Functional genomics of psoriasis

Alicia Lledo Lara
Hertford College
University of Oxford

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

Abstract

Functional genomics of psoriasis

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

This is my abstract...

Acknowledgements

Thank you, thank you, thank you.

Declarations

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

Submitted Abstracts

Title **Year**
Authors

Associated Publications

Title
Journal
Authors

Other Publications

Title
Journal
Authors

Contents

List of Figures

List of Tables

Abbreviations

Abbreviation	Definition
Ab	Antibody
ATAC-seq	
Atopic dermatitis	AD
ChIPm	
CLE	cutaneous lupus erythematosus
DMARDs	disease-modifying antirheumatic drugs
Fast-ATAC	
IDR	
GWAS	Genome-wide association studies
KC	Keratinocytes
NSAID	nonsteroidal antiinflammatory drug
Omni-ATAC	
PCA	
PI	Protein inhibitor
PsA	
QC	
qPCR	quantitative polymerase chain reaction
RA	Rheumatoid arthritis
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: Summary table of ATAC-seq methodology analysis for peak calling, filtering and differential analysis..

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

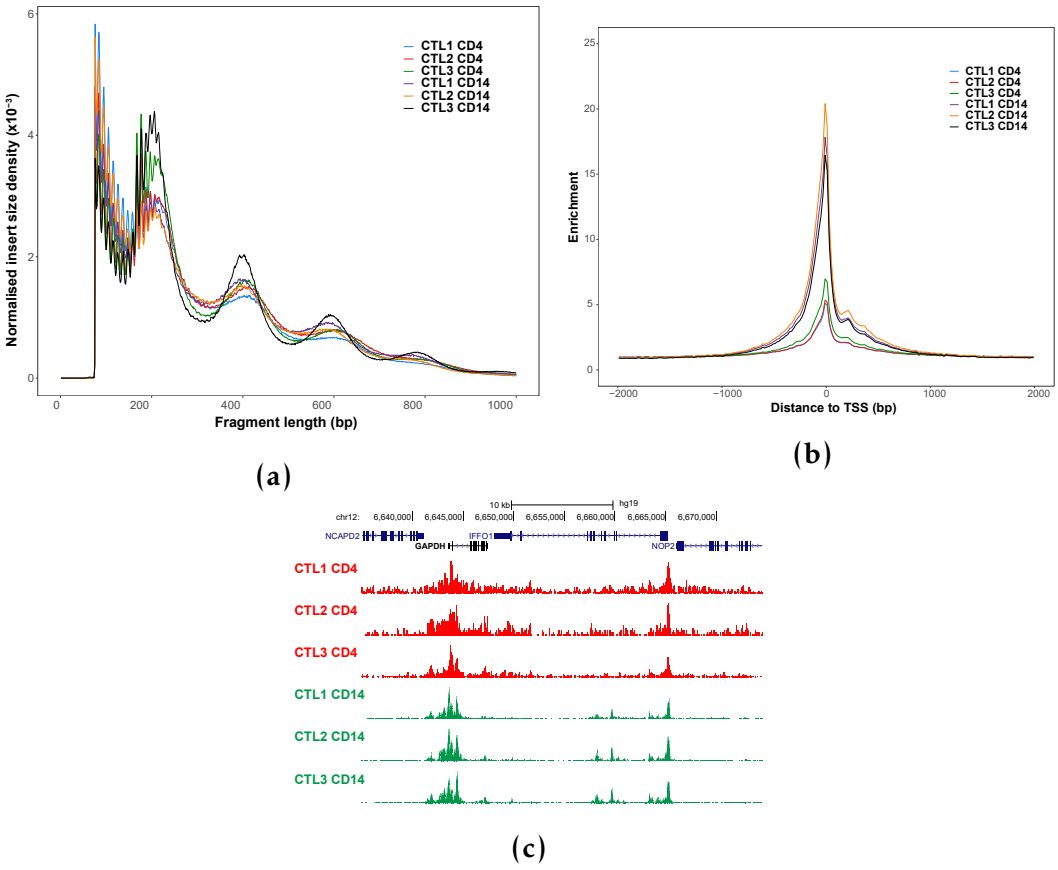


Figure 1.1: Measurements for quality control assessment in ATAC-seq samples

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

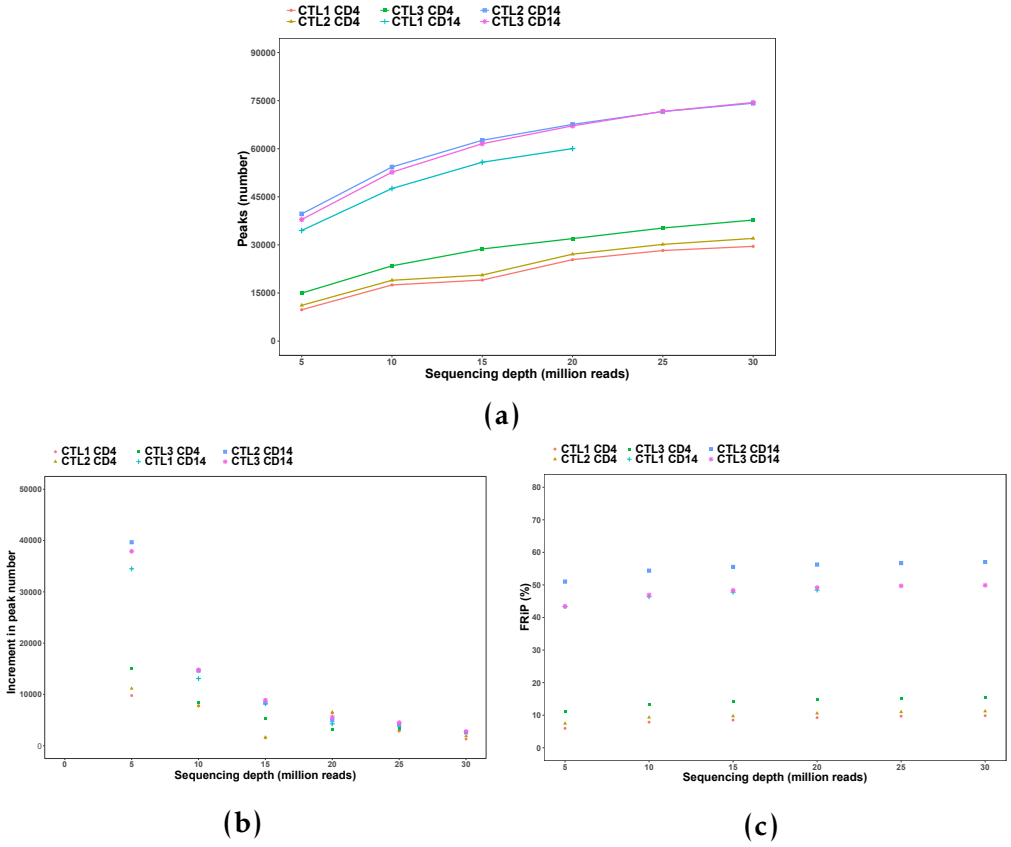


Figure 1.2: Peak calling at different sequencing depth in ATAC-seq samples

number shared peaks) sharing IDR at a particular p-val was calculated (Figure **??** 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 CD4fig:Peak$_{c}alling_{I}DR_{f}iltering_{a}nd_{c}hrom_{s}tated_{A}TACsample(CTL2CD4^{+}$ Figure **??** a) compared to the counterpart CD14^{+} (Figure **??** 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 **??** 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.

