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

Abstract i

Acknowledgements ii

Declarations iii

Submitted Abstracts iv

Associated Publications v

Contents vi

List of Figures viii

List of Tables ix

Abbreviations x

1 Establishment of laboratory methods and analytical tools to assess genome-wide chromatin accessibility in clinical samples 1

1.1 Introduction . . . 1

1.2 Results . . . 2

1.2.1 Establishment of an ATAC-seq data analysis pipeline based on current knowledge . . . 2

1.2.2 Assessment of ATAC-seq transposition times and comparison with FAST-ATAC protocol in relevant cell types 14

1.2.3 Impact of cryopreservation and fixation in the chromatin landscape of immune primary cells . . . 14

1.2.4 Limitations of ATAC-seq and FAST-ATAC to assess chromatin accessibility in KC . . . 17

1.2.5 Discussion . . . 22

2 Cross-tissue comparison of chromatin accessibility, gene expression signature and immunophenotypes in PsA 23

2.1 Introduction . . . 23

2.2 Results . . . 23

2.2.1 PsA patients cohort description and datasets . . . 23

2.2.2 The chromatin accessibility landscape in SF and PB immune cells . . . 26

2.2.3 Pathway and TFBS enrichment analysis highlight functional tissue-specific differences in chromatin accessibility . . . 34

2.2.4 Differential gene expression analysis in paired circulating and synovial immune cells 37
2.3 Discussion 37

Appendices **38**

A Establishment of methods to assess genome-wide chromatin accessibility **39**

B Cross-tissue comparison analysis in PsA **42**

List of Figures

1.1 Measurements for quality control assessment in ATAC-seq samples 7
1.2 Peak calling and sequencing depth in ATAC-seq samples 10
1.3 Peak calling filtering using IDR analysis in ATAC-seq samples . . 11
1.4 Differential chromatin accessibility analysis for different background reads cut-offs. 13
1.5 Exploration of the differential chromatin accessibility analysis using 80% as the empirical cut-off. 14
1.6 Assessment of the effect of transposition times on the ATAC-seq QC parameters 15
1.7 Differences in MT DNA abundance and signal specificity between ATAC-seq and FAST-ATAC protocols 16
1.8 Experimetal design to assess the impact of cryopreservation and fixation in the chromatin accessibility of immune primary cells. . . 17
1.9 QC assessment of ATAC-seq in KC enriched cell suspension derived from a psoriatic lesional skin biopsy 18
1.10 QC assessment of FAST-ATAC and Omni-ATAC in cultured NHEK 20
1.11 QC assessment of Omni-ATAC in NHEK and chromatin accessibility signal for the samples generated with the different ATAC-seq protocols 21

2.1 QC of FAST-ATAC PsA samples in four cell types 28
2.2 Combined PCA analysis of all four cell types isolated from blood and SF. 29
2.3 Enrichment of eQTLs in the combined cell types PsA accessible chromatin master list. 30
2.4 Annotation with genomic regions and chromatin states of the PsA DOCs from the four cell types differential analysis. 32
2.5 Enrichment of PsA DOCs for the FANTOM5 eRNA dataset 34

A.1 FAST-ATAC and Omni-ATAC NHEK tapestation profiles. 40
A.2 Assessment of TSS enrichment from ATAC-seq and FAST-ATAC in healthy and psoriasis skin biopsies samples. 41

B.1 Permutation analysis SF vs PB in $CD14^{+}$,$CD4m^{+}$,$CD8m^{+}$ and NK. 42

List of Tables

1.1 Summary table of ATAC-seq methodology analysis for peak calling, filtering and differential analysis. 3
1.2 ATAC-seq percentage of MT reads and fraction of reads in called peaks 8
1.3 Description of the most relevant parameter from the ATAC-seq and FAST-ATAC protocols assayed in NHEK and skin biopsies. . . 19

2.1 Description of PsA patients cohort recruitment and metadata. . . . 25
2.2 Datasets generated for the PsA cohort samples 26
2.3 Summary results of the chromatin accessibility analysis between SF and PB in PsA samples 31
2.4 Summary results of the chromatin accessibility analysis between SF and PB in PsA samples. 33
2.5 Distinct enriched pathways in $CD14^{+}$, $mCD4^{+}$, $mCD8^{+}$ and NK between SF and PB 35

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
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

Previous and current methods to identify the accessible genome in cells and tissues

Implementation of ATAC-seq to define the chromatin landscape

Technical limitations and recent advances in optimisation

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4473780/

Talk about ATAC being more variable, a native chromatin accessibility assessment without cross-linking. Role of transposase ability in accessing the chromatin, debri and DNA from dead cells adding noise

Paper to justify peak calling: A comparison of peak callers used for DNase-Seq data.

New ATAC but also explanations of the limitations: Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay

Challenges of working with clinical samples

1.2 Results

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

When the first ATAC-seq publication (**Buenrostro2013**) appeared, there were not well established protocols for the complete processing of the data. Since then, several publications have used ATAC-seq and modifications of this protocol together with a wide range of data analysis strategies to answer different biological questions (Table 1.1). There are several limiting aspects in the process of analysing ATAC-seq data, including QC assessment, peak calling/filtering and differential analysis of chromatin accessibility regions between groups. Using the current knowledge in the field as well as on my own analysis, I agreed on the most appropriate criteria and parameters to implement in our in-house pipeline. For this purpose, I used ATAC data generated with the first protocol (**Buenrostro2013**) in paired $CD14^+$ monocytes and $CD4^+$ total T cells from the same three healthy individuals, all of them downsamples to 30 million of reads, in order to facilitate the comparison across all of them.

Table 1.1: .

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 in 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).

Sample quality control

Regarding QC measurements, the variability in performance of the methodology, particularly ATAC-seq and Fast-ATAC, has required to agree on appropriate parameters to determine the quality of the samples before proceeding with downstream differential analysis. After reviewing the different read-outs implemented across different publications as well as the recently ENCODE update, I have identified the most informative ones showing supporting correlation between them.

Firstly, I analysed the fragment size distribution for each of the six samples in order to determine if they recapitulated the expected periodicity of nucleosomes protecting the DNA during the transposition event (Figure 1.1a). All the samples showed periodicity every ~200bp up to 600bp, clearly distinguishing chromatin organisation into mono-, di- and tri-nucleosomes. The relative intensity of nucleosome-free DNA fragments (<~147pb) 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 $CD14^+$). Nucleosome-free fragments(peak<~147bp) are also clearly distinguished in all of the samples, meeting the ENCODE QC recommendations (**ENCODE**).

Another QC measurement 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 TSS to allow formation of the transcriptional machinery and initiation of transcription. Fold-enrichment signals ranged between 5-7 for the $CD4^+$ samples and they were much higher(between 17-20) for the $CD14^+$ samples. The lower sample quality of the $CD4^+$compared to $CD14^+$ shown by the TSS signal were recapitulated by the ATAC-seq signal at the promoter of the constitutively expressed gene glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) (Figure 1.1c).

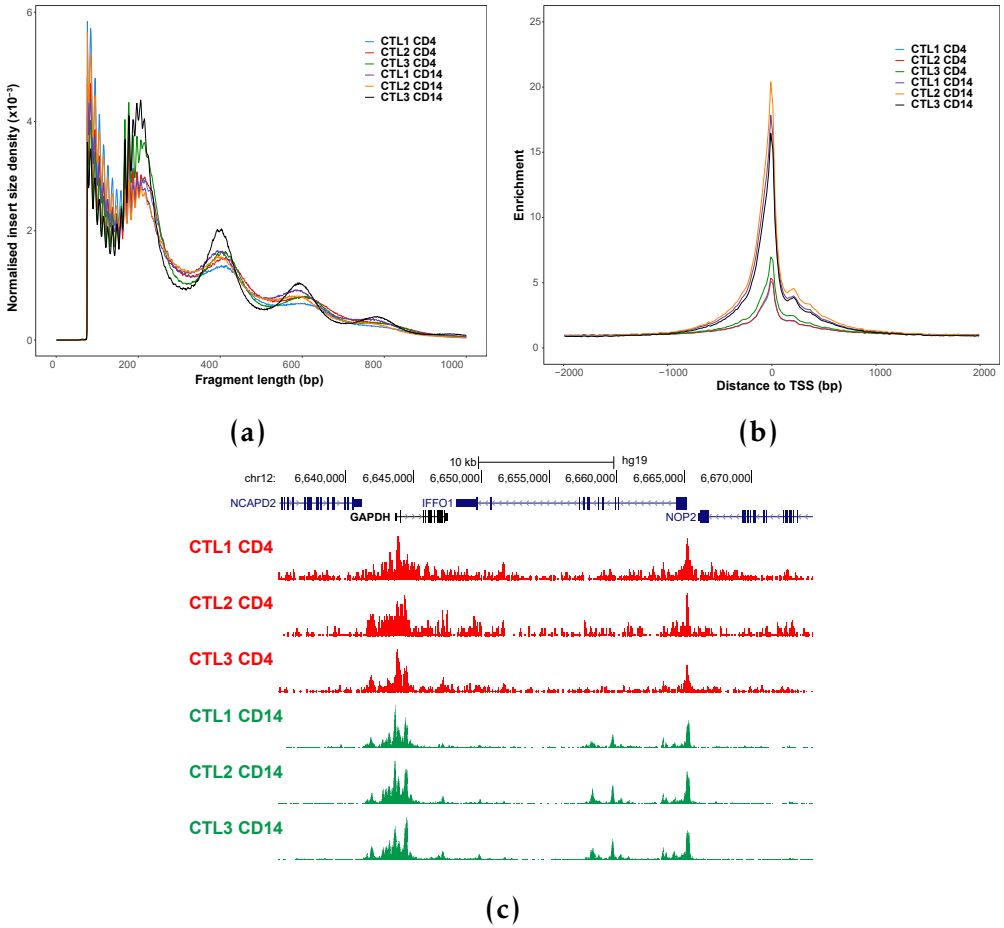


Figure 1.1: Measurements for quality control assessment in ATAC-seq samples

As part of the QC assessment I looked at the percentage of mitochondrial reads and the fraction of reads in peaks (FRiP)(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:

FRiP score is a way 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. Regarding TSS and FRiP cut-off values, Alsoo *et al.*, 2018 and, recently, ENCODE have recommended minimum FRiP between 10-20% and TSS between 6-10. ENCODE has prioritised the use of TSS over FRiP as the measurement to determine the noise in the sample (**ENCODE**). According to this recommendations all these samples passed QC; however clear differences were seen between $CD4^+$ and $CD14^+$ samples. The mitochondrial content ranged between 14.9-43.3% and, alike FRiP and TSS, it was higher in $CD14^+$ than in $CD4^+$ and not directly related with any of the other QC measurements.

