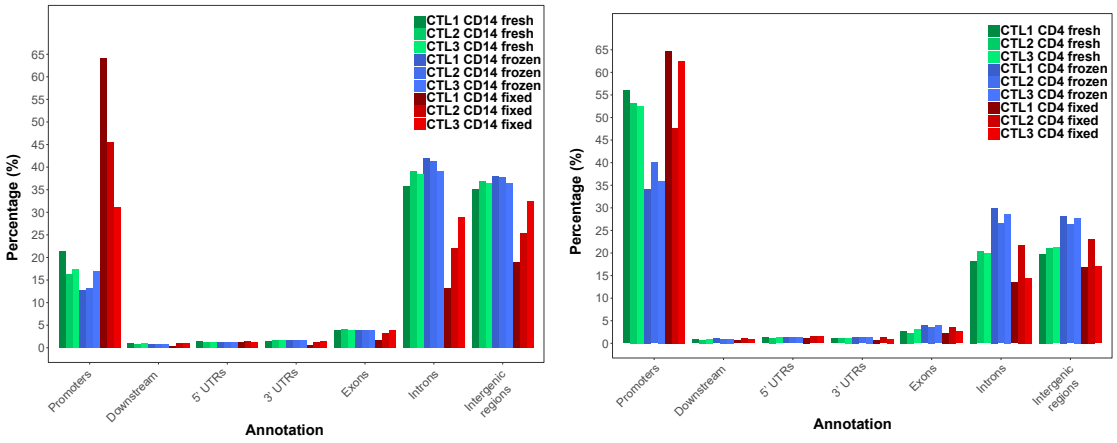


Figure 1.16: ATAC-seq enrichment of nucleosome-free and di-nucleosome fragments at the TSS and surroundings in CD14$^+$ monocytes and tCD4$^+$ samples for the three conditions. Nucleosome-free fragments (<150bp) and di-nucleosome (between 260 and 340bp) were selected *in silico* and enrichment analysis was carried out +/−1Kb across all the Ensembl annotated TSS.

Although fixation in CTL1 CD14$^+$ reduced the abundance of nucleosome-free tagged fragments, the ATAC-seq signal at those regions was clearly enriched when compared to the background. In contrast, DSP in CTL1 CD14$^+$ appeared to increase the efficiency of ATAC-seq to tag NBF (di-, tri- and tetra-nucleosomes) but the loss of chromatin structure around the TSS indicated that those fragments are likely have been displaced from their original location. Conversely, the fresh

$CD4^+$ samples presented low enrichment for the nucleosome-free fragments in the TSS, recapitulating their overall TSS enrichment (Table A.1) and still recapitulated weakly the two nucleosome positioned in the TSS surroundings. Altogether, freezing and, importantly fixing (with exception of the two cell types from CTL1) appeared to maintain the overall chromatin structure.

Annotation of the significant peaks from each sample (filtered for the optimal IDR pval as explained in 1.2.1) demonstrated location of the greatest percentage of ATAC-seq peaks at promoters, introns and intergenic regions(Figure 1.17 a and b), consistently with previous studies (**Buenrostro2013**; **Scharer2016**). The higher percentage of peaks annotated as promoter for some ATAC-seq libraries revealed a preferential location of strong good quality peaks at this feature when compared to other genomic features. Overall, the genomic annotation of the peaks across the three conditions revealed accessible ATAC-seq regions at meaningful genomic features.

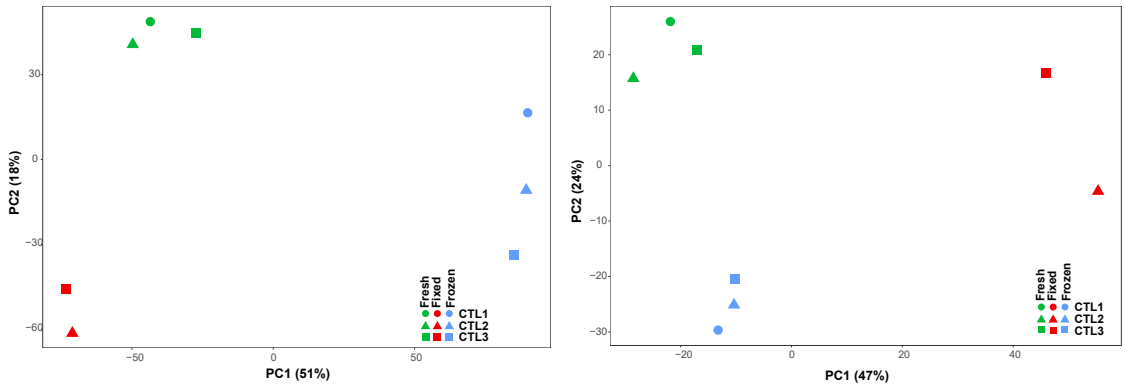


(a) **(b)**

Figure 1.17: Genomic features annotation for the ATAC-seq peaks called in each of the fresh,frozen and fixed samples from $CD14^+$ monocytes and $tCD4^+$. Overlap was performed between the genomic features and the list of a) $CD14^+$ monocytes and b) $tCD4^+$ peaks filtered for FDR<0.01 in each sample from each of the three conditions (fresh=green, frozen=blue and fixed=red).

Differential analysis demonstrates discrete significant changes in chromatin accessibility across conditions

In order to investigate genome wide differences between ATAC-seq fresh (biological reference) and ATAC-seq frozen and fixed, read counts were retrieved at the peaks from a consensus master list including the three conditions for each cell type (ML_CD14_all_cond and ML_CD14_all_cond). For this analysis CTL1 fixed samples from $CD14^+$ and $tCD4^+$ cells were removed from the analysis given the low quality and alterations in the chromatin structure previously described.



(a) (b)

Figure 1.18: PCA analysis based on the ATAC-seq chromatin accessibility landscape in fresh, fixed and frozen samples. PCA analysis was performed using the normalised counts across the consensus master list of the combined fresh, fixed and frozen samples (ML_CD14_all_cond and ML_CD4_all_cond) in a) $CD14^+$ monocytes or b) $tCD4^+$ cells from the same three healthy individuals.