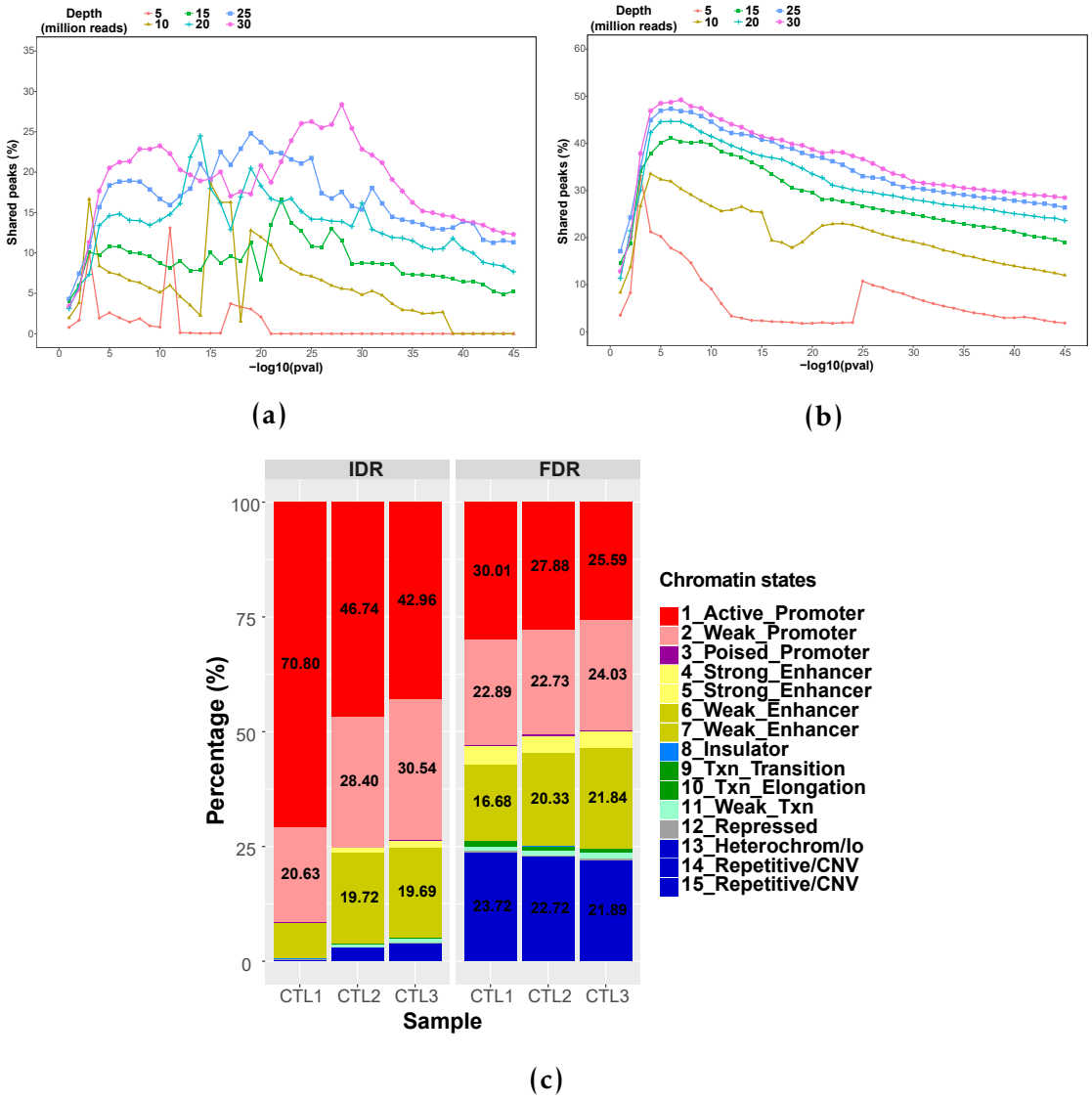


Figure 1.3: Peak calling filtering using IDR analysis in ATAC-seq samples

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, three methods for normalisation and differential analysis were assayed, due to the lack of consistency found across the ATAC-seq publications performing the same type of analysis.

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. Each cut-off corresponded to the number of counts showed by a percentage of absent peaks (supplementary info). Each of the cut-offs 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 3 samples (being 3 