Chapter 3

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

# Introduction

## 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 1. All these techniques require a large number of cells as input material, making them unsuitable for use in a clinical setting. The 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 provided a fast two-step 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 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 and results turn-over time.

## ATAC-seq limitations and advances in optimisation

Despite the advantages in terms of reduced input material and processing time, early data from ATAC-seq revealed two major limitations, namely, a high percentage of mitochrondrial DNA tagged by the Tn5 enzyme, and insufficient sensitivity to detect all the accessible regions, in part due to high background noise (Corces et al. 2016; Sos et al. 2016). An optimised version of the protocol specifically for hematopoietic cells, named Fast-ATAC, replaced the NP-40 detergent used in ATAC-seq with digitonin. This prevents solubilisation of the mitochondrial membrane, and performs lysis and transposition in a single step. This modification efficiently reduced the percentage of mitochondrial reads down to approximately 10% and increased ATAC signal at annotated TSS, using only 5,000 cells as input material (Corces et al. 2016). Optimisation of the ATAC- seq protocol for keratinocytes 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 (Bao et al. 2015).

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 (Corces et al. 2017). The Omni- ATAC protocol consisted of lysis, 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 a higher variability in sample quality and sensitivity in the latter two protocols. Moreover, greater signal-to-noise ratios were also achieved with Omni-ATAC by modifying the transposition buffer. Notably, this versatile protocol represented a particular improvement in keratinocytes with data demonstrating the inability of standard ATAC- seq and Fast-ATAC to yield good quality data in this cell type.

## Challenges of ATAC data analysis

Although some guidelines for DHS data analysis were available, the release of the new ATAC methods also led to a 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 10x106 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 measurements in order to confidently identify good quality samples prior to downstream analysis. Regarding peak calling in ATAC, different algorithms have been applied, with MACS2 being preferred by the majority of studies, including ENCODE. Determining criteria for filtering out poor quality peaks in ATAC is 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

criterion except for ENCODE data, where technical replicates are generated and IDR analysis has been used to identify robust peaks.

The feasibility of generating ATAC data from low numbers of cells and clinical samples presents 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 performing normalisation across samples before conducting differential analysis with microarray or RNA-seq based-methods. An alternative is known as the *de novo* approach, used for ChIP data, which consists of using a sliding window to scan the genome and identify those regions showing read count differences between two groups of samples, avoiding peak calling bias (Shen et al. 2013).

## The challenge of working with clinical samples

The opportunity to apply epigenetic assays to clinical samples has also highlighted a logistic problem regarding the 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 provide a snap shot of the *in vivo* cellular characteristics and avoid confounders of the biological significance of the results generated. The main methods to preserve cell structure and DNA integrity involve 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 nuclei and chromatin organization, and overall yielded comparable ATAC-seq data to those generated in fresh neurons (Milani et al. 2016). 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 Centre at the WHG 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 (Attar et al. 2018). 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 as FACS cell separation (Espina2013).

## Aims

The aim of this chapter is to establish the ATAC protocol(s) in the laboratory, perform a thorough optimisation of the methodological and analytical tools required to study chromatin accessibility in clinical samples of interest, and to determine the suitability of relevant methods for sample preservation to overcome the inherent logistical constraints of working with clinical samples.

The specific aims of this chapter are:

* + - 1. To establish an in-house pipeline to analyse ATAC data, including quality control measurements, peak filtering and a method for differential analysis in order to maximise the use of available samples.
      2. To investigate the effect of transposition time on sample quality in the ATAC-seq protocol.
      3. To validate the reduction of mitochondrial DNA and improvement of signal in Fast-ATAC protocol compared to ATAC-seq.
      4. To test and optimise ATAC-seq protocols in order to adapt them to the mapping of the chromatin accessibility landscape in psoriasis skin biopsies and in cultured keratinocytes.
      5. To determine the effect of cryopreservation and DSP fixation on ATAC-seq data quality and on the overall chromatin accessibility landscape of CD14+ monocytes and total CD4+ (CD4+) cells.

# Results

## Establishment of an ATAC-seq data analysis pipeline

A robust ATAC-seq data analysis pipeline was required as this was a new methodology (Buenrostro2013) for which established pipelines were not available at the time this work was undertaken. Such a pipeline requires appropriate quality control, peak calling and filtering, and a method for identification of differential chromatin accessibility between groups of samples.

Data used for pipeline optimization

The most appropriate criteria and parameters to implement were investigated using ATAC-seq data generated for paired CD14+ monocytes and CD4+ T cells from three healthy individuals (CTL1-3).

Sample QC

The variability in performance of ATAC-seq experiments and protocols requires the identification of the most informative QC measures and parameters to determine the quality of the samples before proceeding with downstream differential analysis.

ATAC-seq fragment size distribution was analysed for each library following down-sampling to 30 million reads after filtering to facilitate the comparison. The observed fragment size distribution demonstrated nucleosome periodicity protecting the DNA during the transposition event (Figure 3.1 a), indicating chromatin integrity and thus good quality ATAC-seq libraries. All 6 libraries showed appropriate nucleosome periodicity (every ∼200bp) up

to 600bp, clearly distinguishing chromatin organisation into mono-, di- and tri-

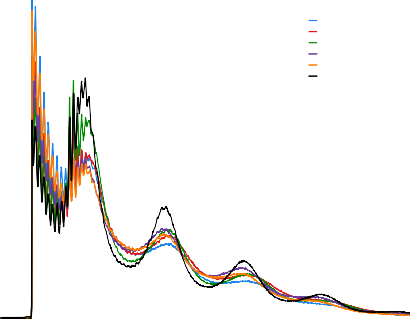
nucleosomes. Some variation in the relative intensity of nucleosome-free fragments (NFF) (≤147pb, approximately) compared to nucleosome-bound DNA was seen across samples however NFF were clearly distinguished in all of the samples, which is considered a compulsory feature for ATAC-seq libraries to pass QC, according to 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 Ensembl genes (Figure 3.1b). This is a useful measure as

nucleosome repositioning and an increase in chromatin accessibility occur at the

TSS to allow TF binding and initiation of gene transcription.

A

6

**CTL1 CD4 CTL2 CD4 CTL3 CD4 CTL1 CD14 CTL2 CD14 CTL3 CD14**

4

**Normalised insert size density (x10−3)**

2

0

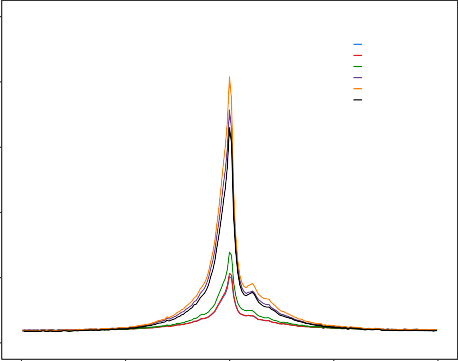
     

0 200 400 600 800 1000

**Fragment length (bp)**

B

25

20

**CTL1 CD4 CTL2 CD4 CTL3 CD4 CTL1 CD14 CTL2 CD14 CTL3 CD14**

15

**Enrichment**

10

5

0

−2000 −1000 0 1000 2000

**Distance to TSS (bp)**

C

chr12:

10 kb hg19

6,640,000  6,645,000  6,650,000 6,655,000 6,660,000 6,665,000  6,670,000 

NCAPD2

**CTL1 CD4 CTL2 CD4 CTL3 CD4 CTL1 CD14 CTL2 CD14 CTL3 CD14**

**GAPDH**

IFFO1

NOP2 



**Figure 3.1: Measurements for QC assessment in ATAC-seq samples.** For each of the CD14+ monocytes and CD4+ samples used to establish the ATAC analysis pipeline, QC measures shown are (A) density distribution of ATAC-seq fragment sizes, and (B) enrichment of ATAC- seq fragments across the TSS of all Ensembl genes. (C) UCSC Genome Browser view illustrating the ATAC-seq normalised read density (y-axis) at the promoters of *GAPDH* and *NOP2* genes. CD14+ monocytes and CD4+ tracks are colour-coded in green and red, respectively.

Fold-enrichment signals over the TSS ranged from 5 to 7 for the CD4+ samples, and were much higher (17 to 20) in the CD14+ samples. The lower sample quality of the CD4+ compared to CD14+ samples indicated by the TSS enrichment values is further evidenced by visualising the ATAC-seq read pile up at the promoters of the glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) and the

NOP2 Nucleolar Protein (*NOP2*) gene which shows more background reads and lower signal for the CD4+ samples in red (Figure 3.1c).