PCA analysis performed using the DESeq2 normalised counts for each of the two master lists demonstrated differences across the three ATAC-seq conditions in the two cell types. PCA analysis showed the sample clustering based on condition, which explained the largest variability in the two cell types. The first PC explaining 51% of the variance in the chromatin accessibility landscape across samples, separated ATAC-seq fresh and fixed from the ATAC-seq frozen samples in $CD14^+$ monocytes (Figure 1.18 a). The second PC (16% variance) showed more moderate changes between fresh and fixed libraries. In contrast, PCA analysis in $tCD4^+$ showed fixed to present the largest differences

with fresh in the genome-wide accessibility landscape in comparison to frozen samples (Figure 1.18 b).

To further compare chromatin accessibility across the three conditions, comparison of the normalised read counts at each of the ML_CD14_all_cond and ML_CD4_all_cond peaks between fresh and fixed or frozen was performed. The majority of the regions showed highly correlated ATAC-seq normalised counts between fresh and frozen or fixed, with the lowest correlation (R=0.918) found between fresh and frozen CD14$^+$ monocytes (Figure 1.19 a, b, c and d).

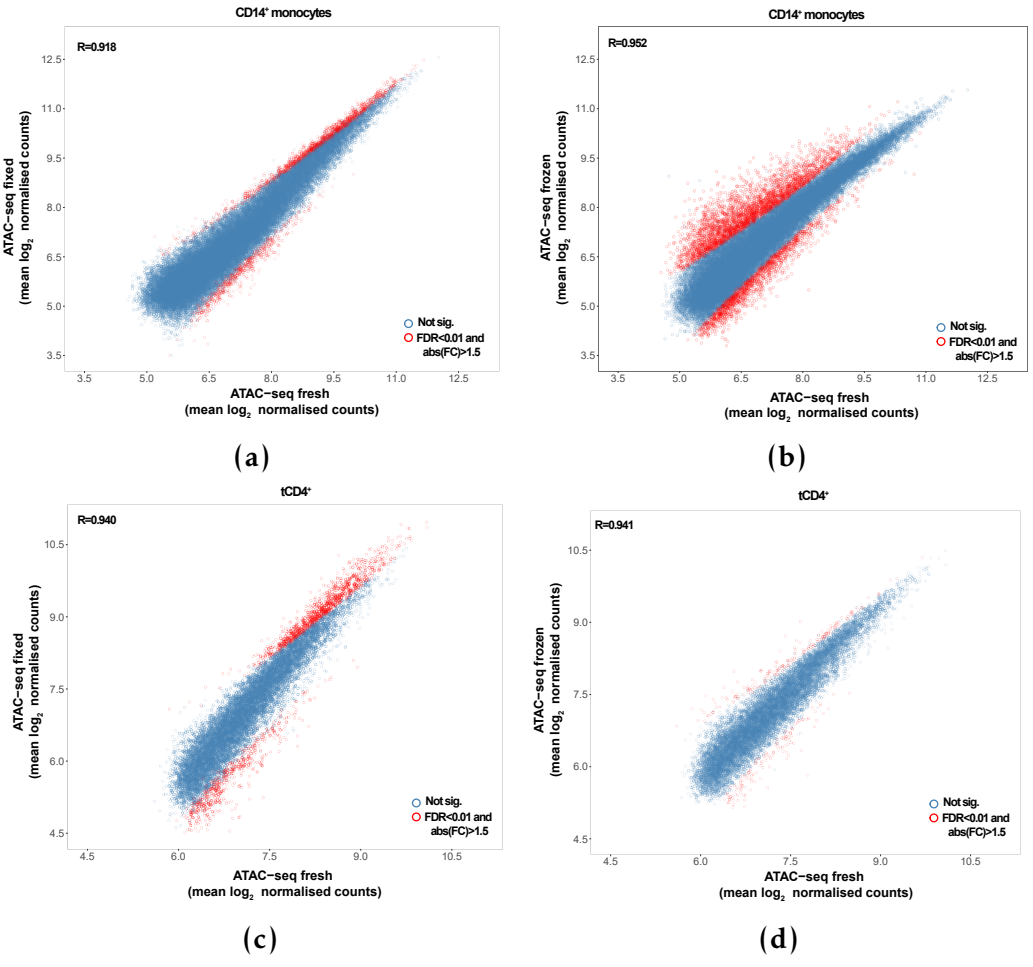


Figure 1.19: Comparison of the $\log_2$ normalised ATAC-seq counts at the consensus master lists peaks in fresh, fixed and frozen conditions. Each plot presents the comparison of ATAC-seq $\log_2$ mean normalised counts from the ML_CD14_all_cond or ML_CD4_all_cond filtered for background noise (80% empirical cut-off) between a) and c) fresh versus fixed or b) and d) fresh versus frozen samples. Pearson correlation coefficient (R) is indicated.

The regions showing the lower correlation in mean counts appeared to be significant DARs (FDR<0.01 and abs(FC)>1.5) when performing differential chromatin accessibility analysis between ATAC-seq fresh and ATAC-seq fixed or frozen at each of the ML_CD14_all_cond and ML_CD14_all_cond regions using DESeq2. The number of significant DARs (FDR<0.01 and abs(FC)¿1.5) reported in each of the comparisons mirrored the PCA analysis results (Table 1.4). In $CD14^+$ monocytes, the largest number of DARs with fresh were reported for the ATAC-seq frozen (5,269 regions in Figure 1.19 b). Conversely, in $tCD4^+$ the greatest differences in chromatin accessibility were found between fresh and fixed ATAC-seq samples (1,564 DARs in Figure 1.19 d). In the performance of the differential analysis, the limited samples size (only two ATAC-seq fixed libraries) and the borderline quality of the ATAc-seq $tCD4^+$ samples partially skwed the normalisation process, as it can be observed by the location of the significant DARs in Figure 1.19 d. Altogether, the number of DARs identified in each comparison did not account for more than 12.2% of the total ATAC-seq regions included in the analysis, representing a discrete proportion of the genome-wide accessible regions studied in this two cell types.

Cell type	Fresh vs Frozen	Fresh vs Fixed
$CD14^+$ monocytes	5,269 (12.9%)	1,838 (4.5%)
$tCD4^+$	282 (2.1%)	1,564 (14.2%)

Table 1.4: Summary results from the differential chromatin accessibility analysis comparing ATAC-seq frozen or fixed chromatin landscape to the reference ATAC-seq fresh. Number of significant DARs (FDR<0.01 and abs(FC)>1.5) identified using ML_CD14_all_cond and ML_CD14_all_cond filtered for background counts using 80% cut-off (optimal cut-off identified for this analysis, data not shown). In brackets are shown the percentage represented by the DARs over the total number of regions included in the differential analysis.