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 of false positive hits 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).

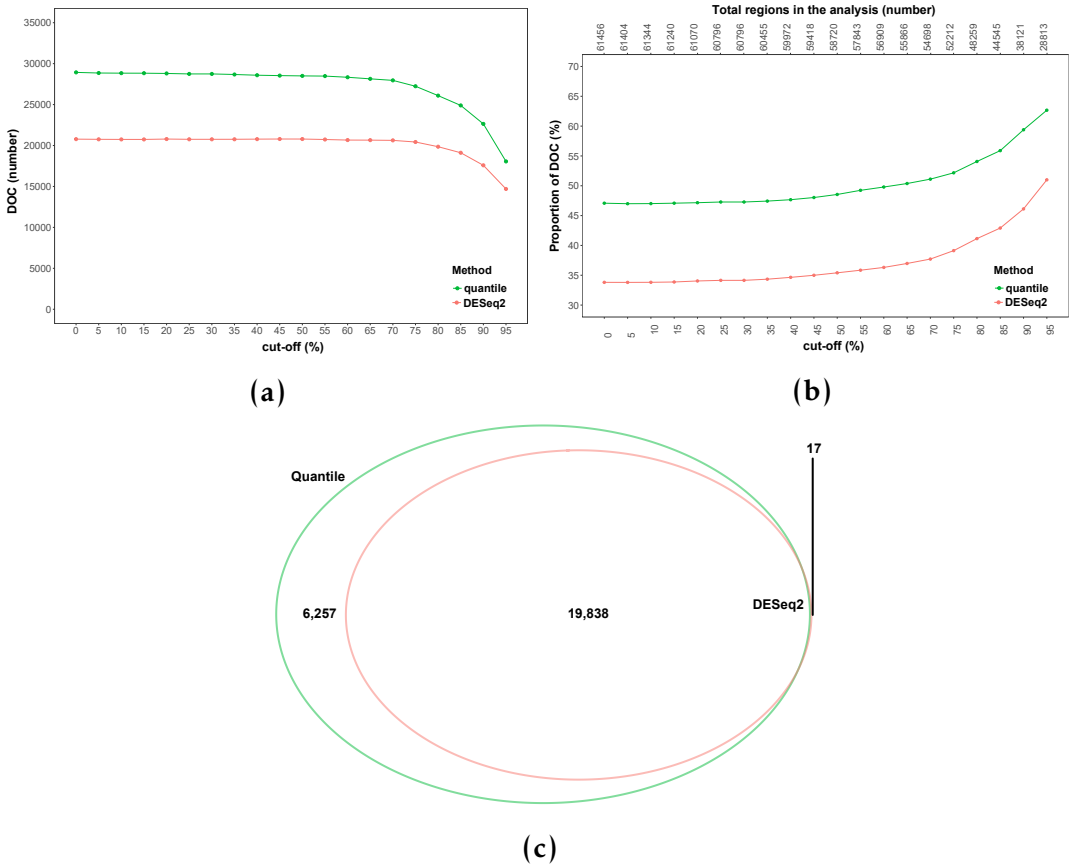


Figure 1.4: Differential chromatin accessibility analysis using limma voom and DESeq2.

Both methods performed appropriate normalisation across the six samples, being the median of the quantile normalised counts slightly more homogenous across the two cell types (Figure 1.5 a). When looking at the first 19,855 limma voom DOCs FDR ranked, 18,768 of them were the same as the retrieved by DESeq2. Moreover, very significant postive 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.