As part of the QC assessment, the percentage of mitochondrial reads and the fraction of reads in peaks (FRiP) were also investigated (Table 3.2).

|  |  |  |
| --- | --- | --- |
| Sample | % mitochondrial 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 3.2: ATAC-seq percentage of mitochondrial reads and fraction of reads in called peaks (FRiP). The percentage of mitochondrial reads was calculated over the total number of sequencing reads (before filtering). FRiP was calculated as the proportion of ATAC-seq fragments overlapping significant peaks with standard filtering for all the samples (FDR*<*0.01).

FRiP score is an alternative to TSS enrichment for 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), suggesting both are appropriate inter-dependent QC measures to evaluate sample noise. For the 6 samples analysed here the mitochondrial content varied (14.9- 43.3%), was higher in CD14+ than in CD4+ cells and was not directly related to any of the other QC measurements. Therefore, mitochondrial reads in this range did not appear to reflect the samples quality and the main issue was the need for deeper sequencing to achieve the desired number of non-mitochondrial reads for downstream analysis.

Both TSS and FRiP are appropriate signal-to-noise measures, with recommended threshold values from ENCODE and Alsoo *et al.*, 2018 being 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 therefore this was the chosen measure for this thesis. In summary from this analysis, all six samples showed appropriate ATAC-seq patterns of fragment size distribution, FRiP and TSS, with exception of CD4+ CTL1 and CTL2, being borderline for the 6 fold-enrichment TSS threshold.

Peak calling and filtering

Next, criteria for peak calling and filtering were investigated. Although different peak callers have been used to analyse ATAC-seq data, MACS2 has been methodology preferred by ENCODE and most publications (Table 3.1). MACS2 was initially 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) parameters, 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, this 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 (Wang et al. 2018).

A systematic analysis of the effect of sequencing depth and the sample quality on peak calling was conducted to better understand the effect of both

variables on the downstream analysis. For each of the six samples, random sub- sampling of reads was performed in 5 million increments, ranging from 5 to 30 total million reads, followed by peak calling with arbitrary filtering for false discovery rate (FDR)*<*0.01. The number of called peaks passing filtering showed a steady increase with read depth (Figure 3.2 a), beginning to plateauat approximately 25 million reads (Figure 3.2 b). Moreover, lower number of peaks were detected

in CD4+ 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 as measured by FRiP (which relies on peak calling) showed very discrete changes with read depth and was stable from 15 million reads onwards for all six samples (Figure 3.2 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.

For peak filtering, an arbitrary FDR*<*0.01 in MACS2 is typically used but may not remove low quality peaks equally successfully in lower quality samples and does not take into account reproducibility of the called peaks. To investigate this, IDR was used to experimentally identify the most appropriate p-value threshold to filter the called peaks in each individual sample. Filtered reads from each sample were partitioned in half to create two pseudoreplicates, peaks were called in each pseudoreplicate and the percentage of peaks sharing IDR rank position when filtered at decreasing pvals was calculated (Figure 3.3 a and b). This strategy was tested across a range of total read counts (as above) to determine the effect of sequencing depth on the suitability of this peak calling filtering approach. The optimal pval giving the largest percentage of IDR shared peaks between the two pseudoreplicates varied more erratically

**CTL1 CD4 CTL2 CD4**

**CTL3 CD4 CTL1 CD14**

**CTL2 CD14 CTL3 CD14**

90000

75000

60000

**Peaks (number)**

45000

30000

15000

0

5 10 15 20 25 30

**Sequencing depth (million reads)**

(a)

50000 80

**CTL1 CD4 CTL3 CD4 CTL2 CD14 CTL2 CD4 CTL1 CD14 CTL3 CD14**

**CTL1 CD4 CTL3 CD4 CTL2 CD14 CTL2 CD4 CTL1 CD14 CTL3 CD14**

70

40000

**Increment in peak number**

60

30000 50

**FRiP (%)**

40

20000

30

10000 20

10

0

0 5 10 15 20 25 30

**Sequencing depth (million reads)**

(b)

0

5 10 15 20 25 30

**Sequencing depth (million reads)**

(c)

Figure 3.2: FRiP and peak calling at different sequencing depths in ATAC-seq libraries. For a series of sequencing depths (from 5 to 30 million 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.

when the sequencing depth was lower than 10 million reads (Figure 3.3 a and b), suggesting this analysis was not appropriate for lower read depths. This variation was more pronounced and extended in CD4+ T cells, which had lower TSS values compared to CD14+ monocytes ATAC-seq libraries (representative examples of CTL2 CD4+ and CD14+ monocytes in Figure 3.3 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 ≥10) samples presented a profile reaching a single maximum of shared IDR peaks for a particular filtering pval (Figure 3.3 b), which was -log10 pval 8 in all three samples (data not shown). In contrast, the same analysis in CD4+ samplesrevealed two pvals for which the percentage of IDR shared peaks reached two local maxima (Figure 3.3 a).

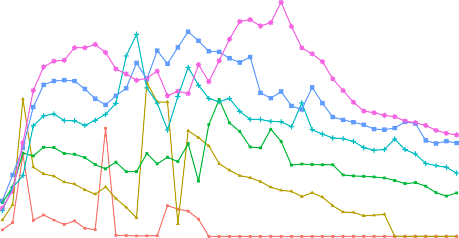
Filtering the CD4+ peaks at the pval of the first local maximum (10−11)

reduced the percentage of peaks overlapping regions, suggestive of noise (e.g heterochromatin, repetitive sequences and repressed regions) when compared to using the list of significant peaks filtered based on FDR*<*0.01 (Figure 3.3 c). In summary, this IDR analysis provided a systematic method to identify an optimum pval with which to perform sample-specific filtering of technically reproducible peaks when the sequencing depth was over 10 million reads. The resulting 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 thesis, 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 section x.x) is the

**Depth (million reads)**



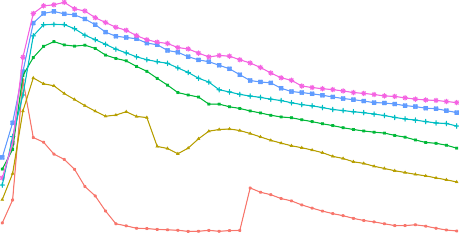
35

**5 15 25**

**10 20 30**

**Depth (million reads)**

60



**5 15 25**

**10 20 30**

30 50

25

**Shared peaks (%)**

**Shared peaks (%)**

40

20

30

15

20

10

10

5

0

0 5 10 15 20 25 30 35 40 45

**−log10(pval)**

(a)

0

0 5 10 15 20 25 30 35 40 45

**−log10(pval)**

(b)

100

**IDR**

**FDR**

**30.01**

**27.88**

**25.59**

**46.74**

**42.96**

**70.80**

**24.03**

**22.89**

**22.73**

**30.54**

**28.40**

**16.68 20.33 21.84**

**20.63**

**19.72**

**19.69**

**23.72 22.72 21.89**

75

**Percentage (%)**

50

25

**Chromatin states**

**1\_Active\_Promoter 2\_Weak\_Promoter 3\_Poised\_Promoter 4\_Strong\_Enhancer 5\_Strong\_Enhancer 6\_Weak\_Enhancer 7\_Weak\_Enhancer 8\_Insulator 9\_Txn\_Transition 10\_Txn\_Elongation 11\_Weak\_Txn 12\_Repressed 13\_Heterochrom/lo 14\_Repetitive/CNV 15\_Repetitive/CNV**

0

CTL1 CTL2 CTL3 CTL1 CTL2 CTL3

**Sample**

(c)

Figure 3.3: Peak calling filtering using IDR analysis in ATAC-seq samples. For each of the sequencing depths tested (from 5 to 30 million reads after filtering), an illustration of the percentage of peaks sharing IDR rank between the two pseuroreplicates is shown when using different pval filtering thresholds in CTL2. (a) CD14+ monocytes and (b) CD4+ (used as representative samples differing in quality for this analysis). (c) Annotation (as percentage of total) of the CTL1, CTL2 and CTL3 CD4+ 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.

background signal. Therefore, in collaboration with Dr Hai Fang, an empirical cut-off was identified to minimise the impact of background read counts on the peaks included for the differential analysis (Xinmin et al. 2005; Jonker et al. 2014). Two strategies were tested.

