~~\section{Computational and statistical analysis}~~

~~\subsection{ATAC data analysis}~~

~~\label{ATAC\_analysis}~~

~~ATAC-seq, Fast-ATAC and Omni-ATAC data were analysed using an in house pipeline, towards which I made an important contribution. The pipeline performs single sample data processing and also builds a combined master list for each of the conditions of interest for chromatin accessibility characterisation and differential analysis.~~

~~\subsubsection{Next generation sequencing data analysis}~~

~~NGS data for each sample was trimmed for low quality base pairs and Nextera adapter sequences using cutadapt \parencite{Martin2011} before general QC assessment using fastqc \parencite{Andrews2010}. Trimmed reads were aligned to the reference genome (build hg19) using bowtie2 \parencite{Langmead2006} and the following parameters -k 4 -X 2000 -I 38 $-$$-$mm -1, consistent with other publications \parencite{Buenrostro2013, Corces2016}. Samtools \parencite{Li2009} was used to remove PCR duplicate reads marked with Picard Tools ( http://broadinstitute.github.io/picard/) as well as low MAPQ (${<}$30), non-uniquely mapping and non-properly paired reads. The resulting bam file was additionally filtered to remove mitochondrial DNA and reads were adjusted by $+$4bp on the plus strand and by $-$5bp on the minus strand to represent the center of the transposition binding event. Pileup tracks (bigWig files) representing the number of reads per bp position were generated using bedtools genomeCoverageBed \parencite{Quinlan2010} and the UCSC genome browser bedGraphToBigWig tool \parencite{Kent2010}. For visualisation purposes, normalised bigWig files were generated from normalised bedgraph files with bedtools genomecov and the library size factor estimated by the differential analysis algorithm.~~

~~\subsubsection{Peak calling, filtering and sample quality assessment}~~

~~Peak calling was performed using MACS2 callpeak \parencite{Zhang2008} and the parameters $-\-$nomodel $-\-$shift -100 $-\-$$extsize 200 $-\-$p 0.1 $-\-$keep-dup all $-$$-$call-summits. Peaks overlapping blacklisted features from the ENCODE project (~~[~~https://www.encodeproject.org/annotations~~](https://www.encodeproject.org/annotations)~~) were removed. The $-\-$shift and $-\-$extsize parameters were set according to the recommendations of MACS2 for DHS and following other ATAC-seq publications \parencite{Buenrostro2015, Corces2016}. The pval cut off for filtering called peaks was determined for each cell type using Irreproducibility Discovery Rate (IDR) analysis (as further detailed in Chapter \ref:{ch:Results1}). For this, the filtered bam file of each sample was partitioned into two equal size files (pseudoreplicates). Peak calling was performed in each pseudoreplicate, followed by filtering for a range of pvals (from 0 to 10$^{-45}$) and IDR analysis done for the resulting pairs of filtered peak sets. For each of the filtering pvals, the percentage of peaks sharing IDR rank between the two pseudoreplicates was determined, and the optimal pval filter identified. When more than one summit was identified in a peak, the median of the summits was used. For all peaks, summits were extended $\pm$250bp to create a non-overlapping homogenous 500bp peak list for each sample \parencite{Buenrostro2015, Corces2016}.~~

~~Sample quality was determined by the fold enrichment of ATAC-seq signal across all the hg19 TSSs annotated by Ensembl (175,114 features), computed as in \parencite{Buenrostro2015} using a script provided by Dr Silvia Salatino. In brief, transposition events were calculated in 1bp bins $\pm$1,000bp surrounding all TSSs and normalised to the mean value of background reads (signal from -1,000 to -800). For overall library quality assessment all ATAC fragments were considered. When assessing chromatin structure within or across the TSS, fragments of $<$150bp or 260–340bp were used, respectively \parencite{Scharer2016}. Fraction of reads in peaks (FRiP) was calculated for samples in Chapter \ref{ch:Results1} as the overlap between the peak list filtered for FDR$<$0.01 and all ATAC fragments using bedtools intersect with the parameter $-\-$f 0.1.~~

~~\subsubsection{Combined peak master list and differential analysis}~~

~~To perform differential open chromatin analysis a non-overlapping 500bp peak master list including all samples for a particular contrast was built. Each master list was built by union of all the peaks present in at least 30\% of the samples included, regardless of the subgroup to which they belonged (e.g patients or controls, SF or PB). Reads overlapping each of the peaks in the master list were retrieved for each sample using the HTSeq-count algorithm \parencite{Anders2015}. Principal component analysis (PCA) was performed on all counts normalised with the vsd function from the DESeq2 v1.20 R package \parencite{Love2014}.~~