Moreover RANK preserved, different way of calculating FP rate with FDR but similar FC in quantile which is more stable to noise than TMM and preferred since TMM identifies even greater number of hits and more variation in the data(boxplot). Maybe justify in the dicussion the use of DESEq2 and limma shared based on Alasoo observation of more noise.

Clustering and heat map and pathway analysis-briefly

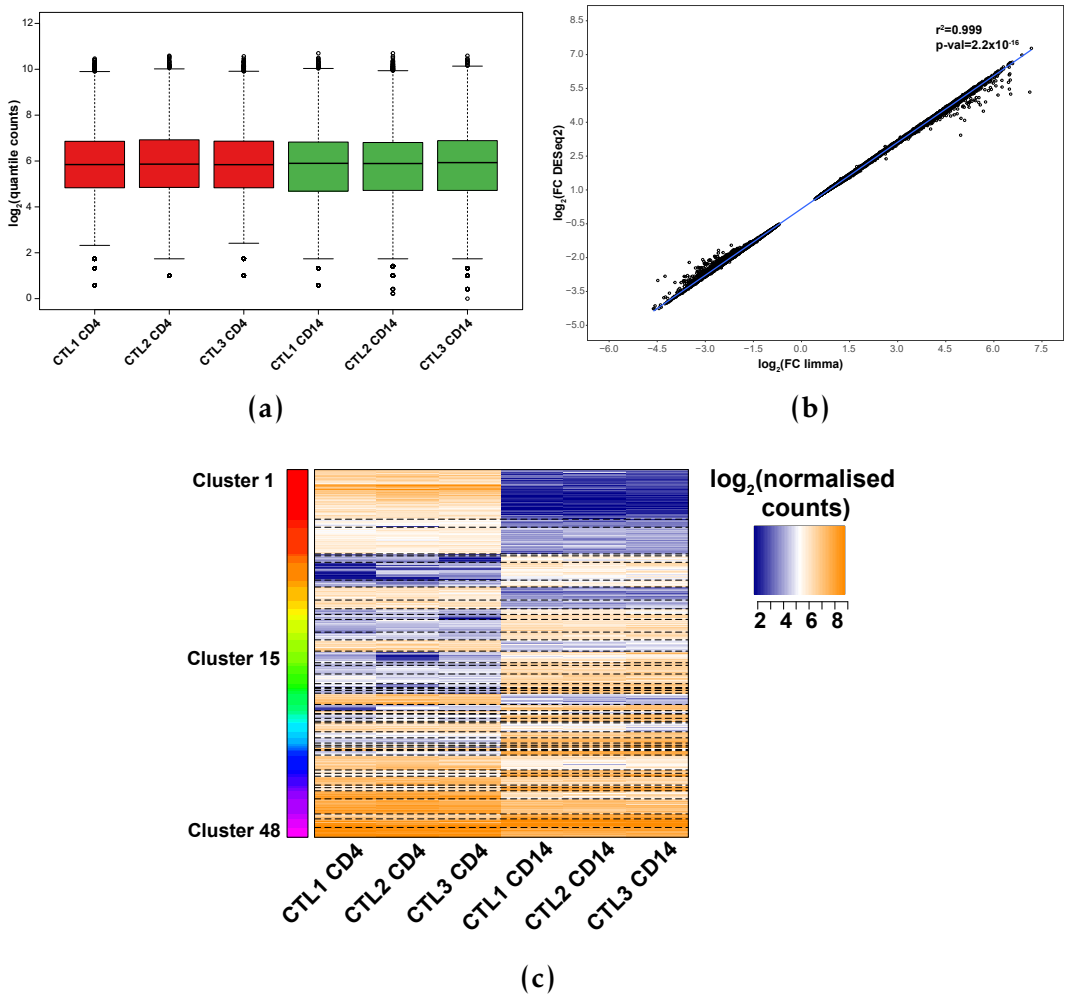
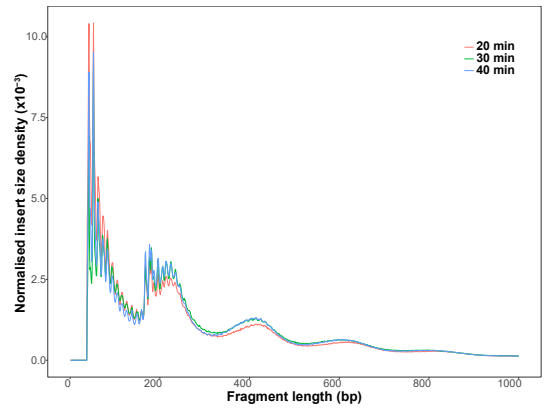
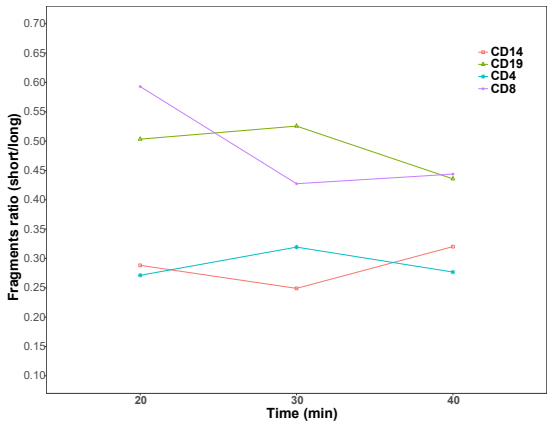