***Step 1***

**ML\_all**

**presence-absence matrix**

Sample

Peak

Peak

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **0** | **1** | **1** | **0** | **1** | **1** |
| **1** | **0** | **0** | **1** | **1** | **0** |
| **1** | **1** | **1** | **0** | **0** | **1** |
| **0** | **1** | **0** | **1** | **1** | **1** |
| **1** | **1** | **0** | **1** | **0** | **0** |

**ML\_all**

**counts matrix**

Sample

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **20** | **56** | **85** | **99** | **10** | **44** |
| **85** | **23** | **2** | **12** | **8** | **77** |
| **0** | **47** | **93** | **25** | **14** | **11** |
| **50** | **3** | **0** | **1** | **25** | **6** |
| **12** | **18** | **26** | **41** | **19** | **27** |

**Cut-off reduced**

**count matrix**

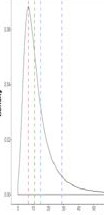
***Step 4***

**Background counts**

**Density**

.

. **Normalisation and**



**Cut-offs**

***Step 2***

. **differential analysis**

***Step 3***

**DESeq2**

. **Quantile**

. **&**

**limma voom**

Establishment of methods to assess genome-wide chromatin accessibility

95

Figure 3.4: 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 2, 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 matrices (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 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 series with 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 ML all includes peaks present in at least 30% of the total samples)(Figure 3.4 Step 1) were used to generate a density distribution plot (Figure 3.4 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 B.1). Each cut-off was used to filter out peaks from the ML all raw count matrix whose values in more than three samples were lower than the background counts (Figure B.1 Step 3). A filter of 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 was normalised using quantile or DESeq2 (library size and variance stabilisation (Love et al. 2014)) and 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 normalisation counts& limma voom showed a greater number of significant (FDR*<*0.01 and absolute FC (abs(FC))*>*1.5) differentially accessible regions (DARs) compared to DESeq2, across all filtering cut-offs (Figure ?? c). The two approaches presented a progressive decrease in the number of DARs from the 75% cut-off onward, suggesting a 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.

From this analysis, 80% was chosen as a conservative filtering cut-off, with the vast majority of the 19,855 DARs called as significant using the

**Quantile normalised counts**

12 12



**DESeq2 normalised counts**

10 10

8 8

**log2(quantile counts)**

**log2(DESeq2 counts)**

6 6

4 4

2 2

0 0

**CTL1 CD4**

**CTL2 CD4**

**CTL3 CD4**

**CTL1 CD14**

**CTL2 CD14**

**CTL3 CD14**

**CTL1 CD4**

**CTL2 CD4**

**CTL3 CD4**

(a)

**CTL1 CD14**

**CTL2 CD14**

**CTL3 CD14**

(b)

35000

**Method**

**quantile DESeq2**

30000

25000

20000

**DAR (number)**

15000

10000

5000

0

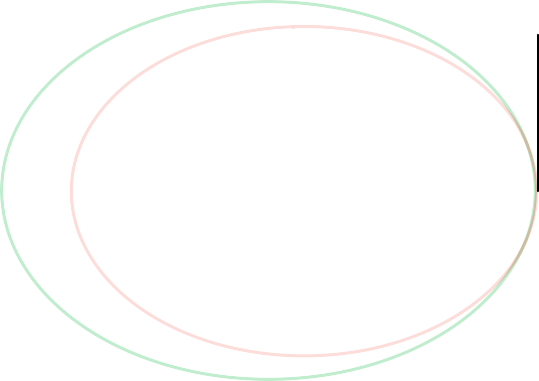
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95

**cut-off (%)**

(c)

Figure 3.5: Normalisation and differential chromatin accessibility analysis for different cut-offs using quantile normalisation&limma voom and DESeq2. Boxplots representing the log2 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) 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.

more conservative method (DESeq2) recapitulated by limma voom at the same significance threshold (FDR*<*0.01 and abs(FC)*>*1.5) (Figure 3.6 a). FDR rank revealed that out of the first 19,855 limma voom DARs, 18,768 were the same as those 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 significant DARs reported by limma voom and DESeq2 could partly be driven by differences in the models used by the two methods to estimate dispersion of counts.



**Quantile & limma voom**

**17**

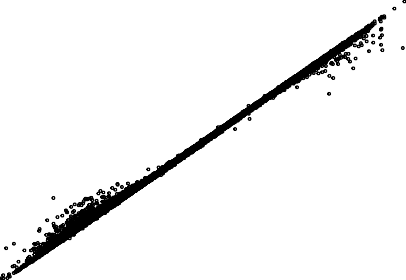
**6,257**

**19,838**

**DESeq2**

(a)

8.5



**R=0.999**

**p-val=2.2x10-16**

7.0

5.5

4.0

**log2( DESeq2 FC)**

2.5

1.0

−0.5

−2.0

−3.5

−5.0

−6.0 −4.5 −3.0 −1.5 0.0 1.5 3.0 4.5 6.0 7.5

**log2(limma FC)**

(b)

Figure 3.6: 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 log2FCs (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.

Lastly, the significant DARs identified by DESeq2 and filtered for FDR*<*0.01 and abs(FC)*>*1.5 were divided in those more accessible in CD14+ monocytes (open in monocytes) or CD4+ (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 3.7). Conversely, the top enriched features for DARs open in monocytes included eRNAs, H3K27ac and DHSs in monocytes. Overall, this enrichment analysis confirmed the ability of this differential analysis method to identify significant and robust DARs that highlight cell-type specific regulatory regions for CD14+ monocytes and CD4+ cells.

Monocytes eRNA

**Open in monocytes Open in CD4+**

Monocytes H3K4me1

Monocytes DHS

Monocytes H3K27ac

T cells eRNA

**Pathway**

CD4+ DHS

CD4+ H3K27ac

CD4+ H3K4me1

CD4+ H3K4me3

0 1 2 3 4

**Fold−change enrichment**

Figure 3.7: Enrichment analysis for the for the significant DARs identified by DESeq2 between CD14+ monocytes and CD4+ 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 CD4+ (open in CD4+).

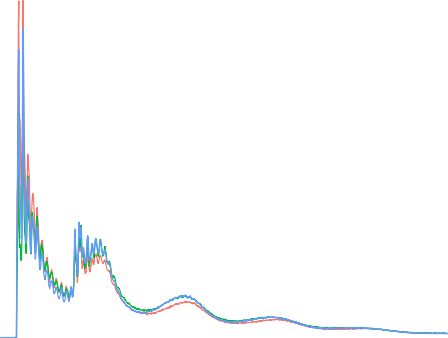
## Assessment of ATAC-seq transposition times in relevant cell types

The duration of the transposition reaction was optimized for the main immune cell types of interest for this thesis. ATAC-seq was performed for three different transposition times (20, 30 and 40 min) in CD14+ monocytes, CD4+, CD8+ and CD19+ cells and the impact on various ATAC-related readouts was explored. All three transpostion times produced appropriate fragment size distributions which recapitulated the nucleosome periodicity pattern (Figure 3.8 a). The duration of the transposition time 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 non-redundant fraction and nucleosome bound fragments (NBF*>*150bp) was calculated (Figure 3.8 b). The relationship of this ratio with transposition time was heterogenous between cell types. For example, CD8+ presented the greatest proportion of NFR for 20 min of transposition whereas the NFR/NBR reached a maximum at 40 min for the CD14+ monocytes. However, the change in this ratio across transposition times was very moderate in all cell types but CD8+ cells.

Transposition times did not significantly affect the signal-to-noise ratio measured as fold-enrichment at the TSS, with the largest enrichment values corresponding to different times across the four cell types (Figure 3.8 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 modest (not more than 4 units difference) across all cell types (data not shown). Before performing this formal comparison of transposition times, some sample recruitment had

0.65



**CD14 CD4 CD8 CD19**

**20 min**

**30 min**

**40 min**

10.0

0.60

0.55

**Normalised insert size density (x10−3)**

7.5 0.50

**Fragment ratio (NFF/NBF)**

0.45

0.40

5.0

0.35

2.5

0.30

0.25

0.20

0.0

0 200 400 600 800 1000

**Fragment length (bp)**

(a)

0.15

0.10

20 30 40

**Time (min)**

(b)

**16**

**20 min**

**30 min**

**40 min**

**14**

**12**

**10**

**Enrichment**

**8**

**6**

**4**

**2**

**0**

**−2000 −1000 0 1000 2000**

**Distance to TSS (bp)**

(c)

Figure 3.8: 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, CD4+, CD8+ and CD19+ cells.