Peak calling and filtering

As part of the ATAC-seq pipeline implementation, peak calling and the criteria for filtering where another two aspects to determine. Although different peak callers have been used, most of the publications as well as ENCODE has been using MACS2 as the preferred methodology (Table 1.1). MACS2 has been initially developed for ChIP but it has also been used for DHS and ATAC-seq

with disabling the model and agreeing in an extension size (–extsize) and a shift (–shift), which indicate the direction and number of bp for reads to be shifted and the number of bp for them to be extended, respectively. The –extsize should correspond to the average fragment size, which in my libraries is ~200bp and the –shift is set to -100, as it is recommended to be set to -1/2 of the fragment size for chromatin accessibility assays. This parameter could be further optimised but it escapes from the aim of this thesis.

I was interested in understanding the effect of sequencing depth and the sample quality on the peak calling to have a better control of both variables in the downstream analysis. I performed random read sub-sampling every 5M total reads (from 5M to 30M) followed by peak calling with arbitrary filtering for FDR<0.01 in each of the six aforementioned samples.

The number of called peaks passing filtering showed an steady increased over the read depth which seemed to reach a *plateau* around 25M reads (Figure 1.2a). This was consistent with the decay in the increments of called peaks over read depth, almost invariable, from 20M reads onwards (Figure 1.2b). Moreover, lower number of peaks were detected in $CD4^+$ samples compared to $CD14^+$ highlighting the influence of sample quality on the total number of called peaks. Interestingly, sample quality measured by FRiP reflected very low changes over read depth and was stable from 15M reads for all six samples (Figure 1.2c). Overall, this confirmed that measurement of sample quality by FRiP or TSS is not biased by sequencing depth.

Regarding peak calling filtering, most of the ATAC-seq publications using MACS2 have arbitrarily used an FDR<0.01 (Table 1.1). In collaboration with Dr. Gabriele Migliorini and following ENCODE pipeline, we explored the use of IDR to experimentally identify the most appropriate p-val for filtering each individual sample. Each sample was partitioned in two, peaks were called in each half and the percentage of peaks (over the total number shared peaks)

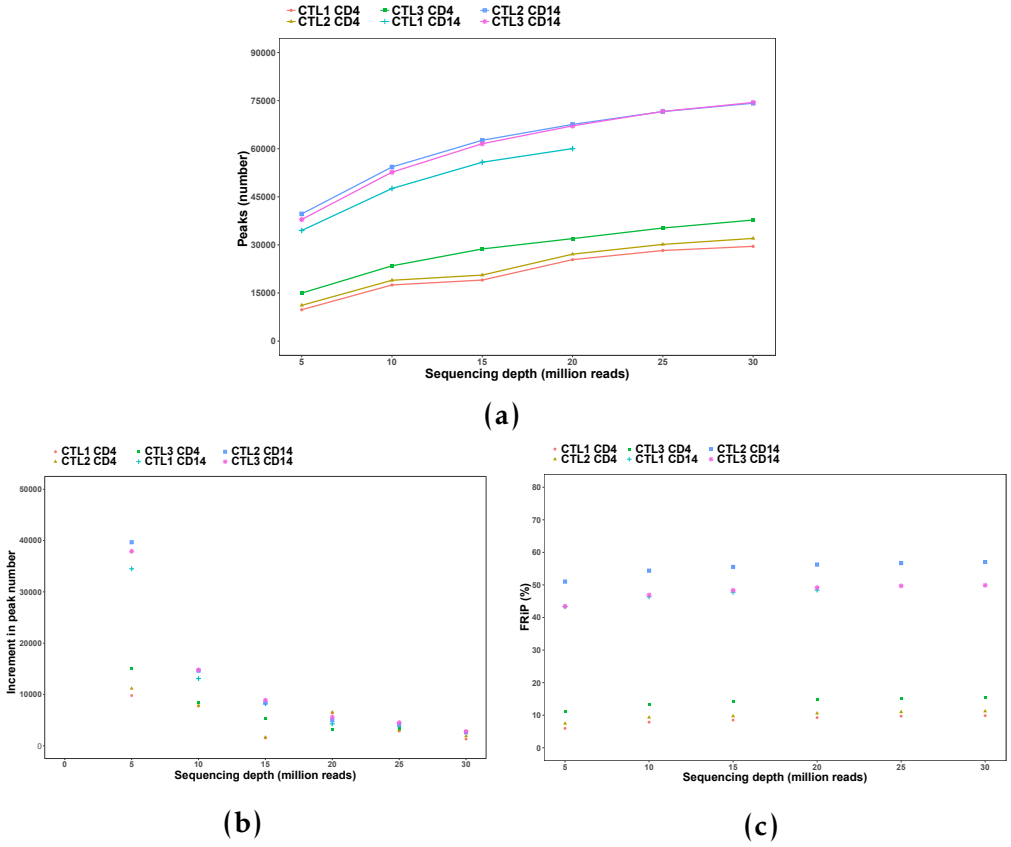


Figure 1.2: Peak calling at different sequencing depth in ATAC-seq samples

sharing IDR at a particular p-val was calculated (Figure 1.3 a and b). Both of the representative samples showed variation in the percentage of shared peaks upon sequencing depth under 10M reads, being the effect more pronounced and extended in the lower quality (CTL2 $CD4^+$ Figure 1.3 a) compared to the counterpart $CD14^+$ (Figure 1.3 b). The shape of the curves was also influenced by the sample quality, presenting a smoother profile reaching a single maximum percentage of shared IDR peaks for samples with TSS enrichment >∼10 compared to samples with lower quality. All the $CD14^+$ samples reached the maximum percentage of IDR shared peaks at approximately -log10 pval 8 (data not shown). Filtering the $CD4^+$ peaks at the -log10 pval of the first maximum of IDR shared peaks reduced the percentage of peaks overlapping noise (e.g heterochromatin, repetitive sequences and repressed regions) when compared to peaks filtered based on FDR<0.01 (Figure 1.3 b). In summary, this

IDR analysis appeared as systematic method to identify an optimum p-val to perform individual filtering in a sample-specific manner and in a less arbitrary way than the extended 1% FDR.

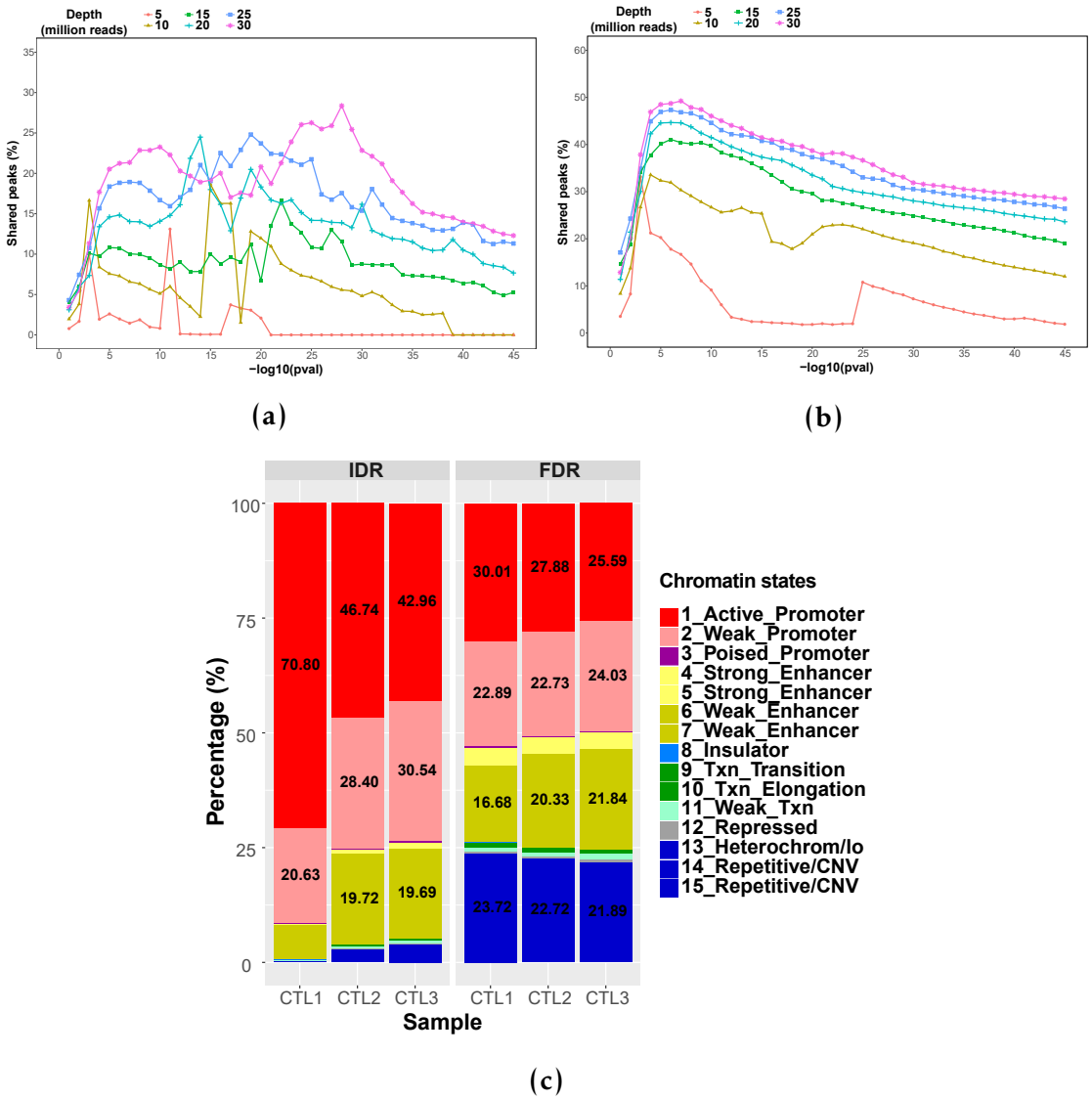


Figure 1.3: Peak calling filtering using IDR analysis in ATAC-seq samples

Differential chromatin accessibility analysis

From the methods that can be used to perform differential chromatin accessibility analysis (Table 1.1), I chose a peak-based approach where a consensus master list between all samples was built and the number of reads overlapping the master list peaks were retrieved for each sample. As previously

mentioned in the Chapter the master list was composed of non-overlapping 500bp with peaks present in at least 30% of the samples, regardless the group they belonged to (e.g patients or controls). One of the main limitations of the ATAC-seq and FAST-ATAC protocols (discussed in the next section) is the background signal. Therefore, it was calculation of an empirical cut-off, similarly to the strategy use in micro-array technology, was performed to minimise the impact of background read counts on the differential analysis (**Xinmin2005**; Jonker et al. 2014). Moreover, due to the lack of consistency found across the ATAC-seq publications, two methods for normalisation/differential analysis were assayed.