~~For differential chromatin accessibility analysis, an empirical 80\% confidence cut-off was calculated and used to pre-filter peaks with high noise that could be confounding the analysis (detailed in Chapter \ref{ch:Results1}). Differential analysis was performed using DESeq2 with a paired design (in Chapter \ref{ch:Results1} and Chapter \ref{ch:Results3}) or including batch as a covariate in the model (in Chapter \ref{ch:Results2}). Peaks were annotated with proximal genes ($\leq$5Kb) using xGR2nGenes function from the XGR R package \parencite{Fang2016}. These gene lists were intersected with differentially expressed genes (DEGs) from RNA-seq/scRNA-seq analysis or genes reported by psoriasis and PsA GWAS.~~

~~\subsection{ChIPm data analysis}~~

~~\subsubsection{Next generation sequencing data analysis}~~

~~ChIPm NGS data from samples and inputs were processed similarly to ATAC-seq (see Section ATAC-seq, Fast-ATAC and Omni-ATAC data analysis) for trimming, mapping and filtering with minor modifications. Specifically, the MAPQ30 score for filtering reads was lowered to 10. For visualisation, bedgraph files with noise subtracted using the control input were generated using MACS2 bdgcmp -m subtract followed by conversion to bigWig with bedGraphToBigWig tools.~~

~~\subsubsection{Peak calling, filtering and sample quality}~~

~~Peak calling for each ChIPm sample was performed accounting for background signal using the control input samples with MACS2 callpeak $-\-$bw 200 $-\-$p 0.1 $-\-$-$keep-dup all $-\-$call-summits. In this case the average library fragment size ($-\-$bw) was used by MACS2 to first empirically find the model that best represents the precise protein-DNA interactions and calculate the appropriate $-\-$shift parameter. For ChIPm PCA, filtering and downstream analysis peak homogenisation was performed similarly to Section \ref(ATAC\_analysis) to build a combined master list for all samples and cell types from cohort 1B (Chapter \ref{ch:Results2} Table \ref{tab:Psoriasis\_cohort\_metadata} and \ref{tab:Control\_cohort\_metadata}).~~

~~Sample quality was determined by a combination of measurements. For library complexity the non-redundant fraction (NRF) and PCR bottleneck coefficients (PBC1 and PBC2) were calculated following ENCODE guidelines (https://www.encodeproject.org/chip-seq/histone/) from unfiltered bam files. Enrichment of the ChIPm signal was evaluated based on the normalised strand cross-correlation coefficient (NSC) and relative strand cross-correlation coefficient (RSC), calculated with SPP using bam files filtered for low MAPQ30, duplicated and non-properly paired reads.~~

~~\subsubsection{Combined peak master list and differential analysis}~~

~~DiffBind (default parameters unless specified) was used to build a peak master list and perform differential H3K27ac analysis between psoriasis patients and healthy controls for each cell type. DiffBind used the unfiltered sample peak files generated by MACS2 and the filtered bam files (from samples and control inputs) to generate a master list including high quality reproducible peaks present in at least 50\% of the samples (modification from default parameters) and retrieve counts of the reads mapping at the location of each peak.~~

~~\subsection{Gene expression analysis}~~

~~\subsubsection{qPCR analysis}~~

~~Pre-processing of the qPCR data up to calculation of fold changes (FCs) for each gene was conducted by UCB collaborators. When comparing SF and PB from patients, FC was calculated for each patient as the ratio between the 2$^{-dCt}$ in each of the tissues. Therefore, in a particular cell type three FCs (one per individual) were provided for each gene. When comparing expression in PB from PsA patients versus healthy controls, the FC for each gene and each individual was calculated as the ratio between the average 2$^{-dCt}$ in the three controls and the 2$^{-dCt}$ of a particular patient. In order to determine the significance of the FCs, one sample t-test was performed with the null hypothesis being $H\_0:\mu=1$ (no change) and the alternative hypothesis $H\_1:\mu\neq 1$.~~

~~\subsubsection{Bulk RNA-seq analysis}~~

~~NGS RNA-seq data processing was performed using an in-house pipeline developed by Dr Katie Burnham. Ribo-depleted RNA-seq data was mapped against the reference genome (build hg19) using the aligner STAR \parencite{Dobin2013}. Mapping allowed multiple alignments and only retained those with the best score and a mis-match percentage lower than 0.04\%. Duplicates were marked and removed using Picard Tools. Gene counts were retrieved using HTSeq-count and the gencode hg19 annotation file comprising 2,840,278 gene entities, including lnc-RNAs. Differential gene expression analysis was performed with DESeq2 on genes with five or more reads in at least eight samples (smallest group size corresponding to the psoriasis patients' samples). Independent filtering of genes with low expression levels, outlier removal using Cook's distance and moderation of log$\_2$FC were enabled when using DESeq2. Differentially expressed transcripts were identified based on False discovery rate (FDR)$<$0.05. Batch effect was included as a covariate in the contrast between psoriasis patients and healthy controls. This effect related to the RNA extraction, library preparation and sequencing of cohort 1A and cohort 1B samples from the psoriasis study (including healthy controls). Lnc-RNAs were annotated using the list provided by gencode.v19 (https://www.gencodegenes.org/releases/19.html). The paired design of the psoriasis skin DGE analysis was taken into account by the DESeq2 model.~~