An example of differences in chromatin accessibility between fresh and frozen ATAC-seq libraries in $CD14^+$ included four regions within and downstream the *TNFSF14* gene (Figure 1.20. TNFSF14 is the ligand for a receptor

from the TNF-receptor superfamily. TNFSF14 is involved in T cell activation, induction of apoptosis and also in bone destruction mediated by monocytes and synovial cells interactions in RA. Three of the DARs were located at the promoter, an exon and the 3'UTR of the gene (respectively) and one was found at approximately 5Kb upstream the gene. All DARs were more open in ATAC-seq frozen libraries when compared to fresh and did not show any changes between fresh and fixed conditions.

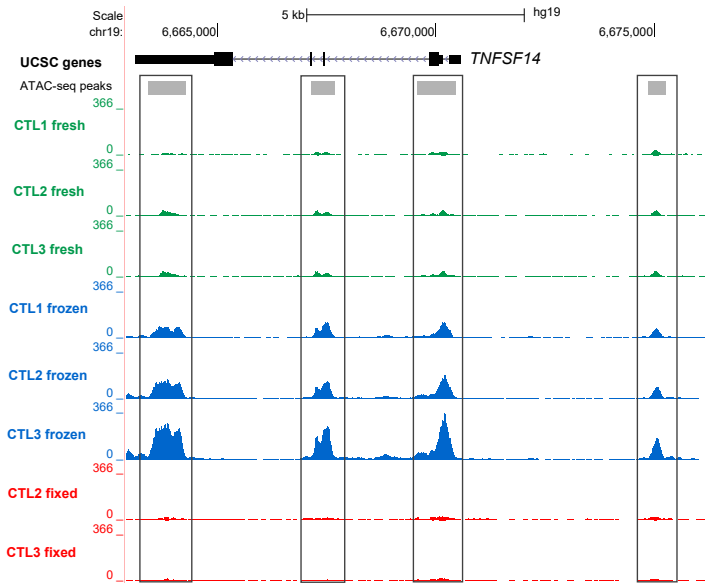


Figure 1.20: Differential chromatin accessibility at the *TNFSF14* gene between ATAC-seq fresh and ATAC-seq frozen in CD14$^+$ monocytes. UCSC Genome Browser view illustrating the normalised read density (y-axis) at four significant (FDR<0.01 and abs(FC)>1.5) DARs (x-axis) within and upstream the *TNFSF14* gene in CD14$^+$ monocytes. The four DARs were more accessible in ATAC-seq frozen when compared to ATAC-seq fresh. ATAC-seq fixed was similar to ATAC-seq fresh at these four locations. Tracks are colour-coded by condition(green=fresh, blue=frozen and red=fixed).

1.3 Discussion

The aim of this chapter was to establish a novel technique and data analysis pipeline for the first time in the group. As such, an exhaustive evaluation of all possible methods was beyond scope of the project. Instead, to select appropriate methods for clinical studies, where sample availability and quality may be

severely limiting, a number of alternative protocols, metrics and algorithms described in early ATAC experiments were evaluated to some extent. This enabled to establish a pipeline to be implemented in **??** and **??** for investigation of psoriasis and PsA chromatin landscape.

ATAC: technical aspects and pipeline establishment

The subsequent release of new ATAC protocols as well as alternative protocols such as THS-seq to assess chromatin accessibility in low number of cells has confirmed some limitations of the ATAC-seq and also Fast-ATAC protocols. Quality assessment and variability across samples was difficult to detect through pre-sequencing library QC based on relative abundance of the different fragment sizes(tapestation profiles). Successful tapestation profiles showing nucleosome patters would still lead to libraries with high background noise when visualising read density in the UCSC Genome Browser. This required the identification and establishment of appropriate data analysis and QC measures beyond pre-sequencing library QC.

In this chapter different quality metrics were explored, including TSS enrichment and FRiP, finding both correlating well with the overall differences in sample quality from the ATAC-seq libraries used as an examplar in the present chapter. Importantly, TSS and FRiP showed to be independent of sequencing depth, and therefore can be applied in low depth sequenced samples when performing optimisation or preliminary QC before increasing the coverage, as also recently shown in other studies (**Corces2017**). Similarly to TSS, FRiP proved to be informative to evaluate signal-to-noise ratios; however it relies on peak calling and thus more likely to be biased. In this line, enrichment of ATAC signal across Ensembl annotated TSS is now recommended by ENCODE as the preferred means of assessing overall sample quality, and was implemented as the metric to evaluate signal-to-noise in our pipeline.

The variability in quality of the ATAC-seq and Fast-ATAC libraries was also addressed at the peak calling level in this chapter, with the implementation of a peak filtering strategy that for each particular sample could identify good quality and reproducible peaks using IDR analysis between pseudoreplicates. This approach demonstrated to allowing to reduce repetitive and non meaningful regions that could be confounder for downstream analysis. In terms of sequencing depth, analysis in this chapter showed 20 to 25 million reads after filtering to be the required minimum sufficient to identify an appropriate proportion of accessible regions (peaks) as well as to obtain meaningful results regarding the peak filtering based on the pseudoreplicates IDR analysis. These observations have also been confirmed by Qu *et al.*,, where IDR analysis used to evaluate consistency across replicates but not implemented for peak filtering.

Establishment of appropriate measurements for post-sequencing library QC allowed to formally test the effect of transposition times, one of the most critical variables in ATAC that can be cell type specific, beyond the conditions from Buenrostro's publication. At the start of the project transposition for 40 min appeared as the most appropriate for all the cell types according to pre-sequencing library QC (tapestation). Assessment of three different transposition times in the ATAC-seq protocol showed heterogenous impact on the ratio of NFF/NBF and overall no major impact on signal-to noise ratios. The release of the improved Fast-ATAC protocol manifested some of the limitations identified by the ATAC-seq data generated within our group. The limitations of this first protocol. Fast-ATAC significantly reduced the percentage MT reads in all four cell types of interest for this project. In contrast, the improvement of signal-to-noise for hematopoietic cells claimed by the Fast-ATAC could not be was not evident. In fact, Corces *et al.*, only showed improved TSS by Fast-ATAC. The publication of Omni-ATAC where a comprehensive comparison of the three ATAC protocols was performed across a large number of cell types demonstrated that Fast-ATAC

did not improve TSS fol-enrichment when compared to ATAC-seq in some of the hematopoietic cells, for example $CD19^{+}$ cells, consistently with this data.