From the count matrix of the same six samples as before, the combined distribution of read density from all the absent peaks in each sample was used to define a sequence of twenty cut-offs. These cut-offs corresponded to the number of counts showed by a particular percentage of absent peaks (supplementary info). Each cut-off was used to filter out from the raw count matrix those peaks from the master list for which the number of counts was <= than that particular cut-off in more than three samples (being three the number of the smallest group of replicates in this particular experimental design). Quantile normalisation followed by differential analysis with limma voom showed greater number of differential open chromatin regions (DOCs) at an FDR<0.01 compared to DESeq2 across all the cut-offs (Figure 1.4 a). The two approaches presented progressive decrease in the number of DOC sites from the 75% cut-off. Conversely, the proportion of DOC calculated over the total number of regions considered in the differential analysis for each cut off significantly increased from the 50% cut-off onwards, indicating a progressive reduction in the false positive hits reported 1.4 b). From this analysis, 80% was chosen as a conservative filtering cut-off for which almost all the 19,855 DOCs identified by the most conservative method

(DESeq2) at an FDR<0.01 were recapitulated by limma voom at the same FDR (Figure 1.4 c).

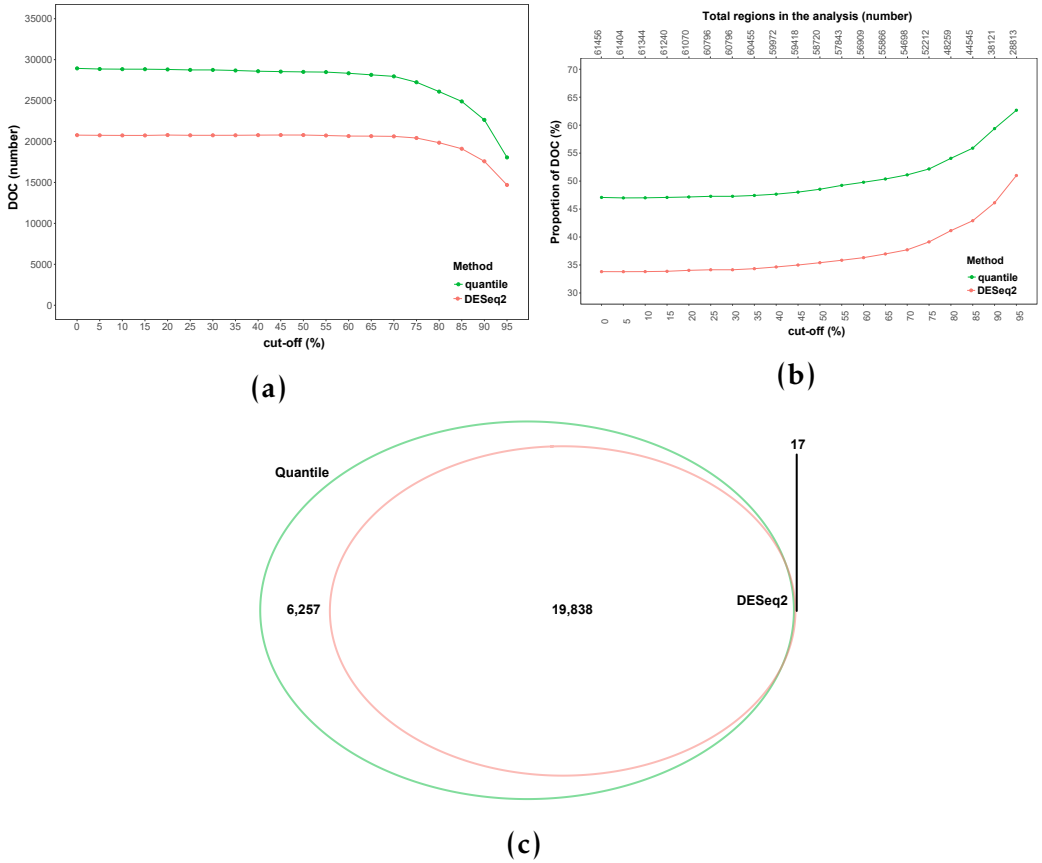


Figure 1.4: Differential chromatin accessibility analysis using limma voom and DESeq2.

Both methods performed appropriate normalisation of the counts at each of the master list peaks across the six samples, being the median of the quantile normalisation slightly more consistent across the two cell types compared to DESeq2 (Figure 1.5 a). When looking at the first FDR ranked 19,855 limma voom DOCs, 18,768 of them were the same as the retrieved by DESeq2. Moreover, very significant positive correlation was found between the fold changed of those 18,768 significant DOCs in both differential analysis methods (r^2=0.999, p-val=2.2^{-}16) (Figure 1.5 b). These observations suggested that the differences in the number of FDR significant DOCs reported by each of the methods could be partly due to differences in the way of calculating the false-positive rate.

Clustering and heat map and pathway analysis-briefly

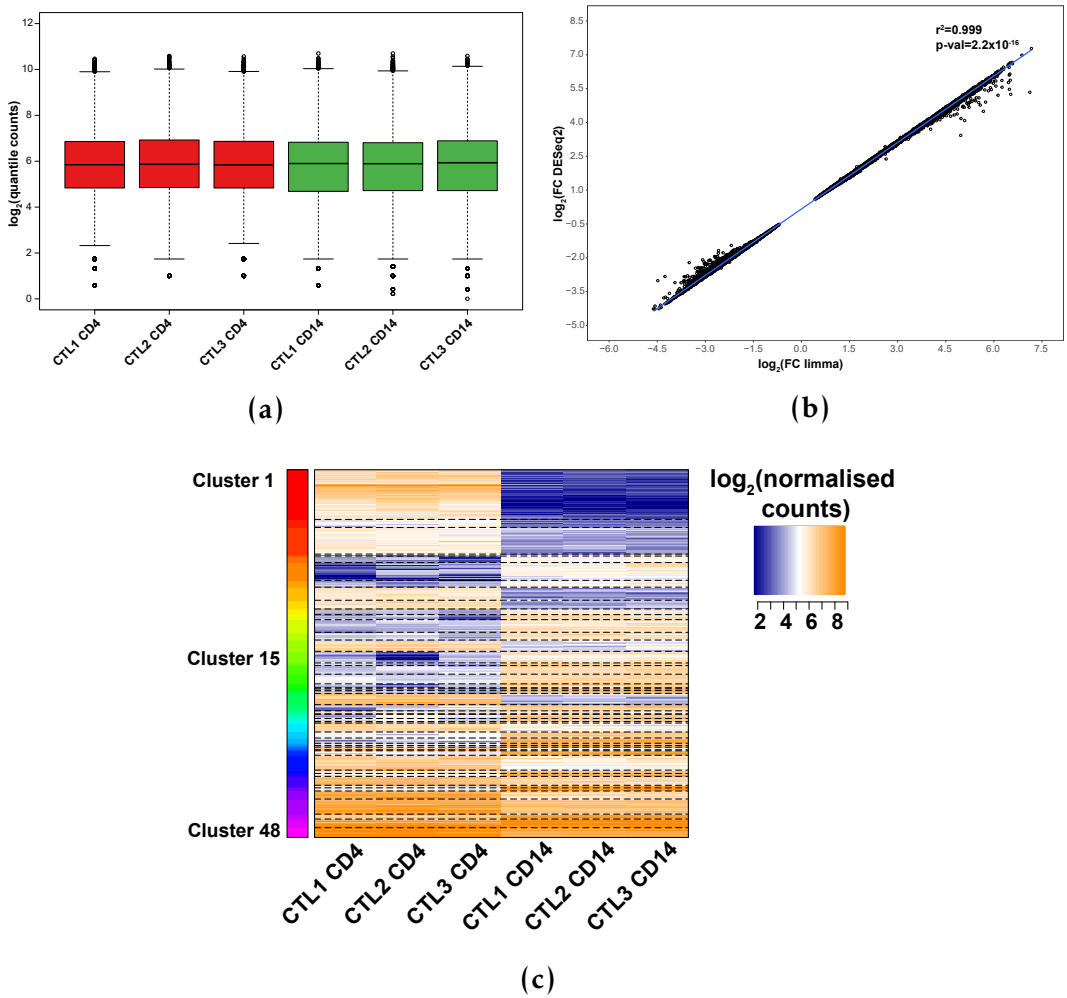


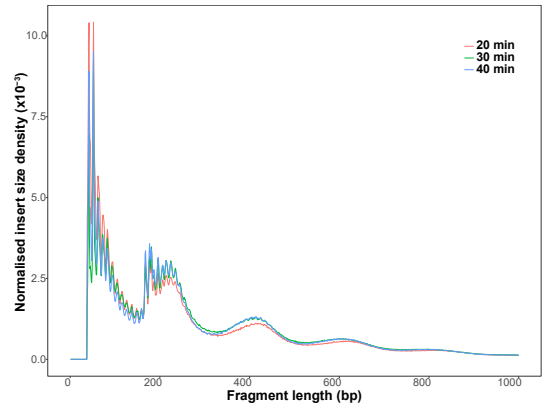
Figure 1.5: Exploration of the differential chromatin accessibility analysis using 80% as the empirical cut-off.