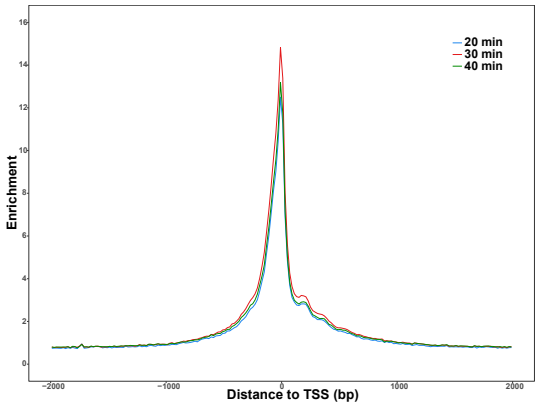
Figure 1.5: Exploration of the differential chromatin accessibility analysis using 80% as the empirical cut-off.



(a)

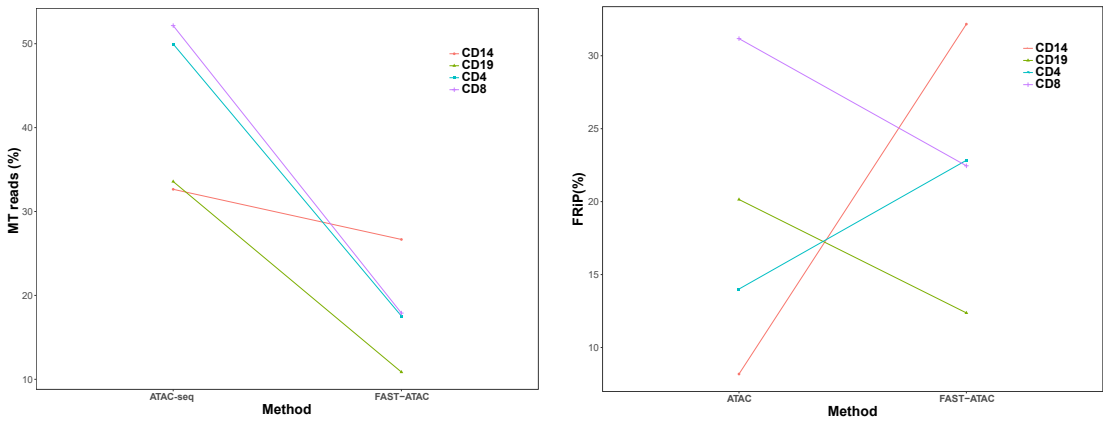


(b)



(c)

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



(a) (b)

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

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

1.2.3 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.