≤

already been conducted using standard ATAC-seq with transposition for 40 min, as it was found to be the most appropriate condition based on the tapestation profiles from library QC (here not shown). Although this analysis suggests 30 min is the best condition for most of the cell types tested, the differences between 30 min and 40 min were minor and therefore for consistency across the cohort 40 min was used for all other patient samples processed using ATAC-seq.

3.2.3 Comparison of standard ATAC-seq with Fast-ATAC protocol

An improved Fast-ATAC protocol (Corces et al. 2016) was reported during the time the work in this thesis was being undertaken and therefore the Fast-ATAC protocol was compared with standard ATAC-seq. The aim was to confirm the two main reported advantages of Fast-ATAC, namely reduction of mitochondrial reads and signal enhancement, before implementing it as a replacement for the ATAC-seq protocol, which had been being used for processing patient samples (Chapter x).

[Describe samples analysed and data generation].

The percentage of mitochondrial reads in the Fast-ATAC libraries was lower for all cell types analysed compared to ATAC-seq (Figure 3.9 a). Importantly, the mitochondrial

percentage in CD4+ and CD8+ cells showed a reduction from 50% to less than

20%. With regard to the background signal, differing trends were observed across cell types (Figure 3.9 b) with no change in TSS enrichment for CD14+ monocytes or CD4+ T cells while CD8+ and CD19+ cells showed a reduction with Fast-ATACseq compared to standard ATAC-seq. Overall, the large decrease in mitochondrial reads together with the reduced duration of the experimental protocol supported the replacement of ATAC-seq by Fast-ATAC for future patient recruitments (see Chapter ??).

25

**CD14**

**CD4**

**CD8 CD19**

**CD14**

**CD4**

**CD8 CD19**

50

20

40

15

mitochondrial**MT reads (%)**

**TSS enrichment**

30

10

20

10

ATAC

**Method**

(a)

FAST−ATAC

5

ATAC−seq

**Method**

(b)

Fast-ATAC

Figure 3.9: Differences in mitochondrial DNA abundance and TSS enrichment between ATAC- seq and Fast-ATAC protocols. Representation of changes in a) percentage of mitochondrial reads and b) TSS fold-enrichment between ATAC-seq and Fast-ATAC libraries for CD14+ monocytes, CD4+, CD8+ and CD19+ cells.

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

A particular aim of this this thesis was to characterise the regulatory landscape in keratinocytes, one of the most relevant cell types in psoriasis pathophysiology. In order to assess the feasibility of using the standard ATAC-seq protocol (Buenrostro et al 2013) ( referred to ATAC 1 in this subsection), a cell suspension from a psoriatic lesional skin biopsy was generated and ATAC-seq was performed using 50,000 cells for two different transposition times (30 and 40 min). Since biopsy handling and lesional epidermal keratinocytes 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-

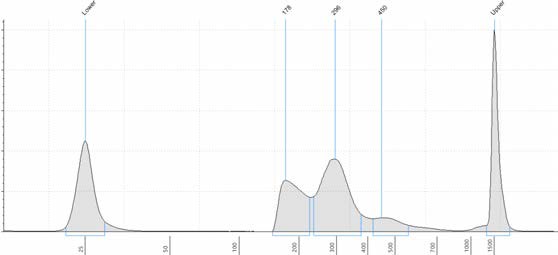
free and nucleosome-bound DNA (Figure 3.11 a). This was consistent with the fragment size distribution from the NGS data, presenting NFF and NBF (mono-and di-nuclosomes only) for both transposition times (Figure 3.11 b).

However the relative abundance of the mono-nucleosome fragments appeared to be approximately equal to the NFF, which is not observed in higher quality libraries. Regarding the signal-to-noise ratio, libraries for both transposition times presented TSS fold enrichment below the acceptable cut-off threshold of 6, showing slightly better signal (3.5 fold-enrichment) in the 30 min library (Figure 3.11 c). Cell suspensions obtained from skin biopsies using trypsinisation of the epidermal layer are enriched in KC, constituting approximately 90% of cells. However, they also contain significant amounts of dead cells and free-DNA released by apoptotic cells which could contribute to the increased background noise observed above.

Following Buenrostro and colleagues’ ATAC-seq protocol, a modified version for keratinocytes by Bao and colleagues (named ATAC-seq 2 in this subsection) and the Fast-ATAC protocols were released (Bao et al. 2015; Corces et al. 2016). 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 2015 paper protocol, to prevent background noise due to presence of dead cells affecting 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 an adherence assay, allows the isolation of viable undifferentiated keratinocytes. 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 (2, ATAC 2 and Fast-ATAC using C1 conditions, Table 3.3) were performed directly on the plate adherent cells with no cell detachment step.

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

**5000**



**Size (bp)**

**Lower**

**178**

**296**

**450**

**Upper**

**4000**

**Intensity**

**3000**

**2000**

**1000**

**0**

**25**

**50**

**100**

**200**

**300**

**400**

**500**

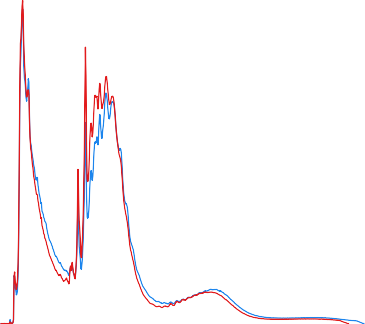
**700**

**1000**

**1500**

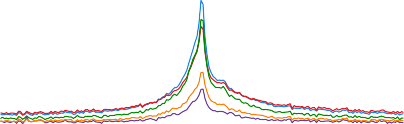
(a)

7.5



**Lesional 30 min**

**Lesional 40 min**



**Lesional 30 min**

**Lesional 40 min**

**NHEK ATAC−seq 1**

**NHEK ATAC−seq 2**

**NHEK Fast−ATAC**

7.0

6.5

7.5

**Normalised insert size density (x10−3)**

6.0

5.5

5.0

4.5

**Enrichment**

5.0 4.0

3.5

3.0

2.5

2.5

2.0

1.5

0.0

0 100 200 300 400 500 600 700

**Fragment length (bp)**

(b)

1.0

0.5

−2000 −1000 0 1000 2000

**Distance to TSS (bp)**

(c)

Figure 3.10: QC assessment of different ATAC protocols in keratinocytes from a psoriatic lesion 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 keratinocytes in suspension isolated from lesional skin from a psoriasis patient. Two transposition times (30 and 40 min) were tested in the same sample keratinocyte suspension and the tapestation profile for the 30 min transposition has been included as representative. c) Fold-enrichment of ATAC fragments across the Ensembl annotated TSS from the ATAC 1 psoriasis lesional keratinocytes libraries (previously described 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 | Two steps | 0.05% NP-40 and 5µL Tn5 |
| (Bao et al. 2015) |  |  |
| Fast-ATAC |  | C1∗: 0.01% digitonin, 2.5µL Tn5 C2: 0.01% digitonin, 0.5µL Tn5 |
| (Corces et al. 2016) | One step | C3: 0.025% digitonin, 0.5µL Tn5 |
|  |  | C4: 0.025% digitonin, 2.5 µL Tn5 |

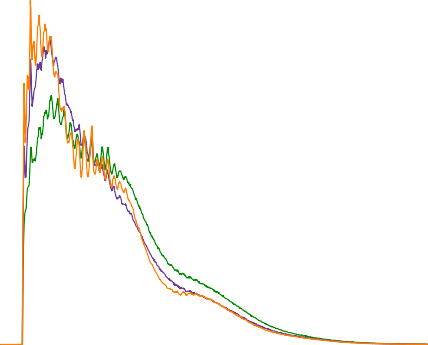
Table 3.3: Description of the most relevant parameters from the ATAC-seq and FAST- ATAC protocols assayed in NHEK and skin biopsies. Transposition times for all protocols was 30 min. ∗corresponds to the original Fast-ATAC conditions from Corces *et al.*, 2016.

psoriatic lesional epidermis, TSS enrichments were very low, particularly for the ATAC 2 protocol, and in all the instances under the acceptable cut-off using the two cell systems (Figure 3.11 c and Figure B.2).