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

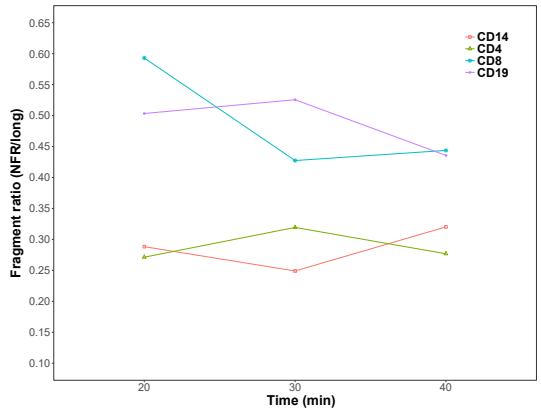
1.2.3 Impact of cryopreservation and fixation in the chromatin landscape of immune primary cells

Experimental design and cohort description

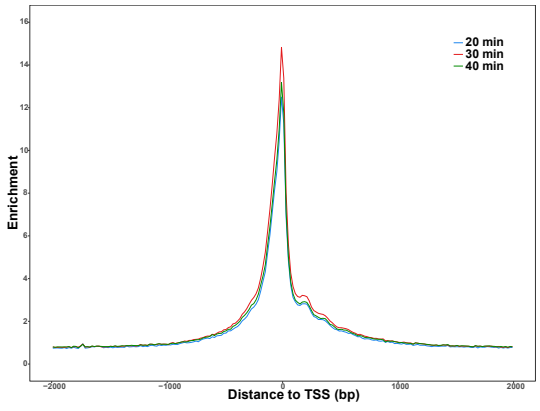
As previously introduced, research using clinical samples represents a logistic challenge precluding immediate sample processing due to geographical



(a)



(b)



(c)

Figure 1.6: Assessment of the effect of transposition times on the ATAC-seq QC parameters

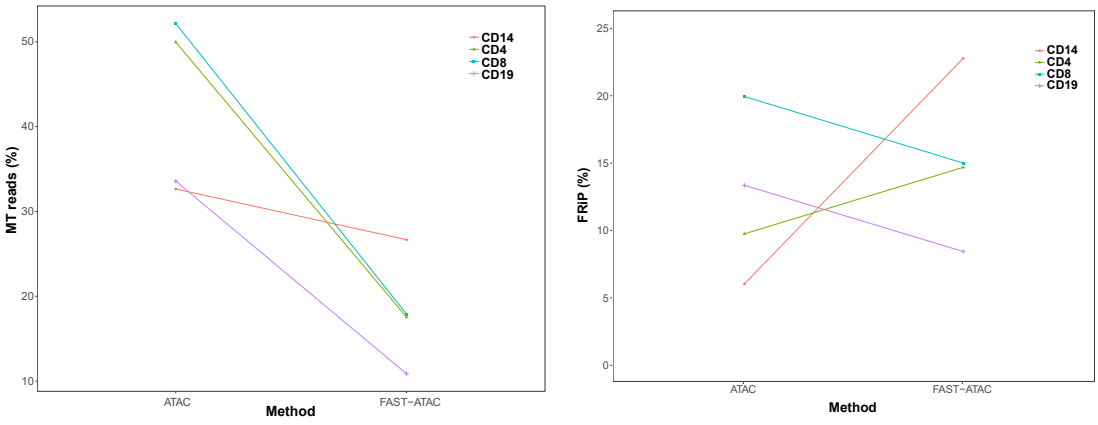


Figure 1.7: Differences in MT DNA abundance and signal specificity between ATAC-seq and FAST-ATAC protocols

and/or temporal limitations. Therefore long-term storage procedures and the use of preservatives such as fixatives facilitate handling of clinical samples. In the context of these thesis two different approaches were of interest and a collaborative project was established with High-Throughput Genomics at the Wellcome Centre for Human Genomics. On one side, the cryopreservation of PBMCs in liquid nitrogen using DMSO would allow long-term preservation and would require PBMCs thawing and recovery followed by FACS isolation of the cell population of interest. On the other hand, the use of a fixative on the FACS-isolated relevant cell types from the clinical sample could be used as a short term preservation method. Regarding fixative, there was an interest to test the performance of an optimised protocol developed by High-Throughput Genomics using DSP in scRNA-seq (). DSP is a cell-permeable and reversible cross-linking fixative that have previously been used to preserve tissue samples for downstream immuno-staining, laser micro-dissection and RNA micro-array expression profiling ().

In order to investigate the effect of the cryopreservation and DSP fixation in the nuclear integrity and chromatin structure, three healthy volunteers sex and age matched were processed on different days to be consistent with the case-

control experimental design in psoriasis and PsA (Figure **??**). For each of the individuals PBMCs were isolated from blood and a fraction was stained with the appropriate panel of Abs to isolate $CD14^+$ monocytes and $CD4^+$ T cells, as detailed in Chapter . ATAC-seq was performed directly in an aliquot of the isolated $CD14^+$ and $CD4^+$ cell. A fraction of the isolated $CD14^+$ and $CD4^+$ cell were appropriately fixed with DPS, stored at 4°C for 24h and processed for ATAC-seq. A fraction of the PBMCs were cryopreserved following the protocol described in Chapter , stored in liquid nitrogen for approximately 1 week followed by thawing and recovery in culture for approximately 30 min before undergoing Ab staining and FACS to retrieve $CD14^+$ and $CD4^+$ cells. Altogether, for each of the three controls three matched ATAC-seq samples were generated: ATAC-seq fresh, ATAC-seq fixed and ATAC-seq frozen.

Figure 1.8: Experimetal design to assess the impact of cryopreservation and fixation in the chromatin accessibility of immune primary cells.

Quality control of ATAC-seq data

Biological relevance of differentially open chromatin across conditions

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

Due to the fact that KC is one of the most relevant cell types in psoriasis pathophysiology, ATAC-seq as described in Buenrostro *et al.*, 2013 (named as ATAC-seq 1 here) was performed in 50,000 cells of a suspensions isolated from a psoriasis lesional skin biopsy. Two different tranposition times (30 and 40 min) where tested. Since biopsy handling and lesional epidermal KC 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. Two tranposition times (30 and 40 min) where tested.

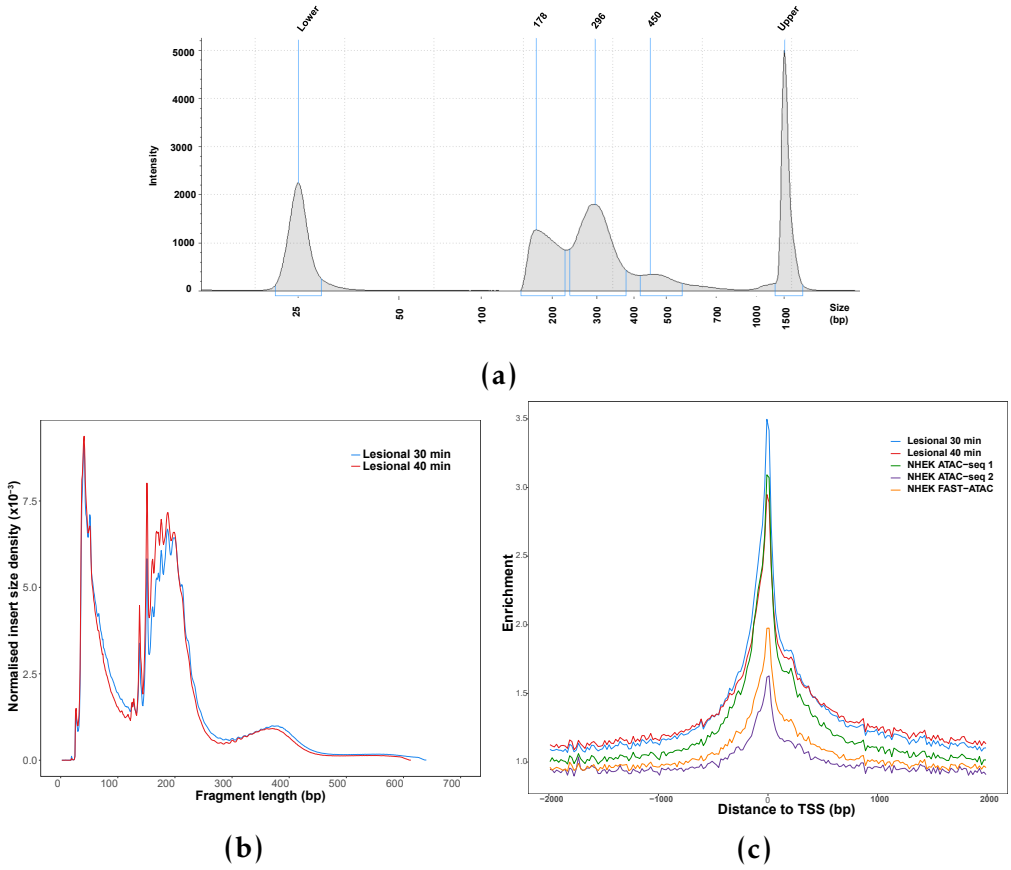


Figure 1.9: QC assessment of ATAC-seq in KC enriched cell suspension derived from a psoriatic lesional skin biopsy. Two transposition times (30 and 40 min) were tested using the standard ATAC-seq protocol (Buenrostro *et al.*, 2013 in 50,000 cells from the same suspension.

Although cell suspension obtained from biopsies using trypsinisation of the epidermal sheet are 90% enriched in KC, they also contain significant amounts of dead cells and free-DNA releases by apoptotic cells. In order to overcome this problem and the impact that it may have over ATAC-seq background signal, viable KC were selected by adherence assay. Biopsy cell suspensions were cultured for 3h in a 96-well plate and washed afterwards to ensure that only the viable and less differentiated KC would remain for down stream analysis. In parallel cultured NHEK were also used to assess the performance of the different ATAC-seq protocols.

Table for the conditions: done Tapestation profiles of the the chosen condition. done Send the others to supplementary. QC measurements: for ATAC1, ATAC2 and NHEK, mention frag size distribution done DHS enrichment for p and q done but not convincing.The complex network of keratin filaments in stratified epithelia is tightly regulated during squamous cell differentiation. Keratin 14 (K14) is expressed in mitotically active basal layer cells, along with its partner keratin 5(K5), and their expression is down-regulated as cells differentiate.

Protocol	Lysis and transposition	Key parameters
Buenrostro et al., 2013	Two steps	0.1% NP-40 and 2.5µL Tn5
Bao et al., 2015	Two steps	0.05% NP-40 and 5µL Tn5
Corces et al., 2016	One step	C1: 0.01% digitonin, 0.5µL Tn5 C2: 0.01% digitonin, 2.5µL 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 for all the different protocols was 30 min.