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

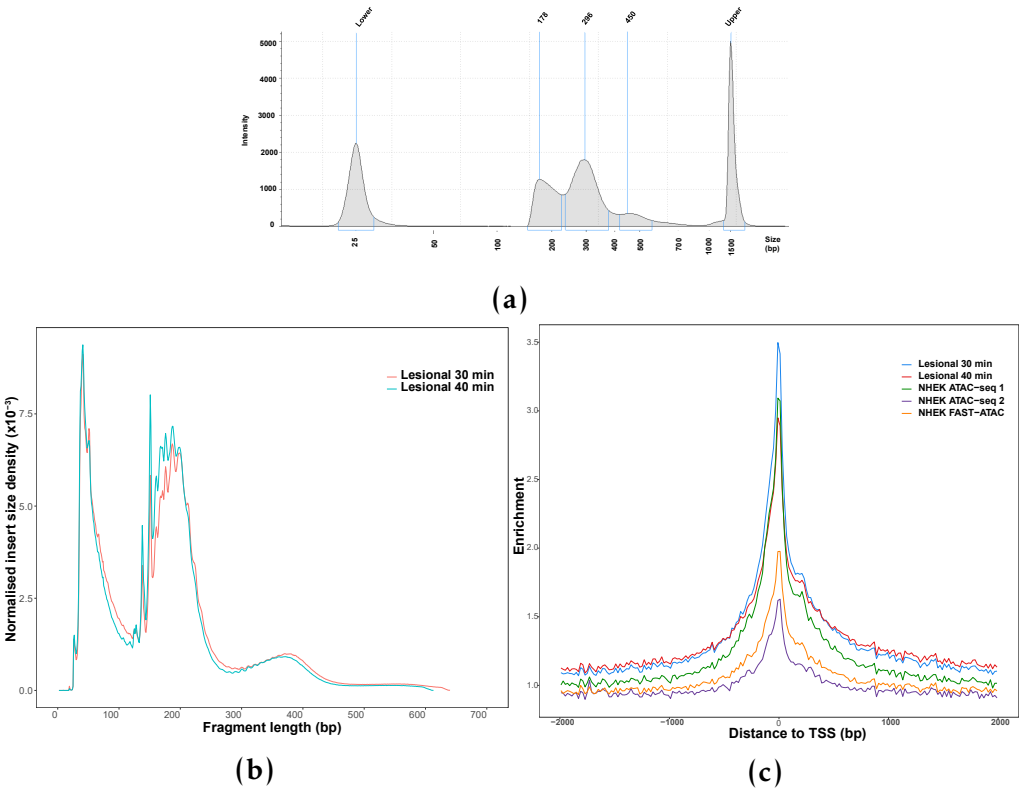


Figure 1.8: QC assessment of ATAC-seq in KC enriched cell suspension derived from a psoriatic lesional skin biopsy. Two transposition times (30 and 40 min) were tested using the standard ATAC-seq protocol (Buenrostro *et al.*, 2013 in 50,000 cells from the same suspension.

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.

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

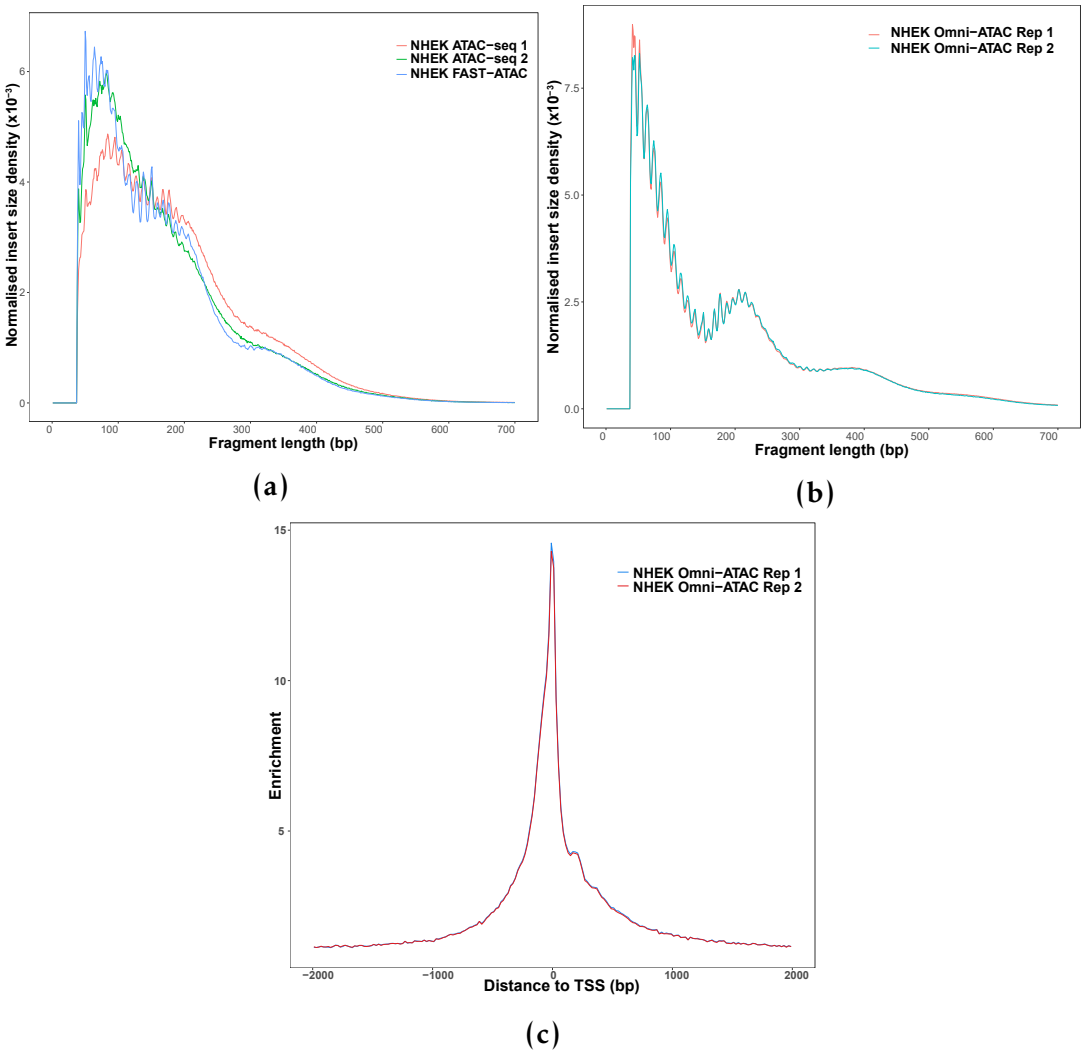


Figure 1.9: QC assessment of FAST-ATAC and Omni-ATAC in cultured NHEK.