Further optimisation of the Fast-ATAC protocol was performed by modifying the original concentration of NP-40 detergent and the Tn5 enzyme (Table 3.3, C1-4). Library QC using tapestation profiles to assess the DNA fragment size distributions failed to show the nucleosome pattern profile expected in ATAC, and therefore the samples did not proceed to NGS (Figure B.3).

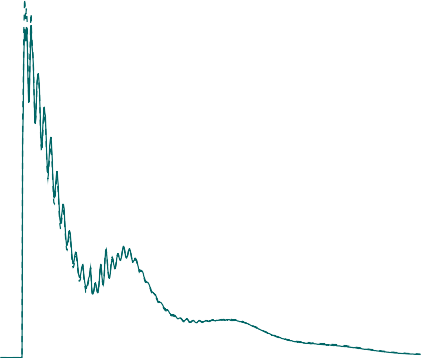
Towards the end of experimental work for this thesis, a new protocol called Omni-ATAC was published (Corces et al. 2017). Omni-ATAC was a protocol suitable for every cell type, in contrast to ATAC 1 and Fast- ATAC which were optimised for hematopoietic cells (Buenrostro2013; Corces et al. 2016). Performance 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 3.11 b). Moreover, high TSS enrichment values (approximately 20 fold) were observed for the two replicates (Figure 3.11 c). When performing overlap between the

6



**NHEK ATAC 1**

**NHEK ATAC 2 NHEK FAST−ATAC**



**NHEK Omni−ATAC Rep 1**

**NHEK Omni−ATAC Rep 2**

**Normalised insert size density (x10−3)**

7.5

**Normalised insert size density (x10−3)**

4

5.0

2 2.5

0

0 100 200 300 400 500 600 700

**Fragment length (bp)**

(a)

0.0

0 100 200 300 400 500 600 700

**Fragment length (bp)**

(b)

15

**NHEK Omni−ATAC Rep 1**

**NHEK Omni−ATAC Rep 2**

10

**Enrichment**

5

−2000 −1000 0 1000 2000

**Distance to TSS (bp)**

(c)

Figure 3.11: 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. c) Fold-enrichment of ATAC fragments across the Ensembl annotated TSS from two Omni-ATAC technical replicates.

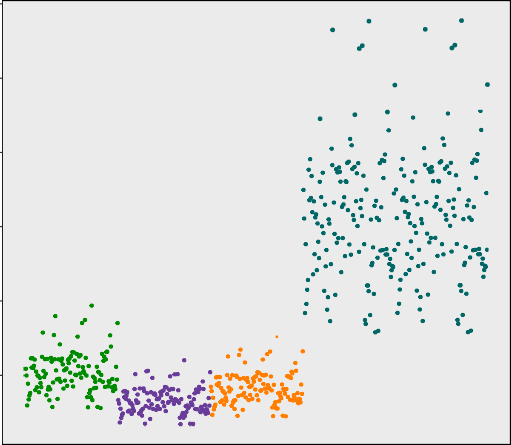
Omni-ATAC sample peaks filtered for a low stringency pval (pval=0.01) and the ENCODE DHSs from 125 cell types, the highest percentage (approximately 55%) was reported for NHEKs (Figure 3.12 a).

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 keratinocytes, even at a very low pval filtering. The differences in the quality of ATAC signal between Omni- ATAC and the prior ATAC protocols was clearly observed at the chr17 locus harbouring a number of keratin genes, which encode the main components of keratinocyte cytoskeleton (Figure 3.12 b). Omni-ATAC clearly showed 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.

## 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 as immediate processing of freshly acquired cells may not be feasible. In the context of this thesis, two different possible approaches involving cryopreservation and fixation were of interest and a collaborative project to investigate these was established with High-Throughput Genomics at the WHG. The first approach was the cryopreservation of PBMCs in liquid nitrogen using DMSO followed by thawing, recovery and FACS isolation of the cell population of interest (Figure 3.13). Secondly, the performance of an optimised protocol developed by High-Throughput Genomics using DSP in scRNA-seq (Attar et al. 2018) was

60

NHEK

NHEK

NHEK

NHEK

NHEK

50

40

**Overlap (%)**

30

20

10

**ATAC 1 ATAC 2 Fast−ATAC Omni−ATAC Omni−ATAC**

(a)

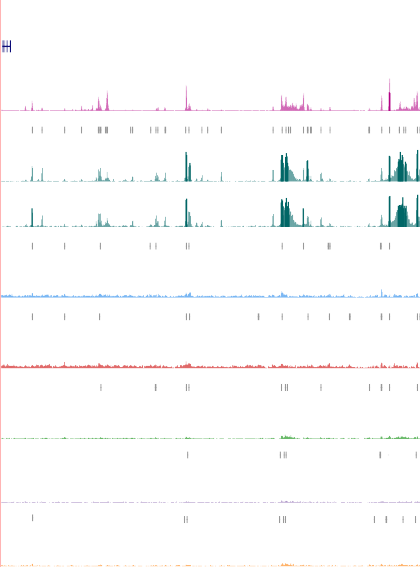
**Rep 1**

**Rep 2**

Scale chr17:

50 kb hg19

39,700,000 39,750,000



**UCSC genes**

266 \_

**NHEK ENCODE**

*KRT16*

*KRT13*

*KRT15*

*KRT9*

*KRT14*

*KRT16*

*KRT17*

1 \_

**Omni-ATAC peaks**

143 \_

**NHEK Omni-ATAC Rep 1**

1 \_

143 \_

**NHEK Omni-ATAC Rep 2**

0 \_

**Peaks lesional 30 min**

143 \_

Lesional 30 min

1 \_

**Peaks lesional 40 min**

143 \_

**Lesional 40 min**

1 \_

**Peaks NHEK ATAC 1**

143 \_

**NHEK ATAC-seq 1**

1 \_

**Peaks NHEK ATAC 2**

143 \_

**NHEK ATAC-seq 2**

1 \_

**Peaks NHEK Fast-ATAC**

143 \_

**NHEK Fast-ATAC**

1 \_

(b)

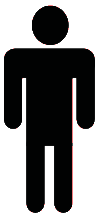
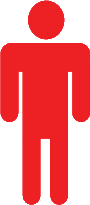
Figure 3.12: Comparison of four different ATAC-seq protocols applied to psoriasis lesional keratinocytes and NHEK cells, and published ENCODE data. (A) Enrichment (% overlap) of observed peaks from each sample (using low stringent pval) with open DHS chromatin regions in 125 ENCODE cell types. (B) Tracks showing open chromatin at the chr17 keratin family gene locus for different ATAC-seq protocols with the normalised ATAC read density (y-axis) shown.

investigated as a short term preservation method for FACS-isolated relevant cell types.

In order to investigate the performance of these two strategies, blood from three healthy volunteers matched for sex and age was processed on different days to simulate the experimental design when using patient samples (experimental design summarised in Figure 3.13). PBMCs were prepared from 90ml blood using a Ficoll gradient and CD14+ monocytes and CD4+ T cells were isolated by FACS as detailed in Chapter 2. ATAC- seq was performed on 50,000 CD14+ and CD4+ cells, either freshly isolated or after fixation with DPS, stored at 4◦C for 24h and then processed for ATAC-seq (Figure 3.13 Day 1, ATAC-seq fixed). To investigate the utility of cryopreservation, 70x106 million 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 2). After recovery, CD14+ and CD4+ cells were isolated from the PBMCs by FACS and then ATAC-seq performed. For each volunteer, three matched ATAC-seq libraries were generated in the two different cell populations: ATAC-seq fresh, ATAC-seq fixed and ATAC-seq frozen.

Day 0

CD14+ CD4+



FACS

separation

Day 1

**ATAC-seq fresh**



90 mL

blood

PBMCs

isolation

CD14+ CD4+

PBMCs

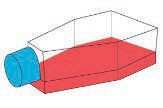
Freezing

~70x106

thawing

Fixation 24h

Culture 30 min



**ATAC-seq fixed**

CD14+ CD4+

FACS

separation

**ATAC-seq frozen**

Establishment of methods to assess genome-wide chromatin accessibility

111

Day 0

Day 7

Figure 3.13: Experimental design to assess the impact of cryopreservation and fixation in the chromatin accessibility of immune primary cells. Three healthy control individuals were recruited on different days, PBMCs were isolated and freshly isolated CD14+ monocytes or CD4+ T cells processed for ATAC-seq immediately (ATAC-seq fresh), after fixation with DPS (ATAC-seq fixed) or after cryopreservation of PBMCs (ATAC-seq frozen).