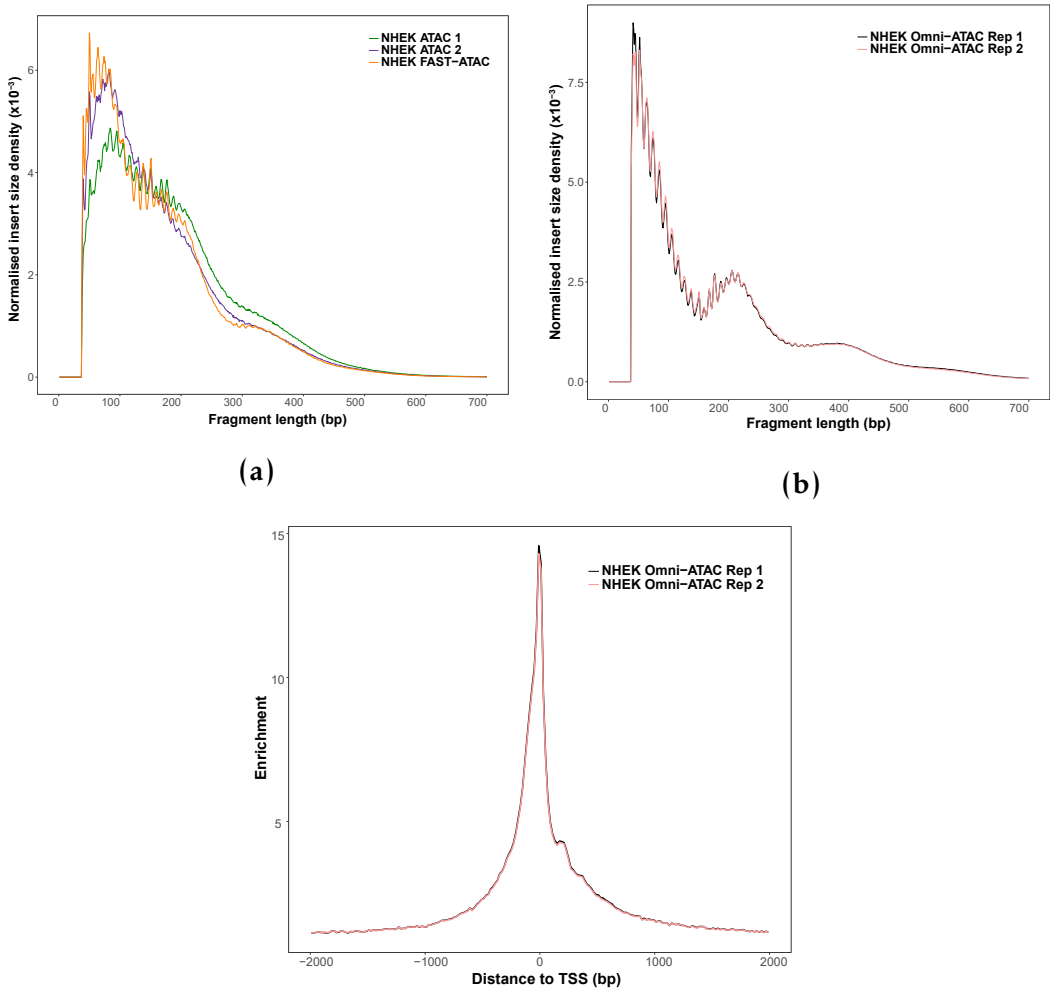


Figure 1.10: QC assessment of FAST-ATAC and Omni-ATAC in cultured NHEK.

Omni-ATAC Tapestation profiles of the the chosen condition include it with the supplementary that includes all other tapestation profiles.done QC measurements: frag size distribution and TSS done Track including all skin samples

Think of what to include about the biopsies in supplementary done

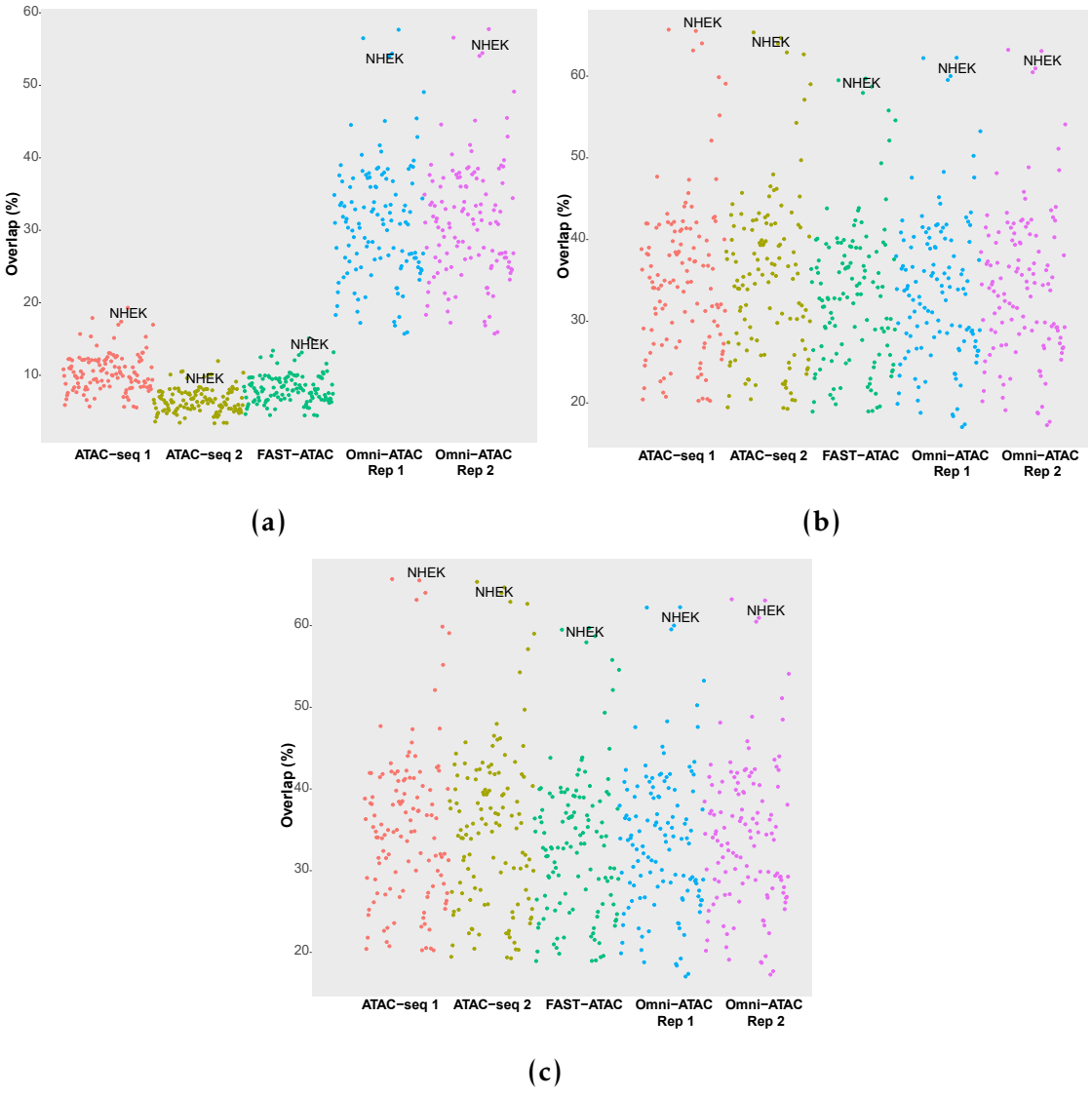


Figure 1.11: QC assessment of Omni-ATAC in NHEK and chromatin accessibility signal for the samples generated with the different ATAC-seq protocols.

1.2.5 Discussion

Maybe justify in the dicussion the use of DESEq2 and limma shared based on Alasoo observation of noise effect in limma

Chapter 2

Cross-tissue comparison of chromatin accessibility, gene expression signature and immunophenotypes in PsA

2.1 Introduction

The techniques to generate the scRNA-seq data have also evolved and SmartSeq2 and 10X Chromium are the two main

2.2 Results

2.2.1 PsA patients cohort description and datasets

For this study blood and SF samples were collected from six PsA patients, with equal number of male and female (Table 2.1). All the patients presented oligoarticular joint affection and had been first diagnosed with psoriasis. Maybe add sth about oligoarticular?

The cohort presented a mean of 1.5 tender or swollen affected joints (TJC66 and SJC66), which is characteristic of the oligoarticular form of disease, and joint pain of xxxx. Regarding global assessment, the mean scores for the patient and physician evaluation were X and 3, respectively, in a scale of 1 to 5. These four measurements including joints and global assessment compose the PsARC

disease activity scores, used by clinicians as the main indicator of response to treatment by recommendation of the xxxx, as previously explained in Chapter **??**.

The mean age of the cohort at the time of diagnosis was 44.3 years old and the mean disease duration 8.8 years. Interestingly, PsA1728 was diagnosed at a later age compared to the other patients in the cohort (late PsA onset clinical significance??). Moreover, CRP levels, other marker of inflammation, was also measured in all the patients presenting an average of 17.45 mg/L and being particularly higher in PSA1719 and PSA1728, compared to the other patients. At the time of sample recruitment all the PsA patients were naive for treatment and only PSA1505 had been on methotrexate therapy in the past for xxx months/years (how many years ago?). Post-visit, most of the patients qualified for TNAi biologic therapy xxxx.

Table 2.1: Description of PsA patients cohort recruitment and metadata.. PsARC disease activity score is composed of tender joint count 66 (TJC66) and swollen joint count 66 (SJC66), joint pain (4 point score) and self-patient and physician global assessment (5 point score). Joint pain and global assessment use a likert scale based on questionnaire answers that measure the level of agreement with each of statements included. C-reactive protein (CRP).

Sample ID	Sex	Age diagnosis	Disease duration (months)	Type	TJC66/SJC66	Physician assessment	CRP (mg/L)
PSA1718	Female	17	180	Oligo	2/2	3	6
PSA1719	Male	33	24	Oligo	1/1	3	36.6
PSA1607	Male	42	108	Oligo	1/1	4	8
PSA1728	Female	72	48	Oligo	2/2	3	43.2
PsA1801	Female	53	168	Oligo	2/2	3	9.9
PsA1505	Male	35	108	Oligo	1/1	2	1
Total	–	44.3	106	–	1.5/1.5	3	17.45

For each of the patients, paired data in blood and SF was generated from bulk mononuclear cells or the isolated cell types of interest (detailed in Table 2.2 and Chapter 1.2.3). However, not all types of data including ATAC-seq, PCR gene expression array, scRNA-seq and mass cytometry were generated for all six individuals of the cohort due to project constrains.

Sample ID	% FAST-ATAC	RNA PCR array	scRNA-seq	mass cytometry
PSA1718	Yes	Yes	No	Yes
PSA1719	Yes	Yes	No	Yes
PSA1607	Yes	No	Yes	Yes
PSA1728	No	Yes	No	Yes
PSA1801	No	No	Yes	Yes
PSA1605	No	No	Yes	Yes

Table 2.2: Datasets generated for the PsA cohort samples. Four types of data were generated in a paired way between blood and SF from the same individual. The types of data available varies between individuals due to project constrains. FAST-ATAC data was generated for $CD14^+$, $mCD4^+$,$mCD8^+$ and NK cells. RNA PCR array was performed in $CD14^+$, $mCD4^+$ and $mCD8^+$. scRNA-seq data was generated using 10X technology in bulk PBMCs, bulk SFMCs and sorted $mCD4^+$ and $mCD8^+$ from both tissues.