The challenges of performing differential chromatin accessibility analysis

Until ATAC-seq release, limited research had been performed to investigate differences in chromatin accessibility, and mainly using cell lines (**Degner2012**). The availability to study chromatin accessibility in clinical samples has required the definition of a consensus master list of accessible regions for which no accepted method has been agreed. In this project, a master list containing all the peaks identified in at least 30% of the samples included in analysis has been chosen. This represents an unbiased approach to include peaks that can vary across individuals (regardless group) but still be differentially accessible across conditions. Other publications have preferred building condition-specific master list or simply include all the significant peaks called in all the analysed samples (**Alasoo2018**; **Turner2018**).

When used for differential analysis, an additional filtering step to account for high read counts at peaks that were absent in some of the samples (background counts) has been implemented. In terms of the algorithm to perform normalisation and differential chromatin accessibility analysis, no consensus has been reached. The majority of the studies reviewed at the time of implementing differential analysis were peak-based and relied on RNA-seq or microarray algorithms such as EdgeR, limma or DESEq2 (Table 1.1). The analysis here performed, revealed DESeq2 as a more stringent method compared to quantile normalisation&limma voom. Limma has been reported to be affected by low quality samples and that may also explain the increase in differential hits observed when compared to DESEq2 (**Alasoo2018**). For both methods, the implementation of the additional filtering cut-off to control for high number of background reads has shown a reduction in the number of significant

differentially accessible regions. Given the difficulties of obtaining large number of high quality samples in a clinical setting, DEseq2 in combination with the additional filtering step to control to some extent for potential false positive appeared as an appropriately stringent method at the time of this project. As specific R packages for ATAC analysis are developed, further comparison of the outputs will be of interest in future work.

1.3.1 Studying the chromatin landscape from psoriasis biopsies

Up to date, only RNA-seq studies have been performed in KCs from psoriasis skin biopsies. The relevance of KCs in psoriasis pathophysiology and the ability to sample the tissue represented a great opportunity to investigate the chromatin accessibility landscape at the main site of inflammation using ATAC. Three different ATAC protocols (ATAC 1, ATAC2 and Fast-ATAC) performed very poorly in KCs isolated from skin biopsies and also in cultured NHEKs. The fact that similar results were obtained in KCs isolated through different systems as well as in NHEKS indicated the main reason for the lack of performance of these protocols be intrinsic to the cell type and not to be driven by compromising the viability upon the system used to isolate the cells. As they differentiate, KCs synthesis an insoluble protein structure that progressively replaces the plasma membrane, which may have been impairing appropriate cell permeabilisation and efficient transposition reaction. Interestingly, Bao's protocol, using increased Tn5 concentration to perform appropriately in NHEKS did not appear to improve ATAC quality libraries in my data. Similarly, the additional optimisation of the Fast-ATAC protocol modifying the concentration of detergent and Tn5 also failed to improve the quality of the data.

Release of the Omni-ATAC protocol and comparison of its performance along with ATAC-seq and Fast-ATAC reproduced the failure to generate good quality data of the former two protocols in KCs. Corces *et al.*, highlighted the

issues in generating good quality data with low signal-to-noise ratios in this cell type. The release of Omni-ATAC proving excellent performance in KCs has opened a real avenue to explore the chromatin landscape in lesional and uninvolved psoriasis biopsies for the continuation of this project.

1.3.2 Characterisation of the effect of preservative techniques in the chromatin landscape

The use of clinical samples sometimes involves logistical limitations that require of sample preservation. At the time of starting this thesis Oxford Genomic Center at the WCHG has implemented the use of DSP as a compatible fixative with microfluidics-based scRNA-seq methods, being of interest to test the ability of this fixative to perform well in ATAC (**Attar2018**). Alternatively cryopreservation of PBMCs had historically been used but formal assessment of the effect of this process in the chromatin landscape of the different cell types had not yet been conducted.

DSP fixed samples presented overall lower abundance of NFF, when compared to the fresh and frozen libraries. Interestingly, DSP performed very poorly in the two cell types from CTL1, which presented extremely low abundance of NFF and predominance of NBF. Interestingly, despite the abundance of di-nucleosome fragments in these two samples, the chromatin structure across the TSS failed to reproduce the position of the TSS flanking nucleosome, which could be due to nucleosome displacement, as DSP does not cross-link DNA to proteins. Since this effect was only observed for CTL1, problems inappropriate performance of the fixation protocol on that particular day may be the cause.

After removing the CTL1 samples, consideration of the chromatin accessibility landscape in all the remaining samples clearly showed that the differences by condition were greater that the differences between individuals.

This was confirmed by perming differential analysis, which revealed a moderate number of DARs when comparing fixed or frozen to the fresh reference ATAC-seq. As expected, $CD14^{+}$monocytes appeared to be more sensitive to cryopreservation than $tCD4^{+}$ cells and present greater differences in the chromatin accessibility landscape when compared to fresh samples.

The results here presented could be limited by the borderline quality of some of the libraries, as previously explained, which was intrinsic to the issues of consistency in performance of the ATAC-seq protocol and could be distorting the findings to some extent. Nevertheless, this study still provides some useful information regarding the effect of using DSP in sorted cell populations or cryopreserved PBMCs. Depending of the sample size foreseen in each particular study, the aim to perform paired ATAC and scRNA-seq based on microfluidics methods in the same samples and the final biological question, these two preservative protocols could be considered for implementation. Additionally, as the ATAC-seq protocol has been improved, the ability to perform ATAC in frozen tissues and formaldehyde samples has also been successfully conducted and can be considered when establishing the experimental design (**Corces2017**; **Chen2016**). In this thesis, since the sample size was limited and the main question was to assess the chromatin landscape as close as possible to *in vivo* disease conditions, fresh cells were used to generate the results presented in the following two chapters.

1.3.3 Conclusions

As ATAC becomes a more commonly used technique, new methods and updates are introduced both in the lab and analysis side. The aim of this chapter was to establish methods that could be used at the time (with the available resources and expertise in the group) in a clinical setting. A comprehensive comparison of all possible methods recently published was not therefore

performed. The data presented throughout this thesis provides a resource for representative study of the chromatin accessibility landscape using patients's samples that could be revisited in the future with further optimised analytical methods. The work described here has compared commonly used strategies from early ATAC publications in order to establish an appropriate method to perform chromatin accessibility analysis in the context of psoriasis and PsA maximising the use of available samples whilst accounting for some quality limitations. The implementation of Omni-ATAC in future sample recruitment will further improve sample quality and confidence in reported differentially open regions.