Chromatin accessibility observed in the different experimental conditions

All samples from each of the two cell types had more than 15 million reads, which have previously been shown as the minimum for successful ATAC-seq analysis and peak calling (Figure 3.14). The median number of reads across the fresh, frozen and fixed were more similar for CD14+ monocytes (58.6, 64.2 and 39.6 million reads, respectively) than in the CD4+ samples, where the frozen and fixed presented lower median total reads compared to the controls (43.8,

32.9 and 28.8 million reads respectively) (Figure 3.14).

90 70

**CTL1 CTL2 CTL3**

**CTL1 CTL2 CTL3**

80

60

70

**Total reads after filtering (million)**

**Total reads after filtering (million)**

50

60

50 40

40

30

30

20

20

10 10

0

fresh frozen fixed

**Cell type**

(a)

0

fresh frozen fixed

**Cell type**

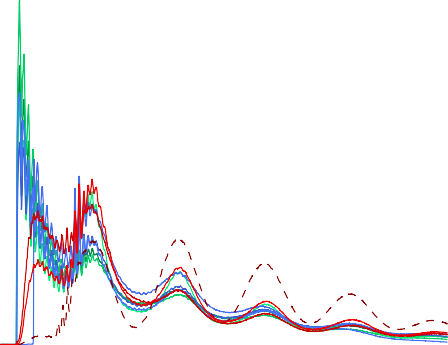
(b)

Figure 3.14: Total number of ATAC-seq reads for the fresh, frozen and fixed CD14+ monocytes and total CD4+ samples for three volunteers (CTL1-3). Representation of million reads after filtering for the fresh, fixed and frozen ATAC-seq libraries in a) CD14+ monocytes and b)CD4+ cells.

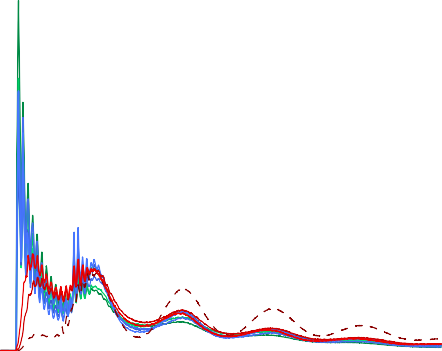
The ATAC-seq signal-to-noise ratios across the TSS showed a similar median for the fresh and fixed CD14+ monocytes libraries (17.4 and 16.5 fold-enrichment, respectively) and was higher for the frozen samples (26.3 fold- enrichment) (Table A.1). The median TSS enrichments in the frozen and fixed CD4+ samples were considerably higher (16.1 and 14.3 fold-enrichment, respectively) than the fresh samples (5.6), which were borderline for the ENCODE recommended threshold. For one volunteer (CTL1) the fixed samples of both cell types showed 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 CD4+ from CTL1, were similar, showing NFR *<*150bp and fragments corresponding to mono-, di-, tri- and tetra- nucleosomes (Figure 3.15 a and b). The fixed samples presented a lower density of NFF when compared to fresh and frozen, which had very similar distributions in both cell types. In particular, fixed CD14+ monocytes and CD4+ in CTL1 had extremely low abundance of NFF (Figure 3.15 red- dashed line in a and b), consistent with the very low TSS enrichment for CTL1 CD14+ ATAC-seq fixed cells as previously highlighted.

12.5



**CTL1 fresh CTL2 fresh CTL3 fresh CTL1 frozen CTL2 frozen CTL3 frozen CTL1 fixed CTL2 fixed CTL3 fixed**



**CTL1 fresh CTL2 fresh CTL3 fresh CTL1 frozen CTL2 frozen CTL3 frozen CTL1 fixed CTL2 fixed CTL3 fixed**

8

10.0

**Normalised insert size density (x10−3)**

6

**Normalised insert size density (x10−3)**

7.5

4

5.0

2

2.5

0

0 200 400 600 800 1000

**Fragment length (bp)**

(a)

0.0

0 200 400 600 800 1000

**Fragment length (bp)**

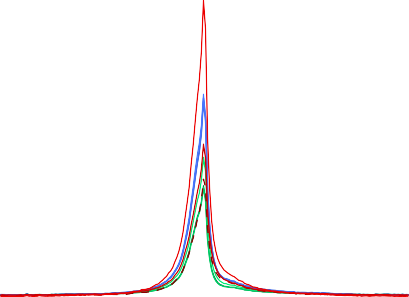
(b)

Figure 3.15: Fragment size density distribution for ATAC-seq fresh, fixed and frozen in CD14+ monocytes and CD4+ cells. The distribution of ATAC-seq fragment lengths are illustrated for a) CD14+ monocytes and b) CD4+ cells and colour-coded by condition (fresh=green, frozen=blue and fixed=red).

Chromatin structure across and within the TSS was then investigated following the approach described in (Scharer et al. 2016). The nucleosome-free fragments (*<*150bp) from all the samples showed a single peak of enrichment at the nucleosome-depleted TSS position, with CD4+ fresh samples presenting the lowest enrichment (Figure 3.16 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 3.16 b and d).

60



**CTL1 CD14 fresh CTL2 CD14 fresh**

**CTL3 CD14 resh CTL1 CD14 frozen CTL2 CD14 frozen CTL3 CD14 frozen CTL1 CD14 fixed CTL2 CD14 fixed CTL3 CD14 fixed**

55

50

45

40

**Enrichment**

35

30

25

20

15

10

5

0

−2000 −1000 0 1000 2000

**Distance to TSS (bp)**

(a)

60

55



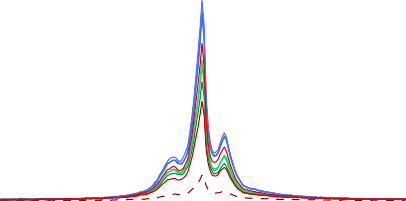
**CTL1 CD14 fresh CTL2 CD14 fresh**

**CTL3 CD14 resh CTL1 CD14 frozen CTL2 CD14 frozen CTL3 CD14 frozen CTL1 CD14 fixed CTL2 CD14 fixed CTL3 CD14 fixed**

50

45

40

35

**Enrichment**

30

25

20

15

10

5

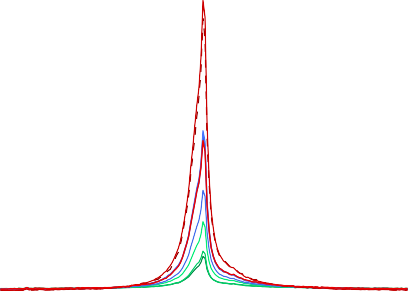
0

−2000 −1000 0 1000 2000

**Distance to TSS (bp)**

(b)

55 55



**CTL1 CD4 fresh CTL2 CD4 fresh CTL3 CD4 fresh CTL1 CD4 frozen CTL2 CD4 frozen CTL3 CD4 frozen CTL1 CD4 fixed CTL2 CD4 fixed CTL3 CD4 fixed**



**CTL1 CD4 fresh**

**CTL2 CD4 fresh CTL3 CD4 fresh CTL1 CD4 frozen CTL2 CD4 frozen CTL3 CD4 frozen CTL1 CD4 fixed CTL2 CD4 fixed CTL3 CD4 fixed**

50 50

45 45

40 40

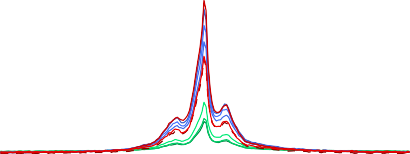
35 35

**Enrichment**

**Enrichment**

30 30

25 25

20 20

15 15

10 10

5 5

0 0

−2000 −1000 0 1000 2000

**Distance to TSS (bp)**

(c)

−2000 −1000 0 1000 2000

**Distance to TSS (bp)**

(d)

Figure 3.16: ATAC-seq enrichment of nucleosome-free and di-nucleosome fragments at the TSS and surroundings in CD14+ monocytes and CD4+ 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 in tagging NBF (di-, tri- and tetra- nucleosomes) but the loss of chromatin structure around the TSS indicated that those fragments are likely to have been displaced from their original location. Conversely, the fresh CD4+ samples showed low enrichment for NFF in the TSS,

in line with their overall TSS enrichment (Table A.1) and indicating weakly the two nucleosomes positioned in the TSS surroundings. In summary, 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 3.2.1) revealed that the highest proportion of ATAC-seq peaks localized to promoters, introns and intergenic regions(Figure 3.17 a and b), consistent with previous studies (Buenrostro2013; Scharer et al. 2016). The higher percentage of peaks annotated as promoters for some ATAC-seq libraries revealed the 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.