2.2.2 The chromatin accessibility landscape in SF and PB immune cells

Quality control of open chromatin regions

The twenty four PsA samples form four cell types and two different tissues (PB and SF) were sequenced to a median of 158M reads (79M paired-end) per sample. After filtering for low quality mapping, duplicates and MT reads, the median total number of reads were 70.2M, 50.6M, 46.6 and 66.7M for $CD14^+$, $mCD4^+$, $mCD8^+$ and NK cells, respectively (Figure 2.1 a). The differences between cell types and samples in the median of total reads remaining after filtering was inversely related to the percentage MT and duplicated reads identified (Figure 2.1 b). For example, $mCD1^+$ and $mCD8^+$, presented the lower

median of total number of reads after filtering concomitantly with the greater percentage of MT and duplicated reads. In combination, MT and duplicated reads accounted for a median of 42, 57.6, 62.2 and 40% in $CD14^+$, $mCD4^+$, $mCD8^+$ and NK cells, respectively, importantly contributing to the loss of reads in this experiment. As previously mentioned, the MT DNA in ATAC-seq is one of the main sources of read loss, which is more accessible to the Tn5transposase due to the absence of nucleosomes. Although the FAST-ATAC protocol represented an improvement, the percentage of MT reads across amongst all the samples ranged between 2.1 and 25.4%. Similarly, despite initial optimisation of the number of PCR cycles used in the library amplification, the duplicated reads still represented between 22.9 to 55% of the total number of the pre-filtered reads.

Regarding sample quality determination, TSS enrichment analysis showed variation in the levels of background noise across cell types and highlighted the variability in performance of FAST-ATAC (Figure 2.1 c). A trend towards greater TSS enrichment in PB samples compared to SF can be observed in all four cell types. In terms of cell types, $mCD4^+$ and $mCD8^+$ presented the best signal-to-noise ratios, with median of 19.1 and 23.1 fold enrichment, respectively. In contrast, NK was the cell type with the lower TSS enrichment values. Particularly, the fold enrichment for PSA1719 and PSA1607 were 7.3 and 6.2, respectively, both just above the 6 to 10 acceptable range from ENCODE. If a bigger sample size was available it would be appropriate to drop these samples from the differential analysis, since high background levels will reduce the power of this approach.

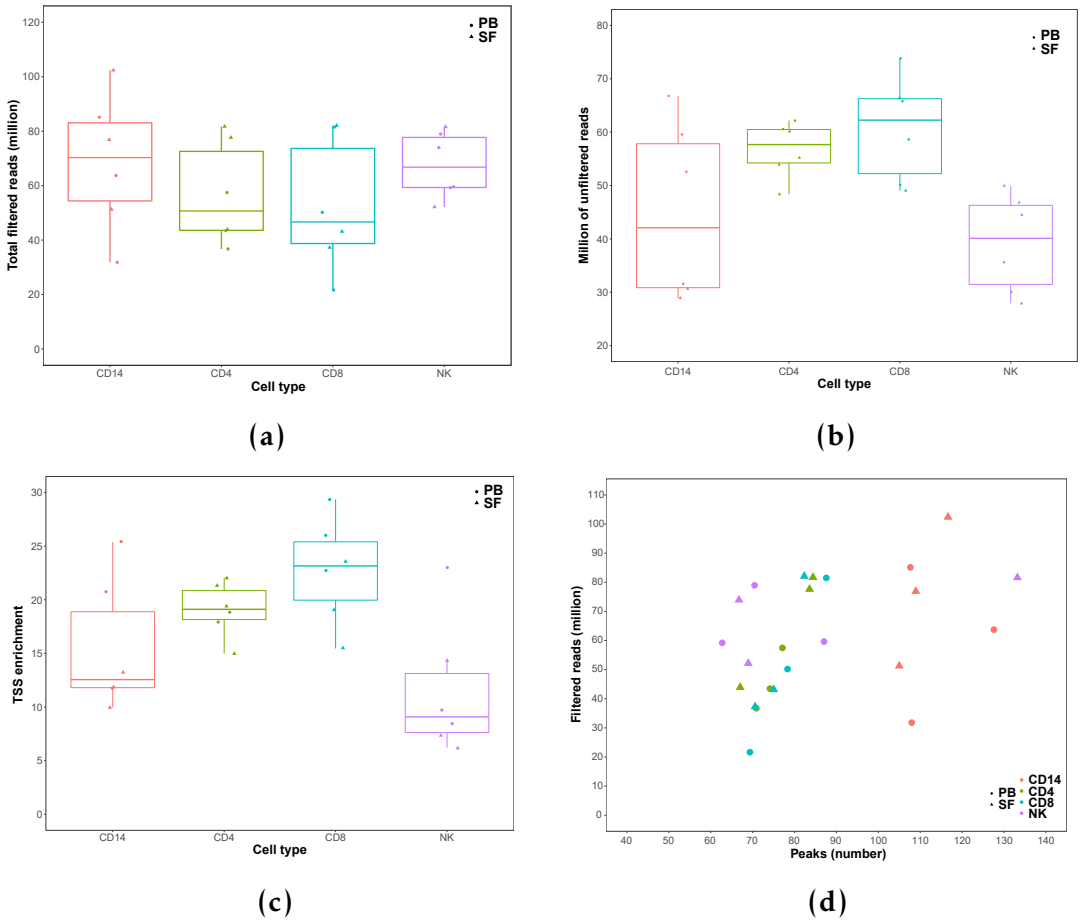


Figure 2.1: QC of FAST-ATAC PsA samples in four cell types.

When identifying open chromatin regions through peak calling and standard filtering for FDR<0.01 (not the IDR sample-specific filtering), the number of peaks ranged from $\sim 62x10^3$ to $\sim 133x10^3$ peaks per sample (Figure 2.1 d). A clear positive correlation between the number of called peaks and number of reads after filtering could be observed in the data. For example, $CD14^+$ was the cell type with greatest number of called peaks ($108.4x10^3$) as well as the greater median of reads remaining after filtering when compared to the other three cell types (Figure 2.1 a). For the NK, the two samples with the greatest TSS enrichment (PSA1718 SF and PB) showed greater number of called peaks when compared to the other NK samples with similar number of reads. This observation was consistent with the correlation between sample quality and the number of identified accessible chromatin regions previously demonstrated in

Chapter 1. Overall, appropriate number of peaks were called in all the samples and no concerning outliers were identified.

Open chromatin reflects cell type specificity and functional relevance

In order to determine the ability of the open chromatin identified by the in house pipeline in the PsA sample cohort, a combined master list including all four cell types and the two tissues was built. Following Chapter 1.2.3 and Chapter 1, the combined master list contained open chromatin regions identified in at least 30% of the samples (in this case 7 samples) regardless cell type and tissue to avoid any bias.

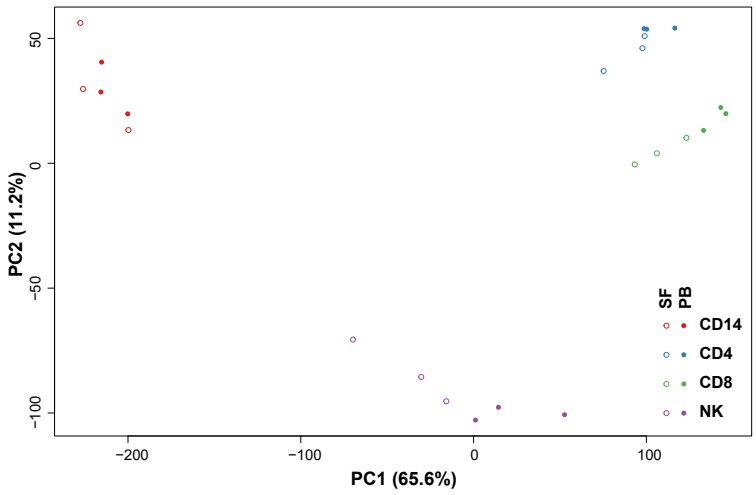


Figure 2.2: Combined PCA analysis of all four cell types isolated from blood and SF.

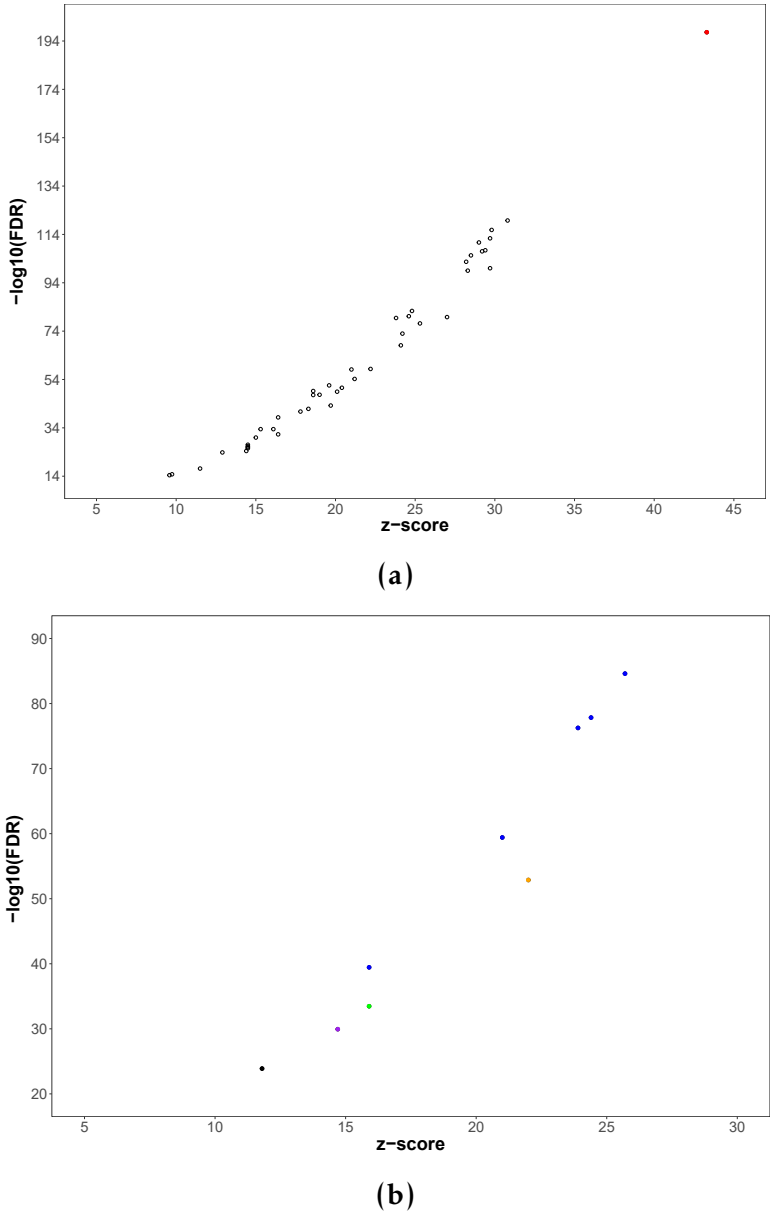


Figure 2.3: Enrichment of eQTLs in the combined cell types PsA accessible chromatin master list xxxx

Differential open chromatin analysis between blood and SF

Differential chromatin accessibility analysis was performed using a paired design between SF and PB for each of the four cell types (Table 2.4. For each of the cell types a master list containing chromatin accessible regions in at leats 30% of the samples (~2 samples), regardless the tissue. In all for analysis an 80% cut-off for background noise was applied in the count matrix as previously

explained in Chapter 3. Only DOCs identified with DESeq2 and also shared with quantile normalisation limma voom analysis where used downstream. The $CD14^+$ monocytes and NK showed a greater proportion of differentially accessible regions (23.3 and 8.9%, respectively) compared to $mCD4^+$ and $mCD8^+$ T cells. In $CD14^+$ monocytes 3,779 out of 5,285 DOCs were more accessible in SF versus the 1,506 more open in PB isolated $CD14^+$. Conversely, the number of DOCs more accessible in each of the tissues were evenly distributed between SF and PB in the remaining three cell types.