[appendix]lof [appendix]lot atoc0

Appendices

[appendix]lof List of Figures

[appendix]lot List of Tables

Appendix A

Appendices

A.1 Additional tables

A.1.1 Chapter 5 Tables

Cell type	Condition	TSS enrichment CTL1	CTL2	CTL3
CD14	Fresh	17.4	19.6	14.11
	Frozen	26.3	25.2	27.1
	Fixed	2.5	16.5	22.4
CD4	Fresh	5.3	5.6	7.7
	Frozen	17.9	14.1	16.1
	Fixed	7.9	23.0	14.3

Table A.1: Enrichment of ATAC-seq reads across the TSS for the $CD14^+$ monocytes and $CD4^+$ samples fresh, frozen and fixed.

Sample ID	NRF	PBC1/PBC2
PS2000 CD14	77.6	0.60/2.5
PS2001 CD14	84.9	0.70/3.0
PS2314 CD14	81.1	0.60/1.8
PS2319 CD14	79.9	0.60/2.2
CTL7 CD14	81.1	0.65/2.2
CTL8 CD14	83.9	0.66/2.3
CTL9 CD14	80.7	0.60/2.3
CTL10 CD14	83.1	0.65/2.1
PS2000 CD4	84.8	0.75/3.4
PS2001 CD4	82.0	0.72/2.9
PS2314 CD4	82.9	0.71/2.8
PS2319 CD4	82.4	0.73/3.2
CTL7 CD4	78.6	0.68/2.5
CTL8 CD4	81.8	0.71/2.9
CTL9 CD4	81.6	0.74/3.3
CTL10 CD4	77.6	0.61/1.9
PS2000 CD8	77.0	0.76/4.5
PS2001 CD8	74.7	0.74/4.0
PS2314 CD8	74.2	0.75/4.1
PS2319 CD8	72.2	0.75/4.0
CTL7 CD8	32.7	0.32/1.5
CTL8 CD8	70.1	0.70/3.3
CTL9 CD8	73.9	0.73/3.7
CTL10 CD8	68.2	0.65/2.9
PS2000 CD19	38.0	0.42/1.9
PS2001 CD19	71.4	0.71/3.7
PS2314 CD19	29.4	0.34/1.8
PS2319 CD19	76.1	0.78/4.8
CTL7 CD19	74.2	0.69/3.1
CTL8 CD19	68.4	0.67/3.2
CTL9 CD19	75.1	0.76/4.6
CTL10 CD19	61.7	0.59/2.6

Table A.2: Evaluation of ChiPm library complexity for the psoriasis and control chort 1B ChIPm assay. NRF, PBC1 and PBC2 are the three measures used according to the ENCODE standards as referred in Chapter **??**. 0.5≤NRF<0.8 acceptable; 0.8≤NRF≤0.9 compliant; NRF>0.9 ideal; 0.5≤PBC1<0.8 and 1≤PBC2<3 moderate bottlenecking; 0.8≤PBC1<0.9 and 3≤PBC2<10 mild bottlenecking.

CD14$^+$ monocytes additional enriched pathways in psoriasis
Generic transcription
RNA transport
GnRH signalling
Ribosome biogenesis in eukaryotes
Neurotrophin signaling
Spliceosome
Autophagy
Protein processing in endoplasmic reticulum
CD8+ additional enriched pathways in psoriasis
Epstein-Barr virus infection
RNA Polymerase I and III, and mitochondrial transcription
Apoptosis

Table A.3: Additional enriched pathways DEGs between psoriasis and healthy controls in CD14$^+$ monocytes and CD8$^+$ cells. Significant pathways for FDR<0.01. All the enriched pathways contained a minimum of ten DEGs FDR<0.05 from the analysis.

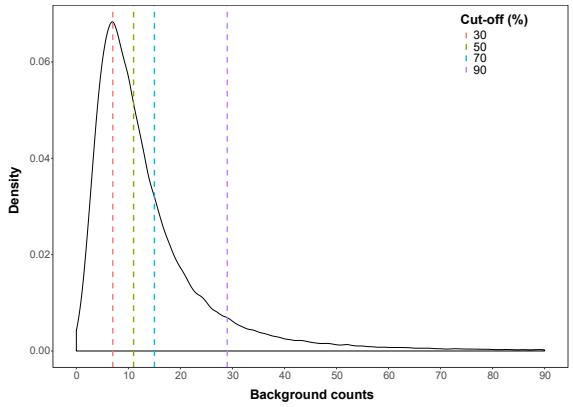


Figure A.1: Distribution of the background read counts from all the master list peaks absent peaks in each sample. Each cut-off corresponds to the number of background counts showed by a particular percentage of the total number of absent peaks.

Lesional versus uninvolved epidermis additional enriched pathways
Genes encoding extracellular matrix and extracellular matrix-associated proteins
Serine/threonine-protein kinase (PLK1) signalling
Genes encoding secreted soluble factors
Glycolysis/gluconeogenesis
FOXM1 transcription factor network
Phase 1 functionalization of compounds
Biological oxidations
G2/M Checkpoints
Biological oxidations
Aurora B signaling
Chemical carcinogenesis
Serotonergic synapse
Drug metabolism-cytochrome P450
Mitotic M-M/G1 phases
DNA Replication
MicroRNAs in cancer
Metabolism of amino acids and derivatives
Metabolism of carbohydrates
Glycosaminoglycan metabolism
E2F transcription factor network
p73 transcription factor network
Genes encoding structural ECM glycoproteins
Transmembrane transport of small molecules
Fc-epsilon receptor I signaling in mast cells
Tight junction
Origin recognition complex subunit 1 (Orc1) removal from chromatin

Table A.4: Additional enriched pathways for DEGs between lesional and uninvolved epidermis isolated from psoriasis patients skin biopsies. Significant pathways for FDR<0.005. All the enriched pathways contained a minimum of ten DEGs FDR>0.05 from the analysis.