65 65



**CTL1 CD14 fresh CTL2 CD14 fresh CTL3 CD14 fresh CTL1 CD14 frozen CTL2 CD14 frozen CTL3 CD14 frozen CTL1 CD14 fixed CTL2 CD14 fixed CTL3 CD14 fixed**



**CTL1 CD4 fresh CTL2 CD4 fresh CTL3 CD4 fresh CTL1 CD4 frozen CTL2 CD4 frozen CTL3 CD4 frozen CTL1 CD4 fixed CTL2 CD4 fixed CTL3 CD4 fixed**

60 60

55 55

50 50

45 45

**Percentage (%)**

**Percentage (%)**

40 40

35 35

30 30

25 25

20 20

15 15

10 10

5 5

0 0

5’ UTRs

3’ UTRs

Exons

Introns

Promoters

5’ UTRs

3’ UTRs

Exons

Introns

Intergenic

**Annotation**

Promoters

Downstream

Intergenic

regions

Downstream

regions

(a)

**Annotation**

(b)

Figure 3.17: Genomic features annotation for the ATAC-seq peaks called in each of the fresh,frozen and fixed samples from CD14+ monocytes and CD4+. Overlap was performed between the genomic features and the list of a) CD14+ monocytes and b) CD4+ 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 the data relating to CTL1 fixed samples from CD14+ and CD4+ cells were removed given the low quality metrics previously described.

20

**CTL1 CTL2 CTL3**

**CTL1 CTL2 CTL3**

30

10

0 0

**PC2 (18%)**

**PC2 (24%)**

−10

−30

−20

**Fresh Fixed Frozen**

**Fresh Fixed Frozen**

−60

−50 0 50

**PC1 (51%)**

(a)

−30

−20 0 20 40

**PC1 (47%)**

(b)

Figure 3.18: 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) CD4+ cells from the same three healthy individuals. The first two PCs (x-axis and y-axis, respectively) of all the regions included in each of the master lists are plotted. Each point represents a sample, where shape codes for individual (CTL1, CTL2, CTL3) and colour means condition (fresh, fixed, frozen). The proportion of variation explained by each principal component is indicated.

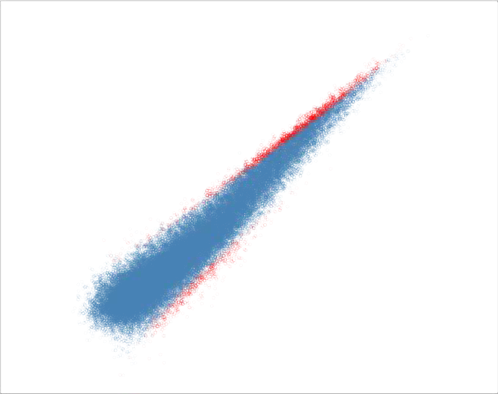
Principal components analysis (PCA) was performed using DESeq2 normalised counts for each of the two master lists demonstrated differences across the three ATAC-seq conditions in the two cell types. Plotting the first two principal components (PCs) showed sample clustering based on condition, which explained the largest variability within 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 3.18). The second PC (16% of variance) showed more moderate changes between fresh and fixed libraries. In contrast, PCA analysis in CD4+ showed fixed to present the largest differences versus fresh in the genome-wide accessibility (Figure 5.3 b).

To further compare chromatin accessibility across the three conditions, the normalised read counts at each of the ML CD14 all cond and ML CD4 all cond master lists peaks were contrasted between fresh and fixed or frozen. 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 3.19 a, b, c and d).

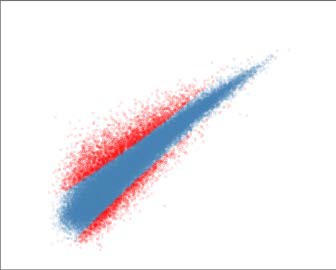
**CD14+ monocytes CD14+ monocytes**

12.5 12.5



**R=0.918**

**Not sig. FDR<0.01 and abs(FC)>1.5**



**R=0.952**

**Not sig. FDR<0.01 and abs(FC)>1.5**

11.0 11.0

**ATAC−seq fixed**

**(mean log2 normalised counts)**

**ATAC−seq frozen**

**(mean log2 normalised counts)**

9.5 9.5

8.0 8.0

6.5 6.5

5.0 5.0

3.5

3.5 5.0 6.5 8.0 9.5 11.0 12.5

**ATAC−seq fresh**

**(mean log2 normalised counts)**

(a)

**CD4+**

3.5

3.5 5.0 6.5 8.0 9.5 11.0 12.5

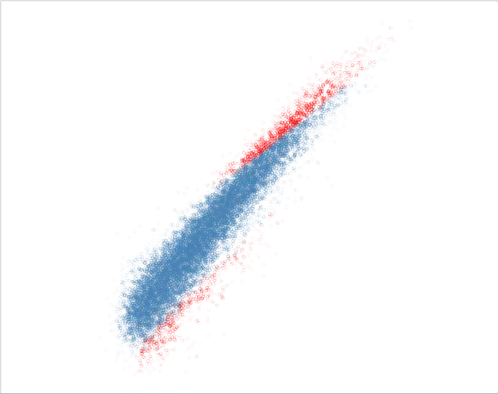
**ATAC−seq fresh**

**(mean log2 normalised counts)**

(b)

**CD4+**

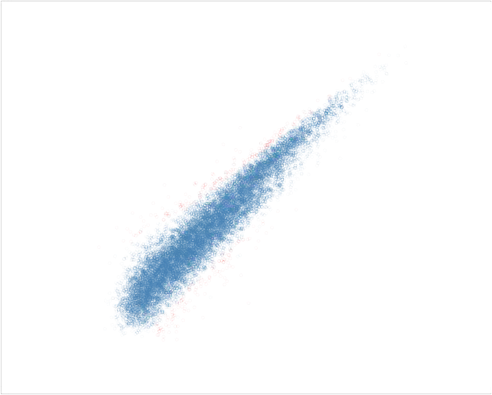
10.5



**R=0.940**

**Not sig. FDR<0.01 and abs(FC)>1.5**

10.5



**R=0.941**

**Not sig. FDR<0.01 and abs(FC)>1.5**

9.0

**(mean log normalised counts)**

9.0

**ATAC−seq frozen**

**(mean log2 normalised counts)**

7.5

**ATAC−seq fixed**

**2**

7.5

6.0

6.0

4.5

4.5 6.0 7.5 9.0 10.5

**ATAC−seq fresh**

**(mean log2 normalised counts)**

(c)

4.5

4.5 6.0 7.5 9.0 10.5

**ATAC−seq fresh**

**(mean log2 normalised counts)**

(d)

Figure 3.19: Comparison of the log2 normalised ATAC-seq counts at the consensus master lists peaks in fresh, fixed and frozen conditions. Each plot shows the comparison of ATAC-seq log2 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 lower correlation in mean counts were 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 3.4). In CD14+ monocytes, the largest number of DARs versus fresh were reported for the ATAC-seq frozen (5,269 regions in Figure 3.19 b). Conversely, in CD4+ the greatest differences in chromatin accessibility were found between fresh and

fixed ATAC-seq samples (1,564 DARs in Figure 3.19 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 CD4+ samples partially skewed the normalisation process, as can be observed by the location of the significant DARs in Figure 3.19 d. Altogether, the number of DARs identified in each comparison did not involve more than 14.2% 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%)

CD4+ 282 (2.1%) 1,564 (14.2%)

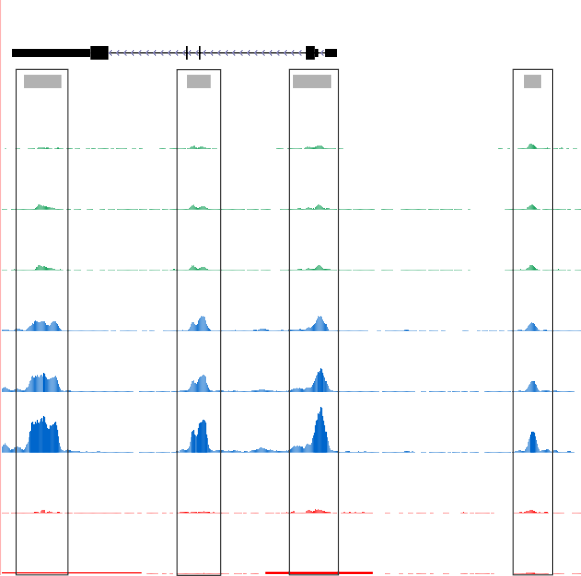
Table 3.4: Summary results from the differential chromatin accessibility analysis comparing ATAC-seq frozen or fixed chromatin landscape to the reference ATAC- seq fresh. Number of significant DARs (FDR*<*0.01 and abs(FC)*>*1.5) identified using ML CD14 all cond and ML CD14 all cond filtered for background counts using 80% cut-off (optimal cut-off identified for this analysis, data not shown). In brackets are shown the percentage of 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 of the *TNFSF14* gene (Figure 3.20. 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.