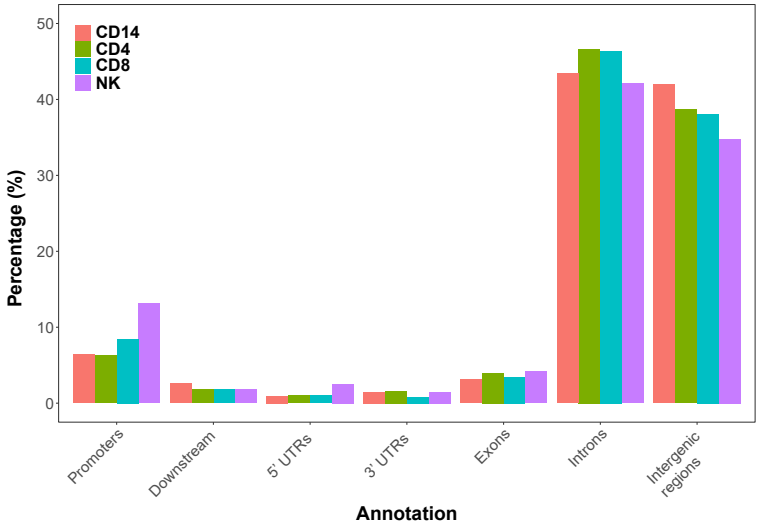
Cell type	Total DOCs	Proportion DOCS (%)	DOCs open in SF	DOCs open in PB
$CD14^+$	5,285	23.3	3,779	1,506
$CD4^+$	1,329	4.3	621	708
$CD8^+$	1,570	4.5	807	763
NK	2,314	8.9	1,223	1,091

Table 2.3: Summary results of the chromatin accessibility analysis between SF and PB in PsA samples.

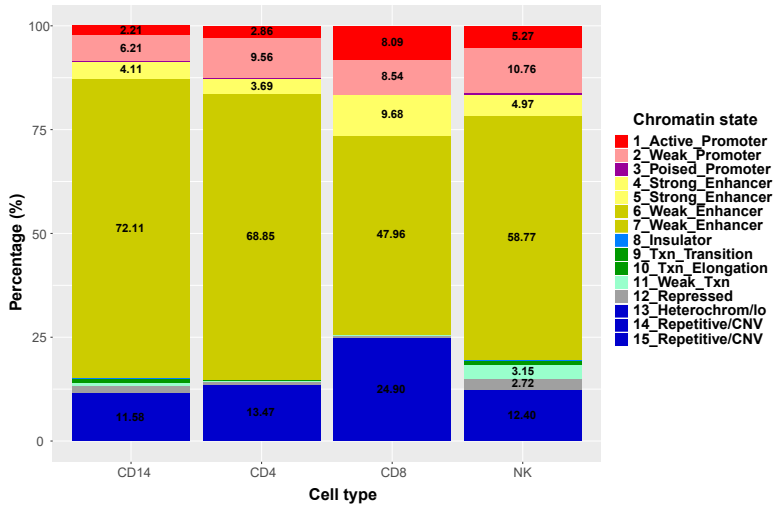
Permutation analysis using the ten unique possible combinations demonstrated that the number of DOCs obtained in the differential analysis for each of the cell types w than expected by chance (Figure B.1). This reinforces the specificity of the identified changes in chromatin accessibility, which are driven by true differences between SF and PB in all the analysed cell types.

When performing genomic annotation of the DOCs, intronic and intergenic regions represented together 80% or more of all the DOCs in the four cell types (Figure 2.4 a). DOCs annotated in universal promoter regions represented approximately between 5 to 15%, constituting the third most represented genomic feature. DOCs were also annotated with the fifteen cell type-specific chromatin states from the Epigenome Roadmap segmentation maps (Figure 2.4 b). For all four cell types between 44.96 and 72.11% of the DOCs where annotated as weak enhancers, which represented the most prominent category. This was consistent with the predominance of introns and intergenic regions

which is the preferred location for enhancers, and also highlighted the cell type specificity of the differences in open chromatin, since enhancers are more cell type specific than promoters.



(a)



(b)

Figure 2.4: Annotation with genomic regions and chromatin states of the PsA DOCs from the four cell types differential analysis. xxxx

Interestingly, from the DOCs located overlapping a gene body, the majority were located within introns instead of untranslated regions (UTRs)and have also been annotated as weak or strong enhancers according to the cell type specific chromatin segmentation map (Table 2.4). For all four cell types, a number of gene

entities contained more than one DOC, showing the same direction of chromatin accessibility between SF and PB. For example, in $CD14^+$ two DOCs located at the 5' and 3'UTRs for IL7R gene where found to be more accessible in SF compared to PB. Similarly, more accessible chromatin in SF compared to PB was identified in five regions of the *IL15* gene, annotated as promoter and enhancers in $CD14^+$ monocytes.

Cell type	DOCs in gene body	Gene with > one DOC	Enhancers	Introns
$CD14^+$	2,357	744	1,775	1,920
$CD4^+$	700	99	504	577
$CD8^+$	831	118	503	666
NK	1,246	235	782	937

Table 2.4: Summary results of the chromatin accessibility analysis between SF and PB in PsA samples.

- Add the enrichment results from the XGR analysis for chromatin segments

The DOCs region from the four cell types presented enrichment for robust and permissive enhancers (Figure **??**). Robust enhancers are those for which transcription was significantly detected at the genome-wide level in one or more primary cell or tissue, whereas the permissive set included also those not passing the filtering criteria (**Andersson2014**). Moreover, amongst the top two most enriched cell type-specific set of eRNAs for each set of DOCs, all four cell types included the appropriate cell entity.

The relevance of the DOC regions identified through differential analysis was also addressed. Enrichment analysis of psoriasis and PsA GWAS hits for the differentially open regions between SF and PB in the four cell types was performed using XGR co-localisation and permutation analysis. Although no significant results were found at the SNP level (lead SNPs and SNPs in LD $r^2 \geq 8$), significant enrichment (2-fold enrichment and empirical p-val 0.043) was observed for psoriasis GWAS LD blocks for the $CD14^+$ DOCs only.

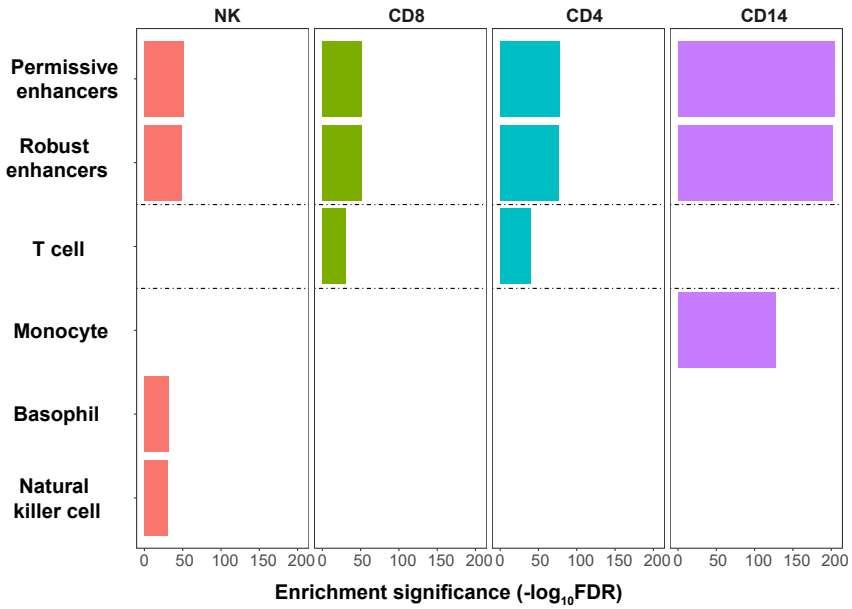


Figure 2.5: Enrichment of PsA DOCs for the FANTOM5 eRNA dataset.

2.2.3 Pathway and TFBS enrichment analysis highlight functional tissue-specific differences in chromatin accessibility

For each of the four sets of DOCs identified by the differential analysis, pathway enrichment analysis was conducted separately for SF and PB open regions. Gene annotation of the DOCs was performed by physical proximity, as detailed in Chapter 1.2.3. Despite commonalities, differences in significant enriched pathways (FDR< 0.01) were identified between SF and PB within the same cell type (Table **??**)

Table 2.5: Distinct enriched pathways in CD14$^+$, mCD4$^+$, mCD8$^+$ and NK between SF and PB. All pathways shown have an FDR <0.01.

Cell type	SF	PB
CD14$^+$	Hemostasis, Platelet activation, Signaling by VEGF GPCR ligand binding, IL-2 signaling pathway, Integrin cell surface interactions,NF-kappa B signaling pathway IL-2 family signaling, IL-3, 5 and GM-CSF signaling.	DAP12 interactions, Metabolism of lipids, Metabolism of vitamins and co-factors, Negative regulation of the PI3K/AKT network.
mCD4$^+$	T cell receptor signaling pathway, Phospholipase D signaling pathway , Chemokine signaling pathway, PI3K-Akt signaling pathway, Signaling by interleukins.	Signaling by Receptor Tyrosine Kinases, Focal adhesion.
mCD8$^+$	Chemokine signaling pathway, Signaling by GPCR Signaling pathways regulating pluripotency of stem cells Regulation of actin cytoskeleton.	Signal transduction, Wnt signaling pathway, Rho GTPase cycle, PI3K-Akt signaling pathway, Signaling by interleukins.
NK$^+$	Extracellular matrix organization, Rap1 signaling pathway, Calcium signaling pathway, PI3K-Akt signaling pathway, Fc gamma R-mediated phagocytosis, HIF-1 signaling pathway.	Th1 and Th2 cell differentiation, Rho GTPase cycle, T cell receptor signaling pathway, Signal Transduction, Natural killer cell mediated cytotoxicity,

MAPK family signaling cascades.