CC-mixed CD14+ monocytes additional enriched pathways
SLE
Translation
3'-UTR-mediated translational regulation
Th-1 and Th-2 cell differentiation
Peptide chain elongation
Rheumatoid arthritis
Metabolism of proteins
Cell adhesion molecules (CAMs)
Th-17 cell differentiation
Nonsense mediated decay enhanced by the exon junction complex
SRP-dependent co-translational protein targeting to membrane
Hemostasis
Metabolism of mRNA
Platelet activation, signalling and aggregation
HTLV-I infection
Innate immune system
Adaptive immune system

CC-IL7R CD14+ monocytes additional enriched pathways
SLE
Tuberculosis
Epstein-Barr virus infection
Immune System

Table A.5: Additional enriched pathways for the DEGs between SF and PB CD14$^+$ monocytes from the CC-mixed and CC-IL7R subpopulations. All the enriched pathways contained a minimum of ten DEGs from the analysis and were significant at an FDR<0.01.

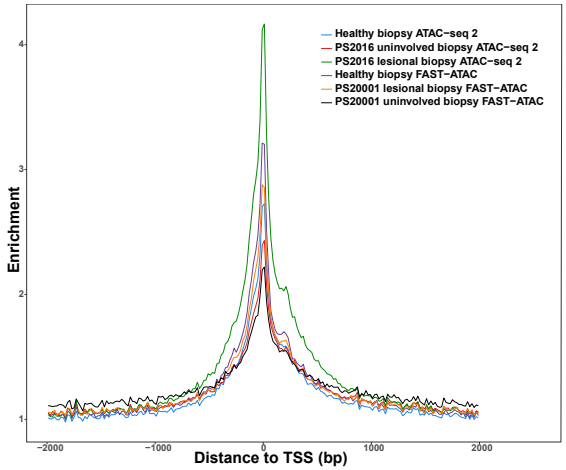
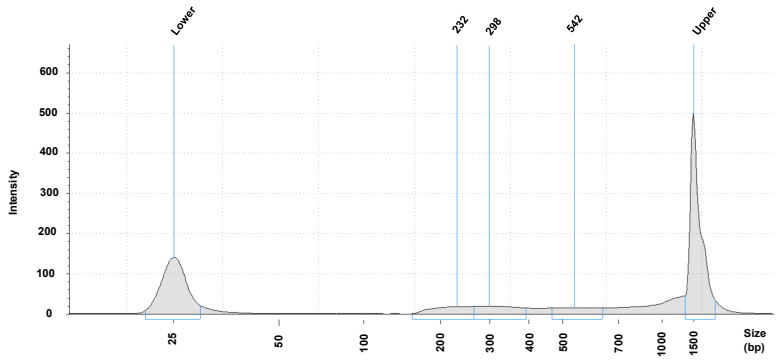
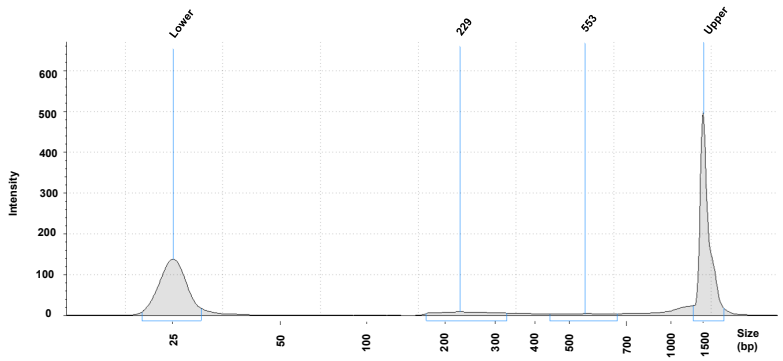


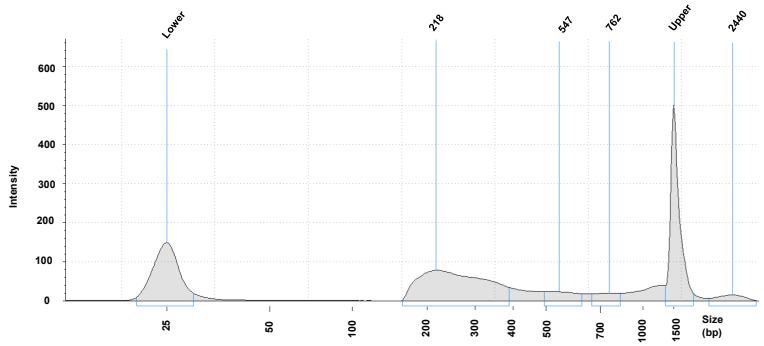
Figure A.2: Assessment of TSS enrichment from ATAC 1 and Fast-ATAC in healthy and psoriasis KCs isolated from skin biopsy samples. Fold-enrichment of ATAC fragments across the Ensembl annotated TSS from the different ATAC libraries.



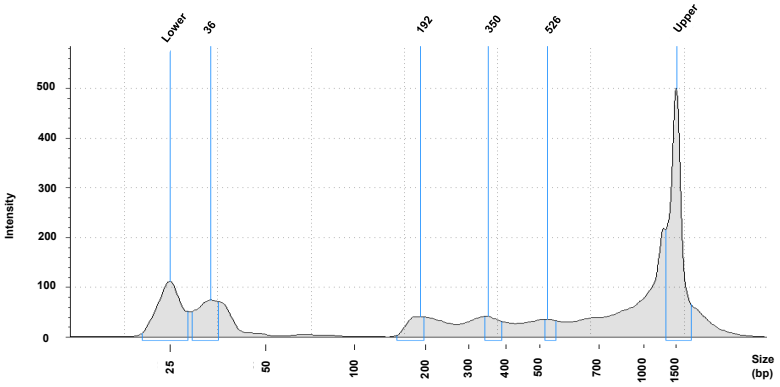
(a)



(b)



(c)



(d)

Figure A.3: Fast-ATAC and Omni-ATAC NHEK tapestation profiles. Pre-sequencing ~~quantification of DNA fragment sizes from the libraries generated using the a) C2, b) C3,~~ and c) C4 versions of the Fast-ATAC protocol based on modifications in the detergent and Tn5 concentration and d) Omni-ATAC. C2, C3 and C4 detergent and Tn5 concentrations are detailed in Table 1.3.