Scale chr19:

5 kb hg19

6,665,000 6,670,000 6,675,000



**UCSC genes** *TNFSF14*

ATAC-seq peaks

366 \_

**CTL1 fresh**

**CTL2 fresh**

**CTL3 fresh**

**CTL1 frozen**

**CTL2 frozen**

**CTL3 frozen**

**CTL2 fixed**

**CTL3 fixed**

0 \_

366 \_

0 \_

366 \_

0 \_

366 \_

0 \_

366 \_

0 \_

366 \_

0 \_

366 \_

0 \_

366 \_

0 \_

Figure 3.20: 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).

# Discussion

The aim of this chapter was to establish a data analysis pipeline for ATAC-seq and compare various experimental protocols. This was the first time ATAC-seq was used in the research group. A particular focus was to consider appropriate methodologies for clinical studies, where sample availability and quality may be severely limiting, and a number of alternative protocols, metrics and algorithms described in early ATAC reports were evaluated in the pilot experiments presented in this chapter. This enabled the establishment a pipeline and approach to be implemented for investigation of psoriasis and PsA chromatin landscape (Chapter x).

3.3.1 ATAC: methodological aspects and pipeline establishment

At the time of the first ATAC-seq publication (Buenrostro2013), well established protocols for complete processing and data analysis of ATAC-seq were lacking. 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 3.x) as well as alternative protocols such as THS-seq to assess chromatin accessibility in low number of cells.

**Table 3.x: 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 NA IDR filtered list or only peaks present in the two samples pseudoreplicates.

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).

Establishment of methods to assess genome-wide chromatin accessibility

81

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

Scharer *et al.*

2016

MACS2 (-nomodel -extsize 147) Merge of peaks from all samples

in an iterative process including permutations

HOMER (-style dnase) Merge of all overlapping peaks

between all samples using HOMER mergePeaks

Peak based: quantile normalisation and Fisher exact text (FDR*<*0.05).

Peak based: TMM normalisation and edgeR package (FDR*<*0.05).

Establishment of methods to assess genome-wide chromatin accessibility

82

The data and analysis presented in this chapter has confirmed some limitations of the ATAC-seq and 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. Both correlated well with the overall differences in sample quality from the ATAC-seq libraries used as an exemplar here. Importantly, TSS and FRiP were shown 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 (Corces et al. 2017). Similarly to TSS, FRiP proved to be informative in evaluating signal-to-noise ratios; however it relies on peak calling and thus is more likely to be biased. In agreement with these findings, 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 was demonstrated 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 in the peak filtering based on pseudoreplicate 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 formal testing of 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 the most appropriate for all cell types according to pre-sequencing library QC (tapestation). Assessment of three different transposition times in the ATAC-seq protocol showed heterogenous impact across cell types on the ratio of NFF/NBF and overall no major impact on signal-to noise ratios. The use of the improved Fast- ATAC protocol addressed some of the limitations identified by the ATAC-seq data generated within our group. Specifically, Fast-ATAC significantly reduced the percentage mitochondrial reads in all four cell types of interest for this thesis. In contrast, the improvement of signal-to-noise for hematopoietic cells claimed by the Fast- ATAC paper was not evident. In fact, Corces *et al.,* only showed improved TSS by Fast-ATAC. The publication of Omni-ATAC and a comprehensive comparison of the three ATAC protocols across a large number of cell types demonstrated that Fast-ATAC did not improve TSS fold-enrichment when compared to ATAC-seq

in some of the hematopoietic cells, for example CD19+ cells, consistent with the

finding in the systematic comparison using pilot data in this thesis.

3.3.2 The challenges of performing differential chromatin accessibility analysis

Studying chromatin accessibility in clinical samples first requires the definition of a consensus master list of accessible regions for which

no accepted method has been agreed. In this work, 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 of biological subgroup) but still be differentially accessible across conditions. Other publications have preferred building condition-specific master lists or simply include all the significant peaks called in all the analysed samples (Turner2018; Alasoo et al. 2018).

When used for differential analysis, an additional filtering step to account for high read counts in 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 in the literature. 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 3.1). 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 (Alasoo et al. 2018). 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 numbers 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 was chosen as an appropriately stringent method at the time study. As specific tools for ATAC analysis are developed, further comparison of the outputs will be of interest in future work.

## 3.3.3 Studying the chromatin landscape from psoriasis biopsies

At the time of writing, only RNA-seq studies have been performed in keratinocytes from psoriasis skin biopsies. The relevance of keratinocytes 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. Standard ATAC- (with different NP40 and Tn5 conditions) (ATAC 1, ATAC2) and Fast-ATAC were found to perform very poorly in keratinocytes isolated from skin biopsies and also in cultured NHEKs. The fact that similar results were obtained in keratinocytes isolated through different systems as well as in NHEKS indicated that the main reason for poor performance of these protocols was intrinsic to the cell type and not driven by compromising cell viability through the system used to isolate the cells. As differentiation progresses, keratinocytes synthesis an insoluble protein structure that progressively replaces the plasma membrane, which may have been impairing appropriate cell permeabilisation and efficient transposition. Interestingly, using increased Tn5 concentration to perform appropriately in NHEKs did not appear to improve ATAC quality libraries in the data presented. 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.

Recent release of the Omni-ATAC protocol and testing here showed substantially improved performance vs ATAC-seq and Fast-ATAC in keratinocytes. The 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 keratinocytes from psoriasis patients biopsies processed through adherent assay to minimise the presence of dead cells. This opens a new avenue to explore the chromatin landscape in lesional and uninvolved psoriasis biopsies.

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

The use of clinical samples sometimes involves logistical limitations that require sample preservation. At the time of starting this thesis the Oxford Genomics Centre at the WHG had implemented the use of DSP as a compatible fixative for microfluidics-based scRNA-seq methods. It was therefore of interest to test the ability of this fixative to perform well in ATAC (Attar et al. 2018). Alternatively cryopreservation of PBMCs had historically been used but formal assessment of the effect of this process in the chromatin landscape of 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. 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 relating to 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 than the differences between individuals. This was confirmed by performing 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 CD4+ cells and present greater differences in the chromatin accessibility landscape when compared to fresh samples. This has been previously reported in the context of cytokine response, where

cryopreservation of monocytes and immunature DCs have shown a skewed profile compared to fresh counterparts (Meijerink et al. 2011).

The results presented here could be limited by the borderline quality of some of the libraries, as previously explained, which was due to the issues of consistency intrinsic to 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 on the sample size foreseen in each particular study, if the aim is 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, ATAC has also been successfully conducted in frozen tissues and formaldehyde samples and can be considered when establishing the experimental design (Corces et al. 2017; Chen et al. 2016). 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.

## Conclusions

The work described here has compared commonly used strategies from 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 samples in a clinical setting whilst accounting for some quality limitations and the resources and expertise in the group available at the time. A robust pipeline was established and conclusions drawn regarding optimized protocols and conditions. As ATAC has become a more commonly used technique,

new methods and updates have been introduced on both the lab and analysis side which means that data presented throughout this thesis could be revisited in the future with further optimised analytical methods. Moreover, the implementation of Omni-ATAC for future sample recruitment is likely to further improve the quality and confidence in reported differentially open regions.

Later publications based on clinical samples have also used similar parameter for ATAC peak calling \parencite{Wang2018}.