-pathway enrichment analysis if possible per open chromatin in each cell type -maybe include an A2 pathway which is different and unique between open in SF and PB in one cell type -TFBS

2.2.4 Differential gene expression analysis in paired circulating and synovial immune cells

Array data

2.3 Discussion

fGWAS analysis as Matthias did would be of interest but needs appropriate GWAS data I am going to try using XGR to do some of this

Appendices

A Establishment of methods to assess genome-wide chromatin accessibility 39

B Cross-tissue comparison analysis in PsA 42

List of Figures

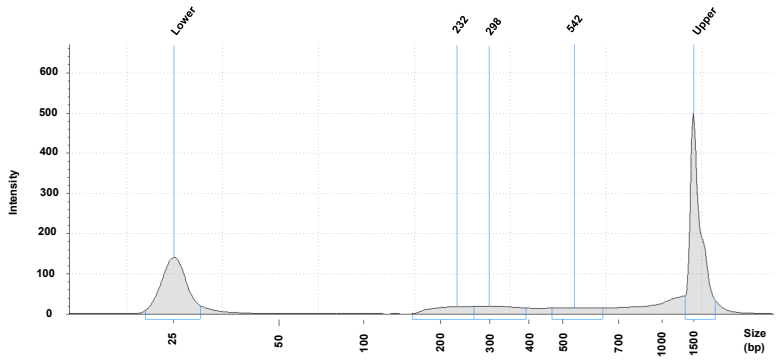
A.1 FAST-ATAC and Omni-ATAC NHEK tapestation profiles. 40
A.2 Assessment of TSS enrichment from ATAC-seq and FAST-ATAC in healthy and psoriasis skin biopsies samples. 41

B.1 Permutation analysis SF vs PB in $CD14^+$,$CD4m^+$,$CD8m^+$ and NK. 42

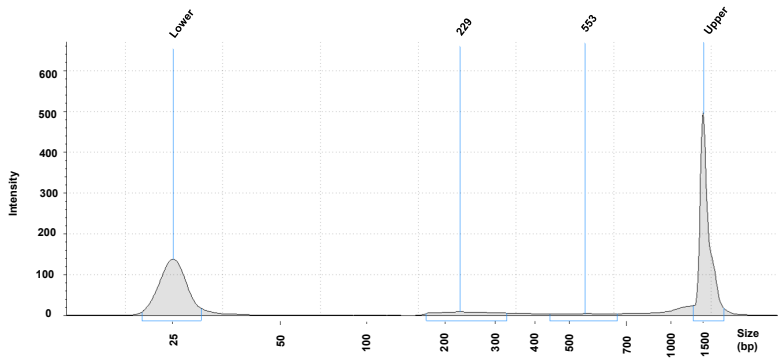
List of Tables

Appendix A

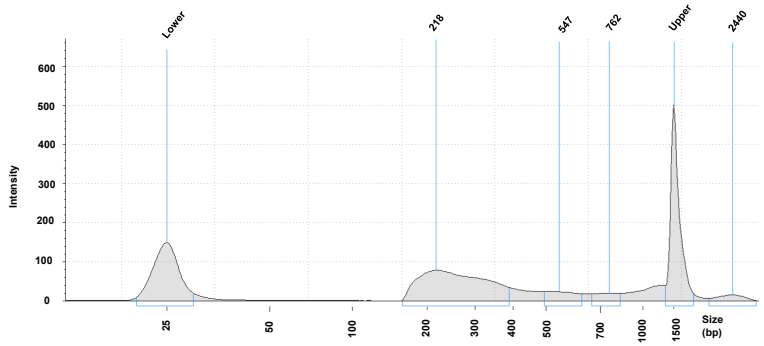
Establishment of methods to assess genome-wide chromatin accessibility



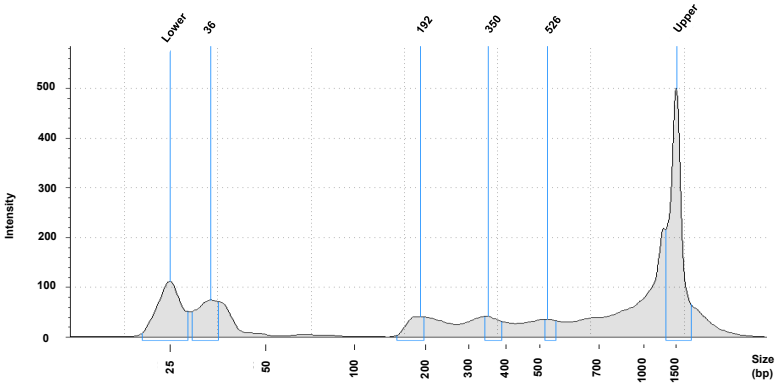
(a)



(b)



(c)



(d)

Figure A.1: FAST-ATAC and Omni-ATAC NHEK tapestation profiles.

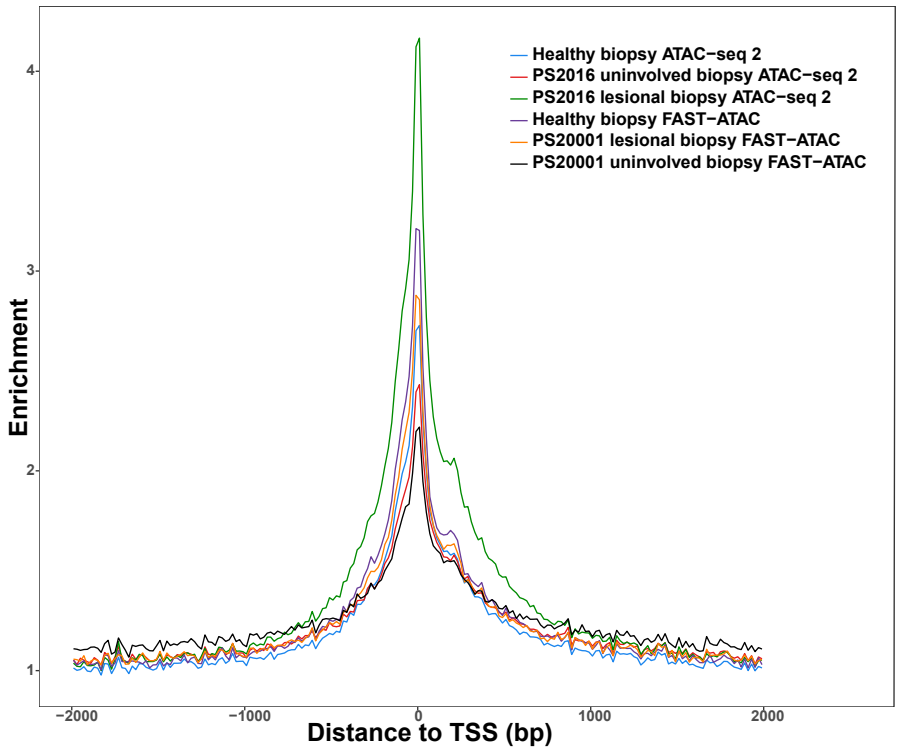


Figure A.2: Assessment of TSS enrichment from ATAC-seq and FAST-ATAC in healthy and psoriasis skin biopsies samples.

Appendix B

Cross-tissue comparison analysis in PsA

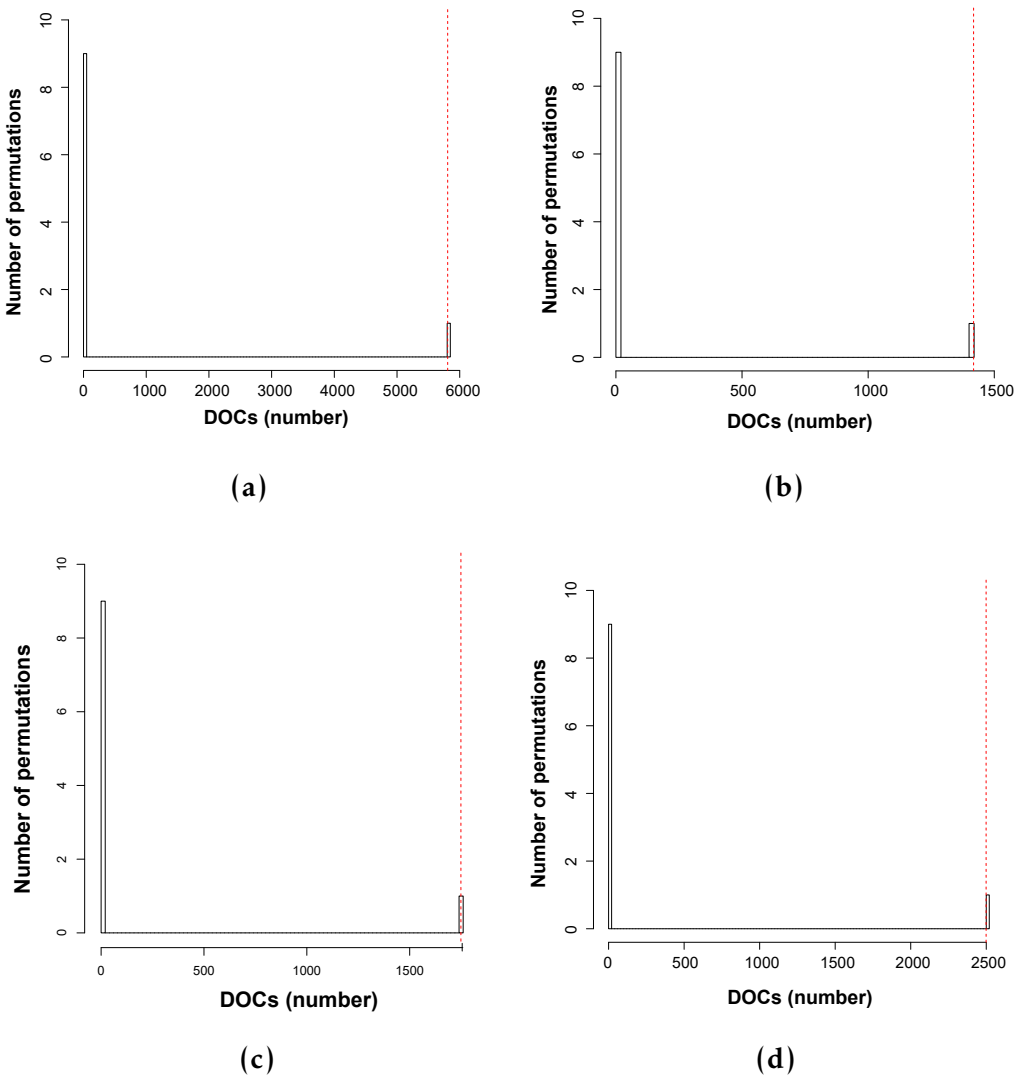


Figure B.1: Permutation analysis SF vs PB in $CD14^+$,$CD4m^+$,$CD8m^+$ and NK.