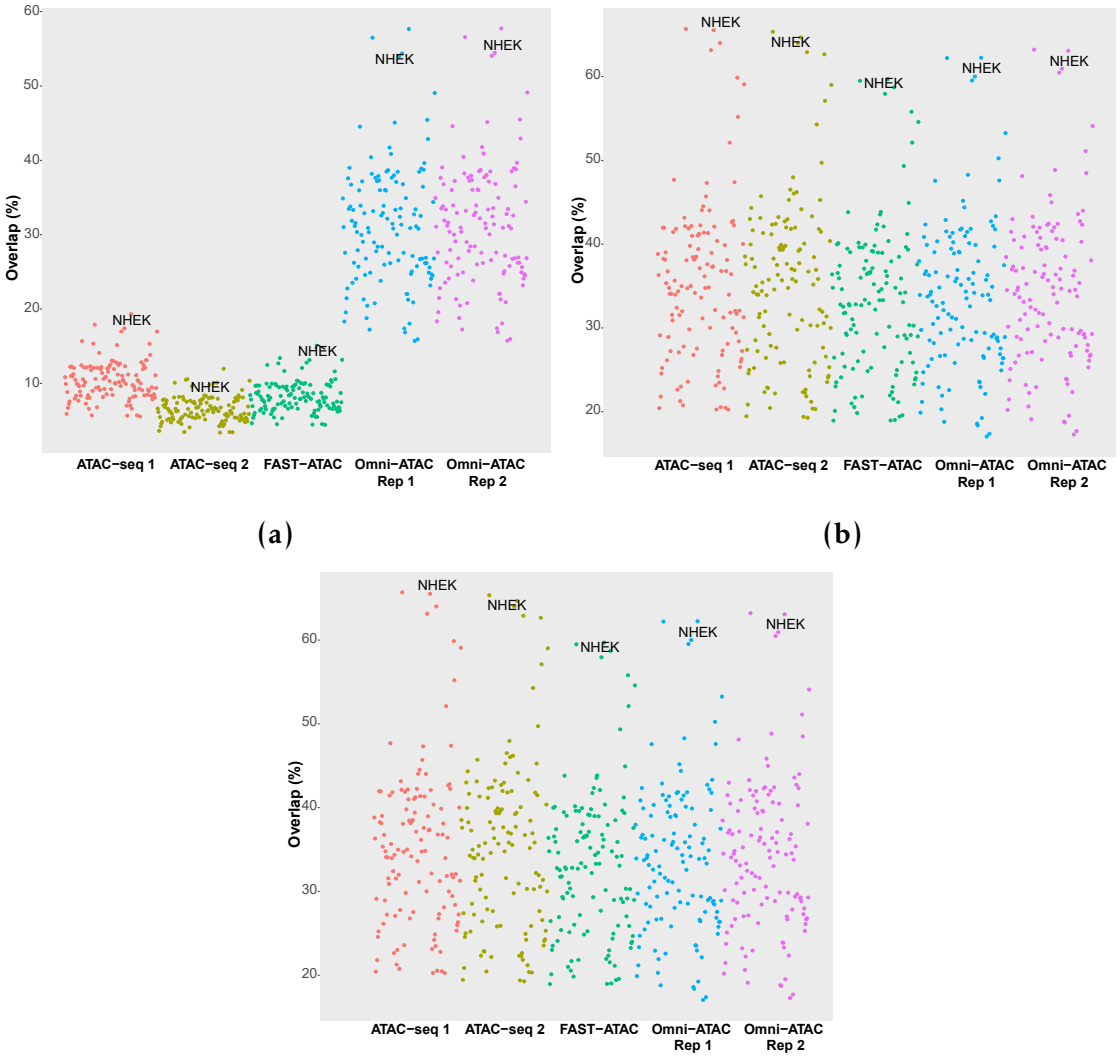


Figure 1.10: QC assessment of Omni-ATAC in NHEK and chromatin accessibility signal for the samples generated with the different ATAC-seq protocols.