A.1.2 Chapter 4 Tables

A.1.3 Chapter 5 Tables

A.2 Additional figures

A.2.1 Chapter 3 Figures

A.2.2 Chapter 4 Figures

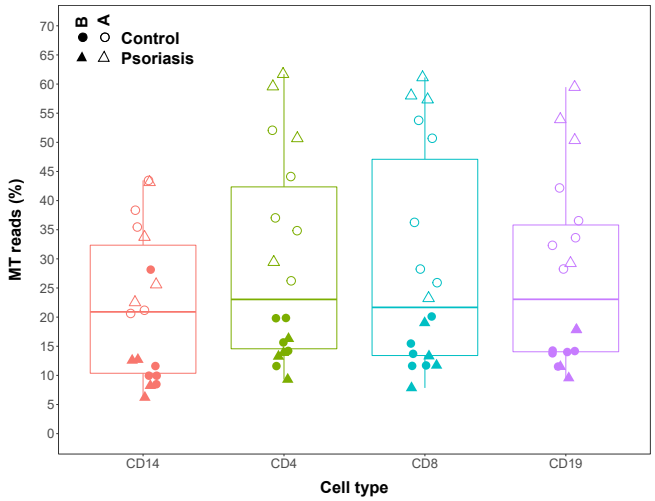


Figure A.4: Percentage of MT reads in the ATAC-seq samples generated in CD14$^+$ monocytes, CD4$^+$, CD8$^+$ and CD19$^+$ isolated from psoriasis patients and healthy controls. Samples from cohort 1A (open circlues and triangles) were generated with the standard ATAC-seq protocol from Buenrostro *et al.*, 2013 whereas samples from cohort 1B (filles circles and triangles) were processed using FAST-ATAC (**Corces2016**).

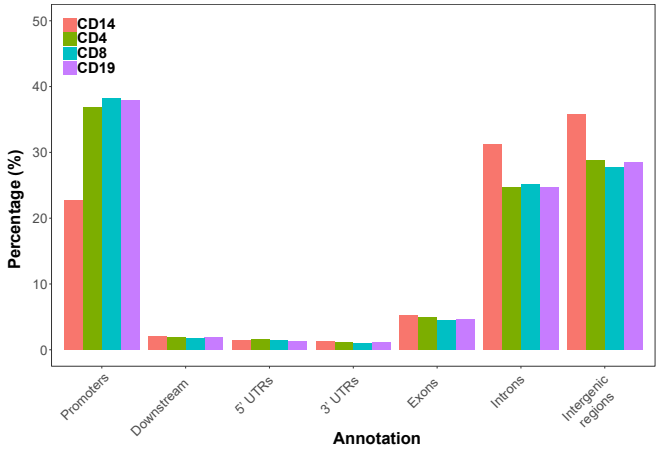
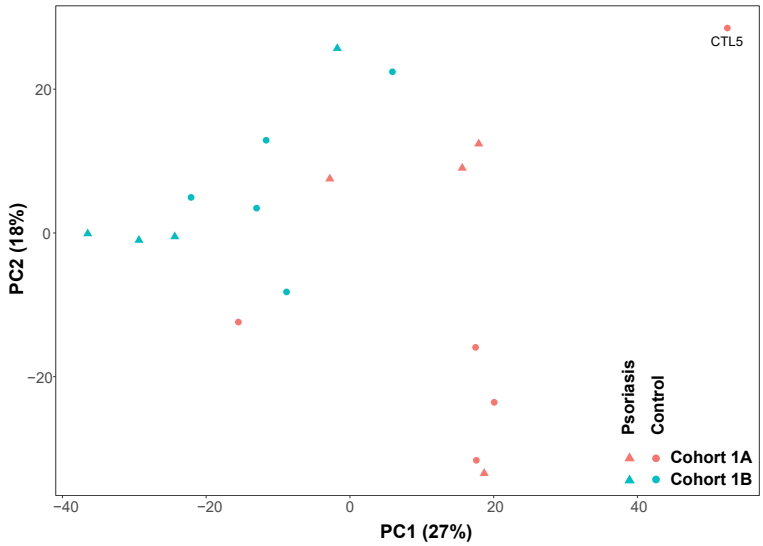
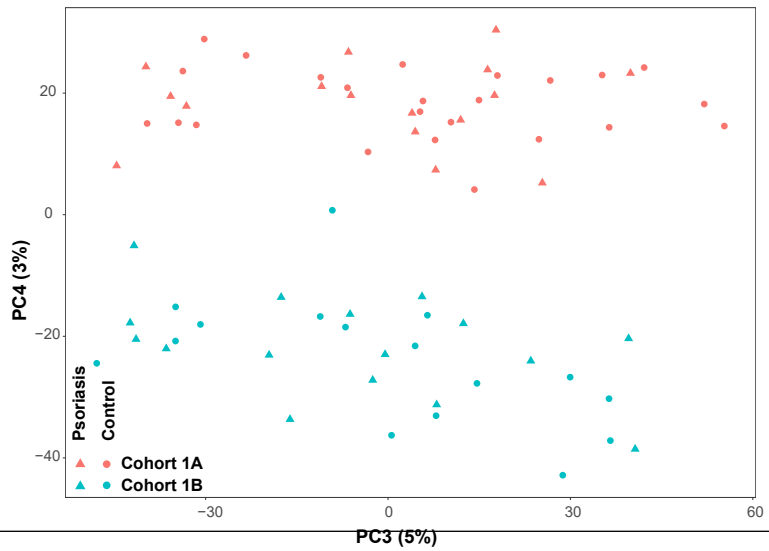


Figure A.5: Genomic annotation of the consensus master list of ATAC-seq enriched sites built for downstream differential chromatin accessibility analysis in CD14$^+$ monocytes, CD4$^+$, CD8$^+$ and CD19$^+$. Annotation is expressed in percentage over the total number of ATAC-seq sites included in each particular cell type master list.