~~\subsubsection{Single-cell RNA-seq analysis}~~

~~Raw Illumina sequencing data from the 10X Genomics technology Chromium single cell 3' expression libraries generated in bulk PBMCs and SFMCs from three PsA patients (see \ref{ch:Results3}) were first processed using Cell Ranger v2.2 software provided by 10X Genomics technology (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/). Illumina sequencing base call files (BCLs) were demultiplexed and converted into fastq files using cellranger mkfastq. For each of the samples, mapping of the fastq files to the compatible human transcriptome reference (GRCh38-1.2.0) and retrieval of counts for each transcript included in the reference genome were performed with cellranger count using default parameters.~~

~~The count matrix files were fully processed downstream using the R package Seurat 2.3.4 \parencite{Butler2018}. Each of the PBMCs and SFMCs individual count matrices were downsampled to 3,500 cells after removing all the genes expressed in fewer than 30 cells (drop-out events). Additional filtering was conducted to remove cells presenting more than 7.5\% of mitochondrial reads, a number of genes larger than 500$\pm$1SD (approximately 1,800 genes in all the samples) and a maximum of 10,000 UMI. After filtering, individual PBMCs and SFMCs count matrices were processed for data scaling and normalisation (regressing number of UMI and percentage of mitochondrial counts) and PCA analysis. The first eight PCs (capturing most of the variation in the data according to jackstraw analysis) were then used to identify clusters (groups of cells with similar expression profiles) built by a shared nearest neighbor (SNN) modularity optimisation algorithm (default resolution 0.6) followed by visualisation using t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensional reduction. CD14$^+$ monocyte clusters were selected based on co-expression of appropriate cell specific markers (\textit{CD14} and \textit{LYZ}). For each of the CD14$^+$ monocyte populations, the top variable genes were quantified by dispersion of expression (variance/mean) ratio, using a cut-off for maximum mean expression of 4 and minimum dispersion of 0.25. The union of the 1,000 most variable genes across the six CD14$^+$ monocyte samples (three from PBMCs and three from SFMCs) were used to perform canonical correlation analysis (CCA). The first nine canonical correlation vectors (CCs) were used to align all the CD14$^+$ monocyte populations. This alignment using CCA was applied in order to merge all the CD14$^+$ monocyte datasets, removing batch effects and allowing further downstream analysis. Cluster identification and visualisation using t-SNE in the combined CD14$^+$ monocytes from the six samples was performed (further detailed in Chapter \ref{ch:Results3}) followed by DGE analysis between the SFMCs and PBMCs CD14$^+$ monocyte populations.~~

~~\subsection{Genomic region annotation, enrichment analysis, gene-network analysis and pathway visualisation }~~

~~Genomic region annotation with chromatin states was performed by overlapping peaks with the appropriate cell type chromatin segmentation map from Epigenome Roadmap (https://egg2.wustl.edu/roadmap/web\_portal/chr\_state\_learning.html).~~

~~Genomic feature enrichment analysis was performed using XGR built-in data (including eQTLs, FANTOM enhancers, histone marks and TFBS, amongst others) and the function xGRviaGenomicAnno. When annotating regions (ATAC or ChIPm peaks) the summit of the peaks were used to increase the specificity of the analysis. Pathway enrichment analysis was conducted for the built-in KEGG, Reactome MsigdbC2BIOCARTA and MsigdbC2CPall pathways with the xEnricherGenesAdv XGR functionality. Input data used were genes annotating differentially accessible ATAC peaks or differentially expressed genes, and background was defined as all the annotated ATAC peaks from the differential analysis master list or all the gencode hg19 detected genes, depending on the analysis.~~

~~Gene network analysis was carried out with xSubneterGenes XGR functionality using as the input list all the qPCR array genes and as significance-level the best pval across the three cell types where the expression was assayed. This list of genes was superposed onto the STRING interaction network (including known and predicted protein–protein interactions) to obtain a maximum-scoring gene subnetwork (30 genes) containing as many highly significant (highly scored) genes as possible and a lesser number of non-significant genes as linkers.~~

~~Visualisation of the signalling-pathway enrichment from the RNA-seq results was performed using the R package Atlas and Analysis of systems-biology-led pathways (A2), developed by Dr. Hai Fang and towards which I have contributed (manuscript in preparation). The manually curated KEGG pathways expanded with genes for each gene family were coloured based on the FC from the corresponding RNA-seq differential analysis and highlighted in bold when passing the FDR threshold for significance.~~