1.2.4 Discussion

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

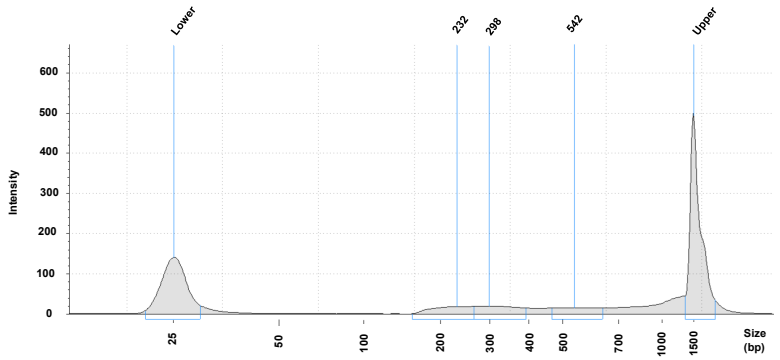
Appendices

List of Figures

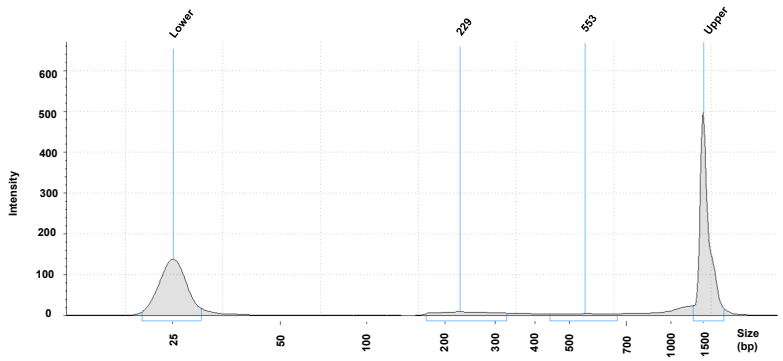
List of Tables

Appendix A

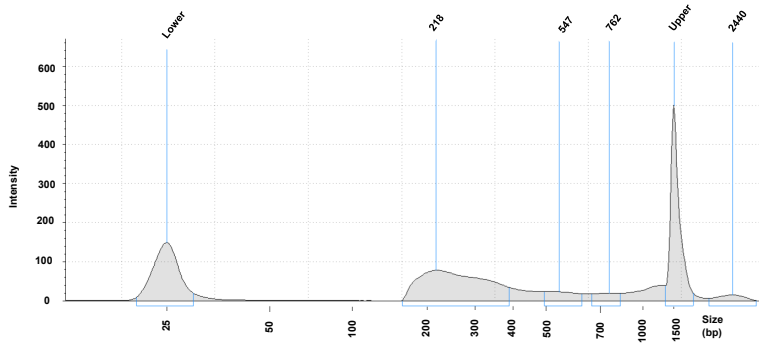
Establishment of methods to assess genome-wide chromatin accessibility



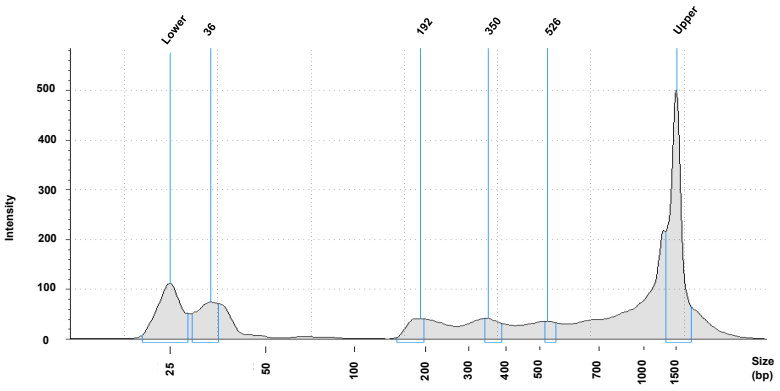
(a)



(b)



(c)



(d)

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

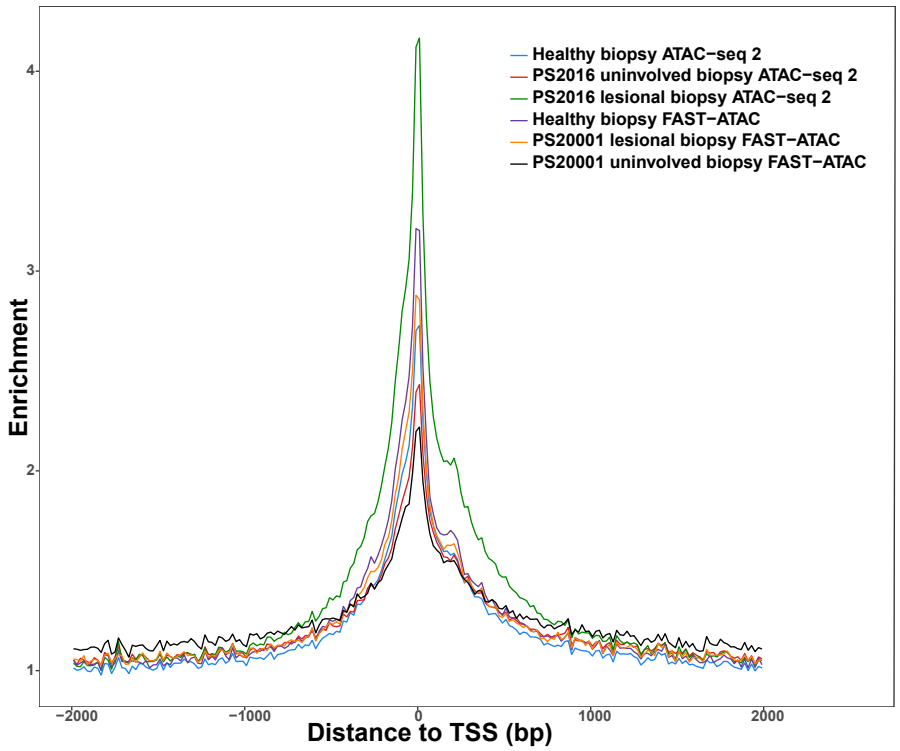


Figure A.2: Assessment of TSS enrichment from ATAC-seq and FAST-ATAC in healthy and psoriasis skin biopsies samples.

Bibliography

Jonker, M. J., Leeuw, W. C. de, and acids, Marinkovi-M. (2014). "Absence/presence calling in microarray-based CGH experiments with non-model organisms". *Nucleic acids*. DOI: 10.1093/nar/gku343. URL: http://dx.doi.org/10.1093/nar/gku343.