(a)



(b)

A.2.3 Chapter 5 Figures

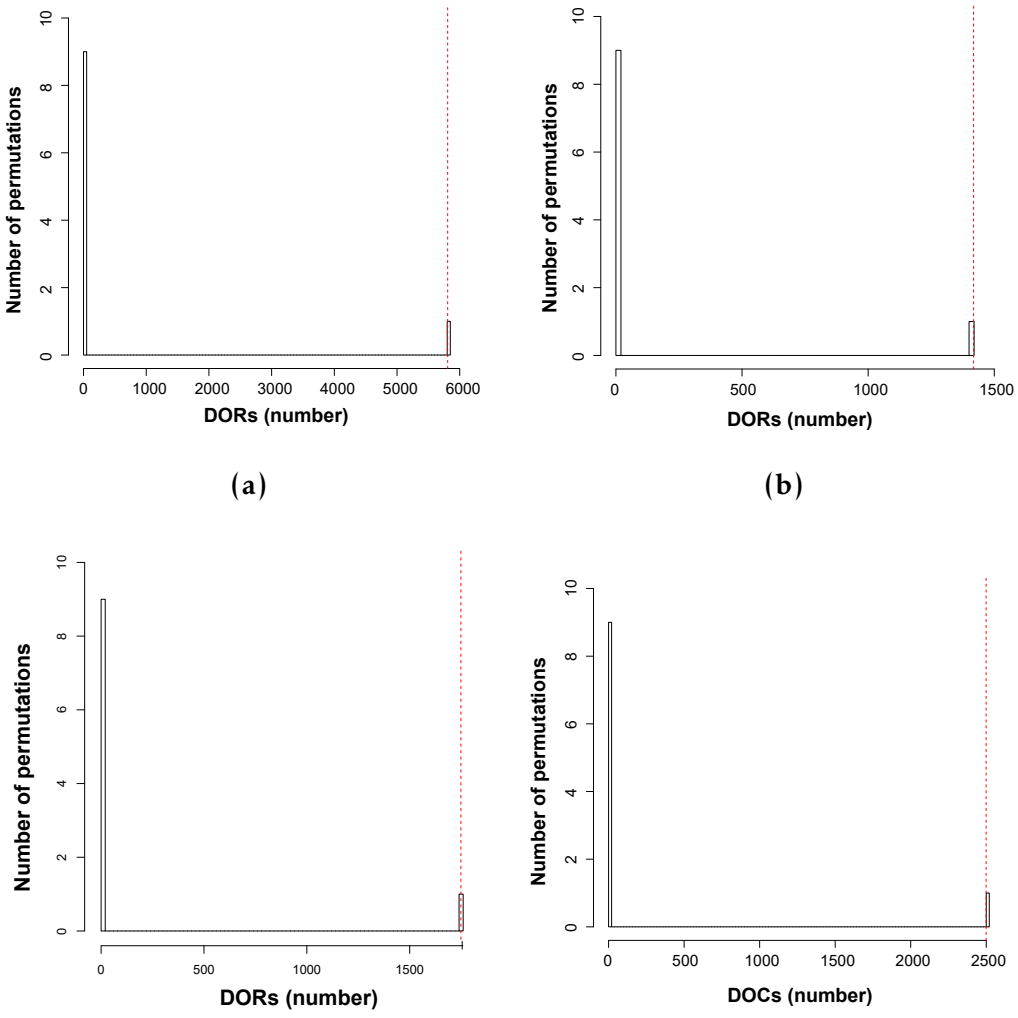


Figure A.7: Permutation analysis SF vs PB in $CD14^+$,$CD4m^+$,$CD8m^+$ and NK.

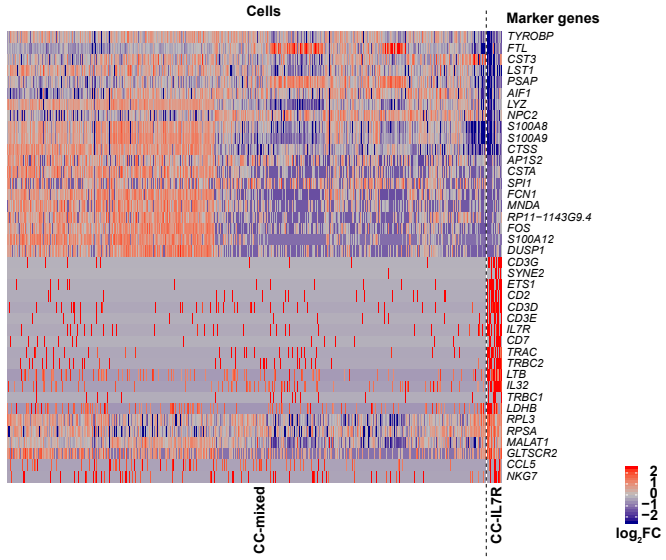
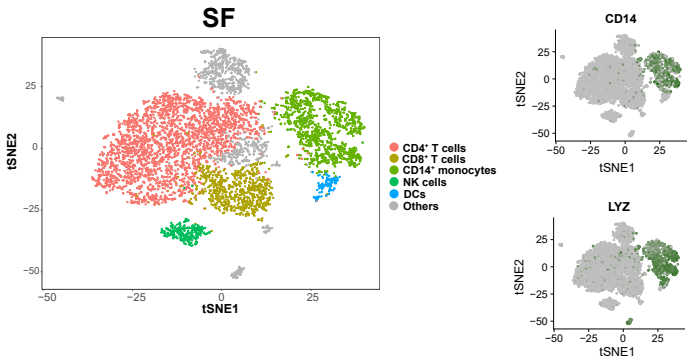
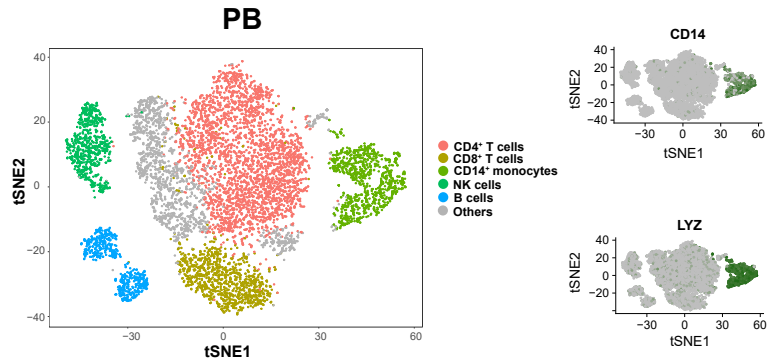


Figure A.8: Heatmap for the top 20 marker genes of the CC-mixed and CC-IL7R CD14$^+$ monocytes subpopulations. Rows are the top 20 marker genes for each of the two subpopulations (total of 40 genes). The columns represent each of the cells members of the CC-mixed (left) or CC-IL7R (right) clusters. The colour scale represents the $\log_2$FC in the expression of the marker gene in a particular cell of the cluster compared to the average expression of all the cells from the other cluster.



(a)



(b)

Figure A.9: Identification of the CD14$^+$ monocytes populations from bulk SFMCs and PBMCs using scRNA-seq transcriptomes. xxxx