~~\subsection{Statistical fine-mapping}~~

~~Fine mapping of the psoriasis and PsA GWAS signals was performed using a Bayesian approach, aiming to overcome the incomplete coverage of genotyping arrays and the hundreds of associations per locus due to the LD structure of the genome. Fine-mapping was conducted using two different strategies due to availability of summary statistics or genotyping data from the psoriasis and PsA Immunochip GWAS studies, respectively. Both strategies include the same main steps of imputation, association testing and calculation of PP and were implemented in collaboration with Dr Adri\'{a}n Cort\'{e}s.~~

~~\subsubsection{PsA fine-mapping using Immunochip genotyping data}~~

~~Fine-mapping was performed for a number of non-MHC PsA Immunochip GWAS susceptibility loci using 1,103 patients and 8,900 controls from the Bowes {et al.} PsA Immunochip UK cohort. Access to the data post-quality control was kindly provided by Dr Anne Barton (The University of Manchester) PCA analysis was performed using only pruned SNPs with flashpca \parencite{Abraham2014} and the calculated PCs were used as covariates in the association analysis to correct for population stratification.~~

~~.~~

~~%previously selected by Bowes \textit{et al.} based on the lead SNP FDR$<$10$^-4$ and a marker density in the region of at least 100 SNPs.~~

~~For each of the fine-mapped loci a 2Mb window around the lead SNP was defined and SNPs extracted from the data using PLINK 1.9 \parencite{Chang2015}. Phasing of the genotype data was performed with SHAPEIT \parencite{Delaneau2012} and used to impute missing genotypes with IMPUTE2 \parencite{Howie2009} and the 1000 Genomes Project Version 3 as the reference panel (October 2015 release). SNPs for which imputation was not successful in at least 70\% of the samples (info-score$<$0.7) were filtered out using QCtool.~~

~~The association and conditional analysis was conducted using a Bayesian additive model implemented with SNPTEST and including the previously calculated PCs as covariates \parencite{Burton2007}. Approximated Bayes Factors (ABF) were calculated for the lead signal and step-wise conditional analysis was performed if ABF$>$3. Credible sets of SNPs containing the variants likely to explain 50\% and 90\% of that association were identified for each of the signals in the locus, along with their corresponding posterior probabilities (PP) as further detailed in Bunt \textit{et al.}, 2015.~~

~~\subsubsection{Psoriasis fine-mapping using Immunochip summary statistics}~~

~~Fine mapping was performed for X of the risk loci reported by the psoriasis Immunochip GWAS study from Tsoi \textit{et al.}, 2012, for which only the summary statistics of the GAPC Immunochip cohort (2,997 cases and 9,183 controls) were available. The summary statistics file included the pval and the OR calculated for each of the genotyped SNPs using a logarithmic regression model and correcting for ten principal components. The statistic z-scores from the genotyped SNPs were used in the direct imputation of summary statistics (DIST) method to impute the z-scores for allele 1 of the missing SNPs based on the correlation in linkage disequilibrium (r$^2$) from the 1000 Genome Project Version 3 \parencite{Lee2013}. Imputation was performed genome-wide for all autosomes and the results were filtered based on the quality of the imputation ($>$ 0.8).~~

~~Association analysis and calculation of the ABF were performed using Wakefield approximation for a 2Mb window around the GWAS lead SNP of each locus of interest. This approximation was applied under the priors of (1) normally distributed OR with mean and variance (v$^$2); (2) the greater the variance the bigger the size effects obtained will be; (3) mean is 0 and variance is fixed to 0.2 (accepted variance for GWAS studies). In this approach ABF was calculated using effect size (\beta) and standard error (SE) derived from the variance formula. $\beta$ was calculated using the z-scores of each of the interrogated SNPs as: $\beta$=z-score\*SE. It is important to note that step-wise conditional analysis is not performed when using summary statistics imputation. Similarly to the genotyping fine-mapping approach, PP for the SNPs in a particular window (signal) were calculated and ranked to set the threshold of the 50 and 90 \% credible set of SNPs.~~

~~\subsection{Mass cytometry data analysis}~~

~~Mass cytometry analysis was performed by Dr Nicole Yager. Mean expression of cytokine release was calculated following unsupervised clustering analysis in a workflow similar to Nowicka \textit{et al.} with minor modifications \parencite{Nowicka 2017}. For each population of cells, cytokine production in SF and PB was calculated as the difference in the mean signal intensity between the 0h and 6h aliquots. The percentage of cytokine release was calculated following manual gating for the CD14$^+$ population based on surface marker expression. The percentage of TNF-$\alpha$ positive staining cells were calculated between the 0h and the 6h samples within each tissue.~~