School of Biological Sciences

ASSESSMENT COVER SHEET AND TEMPLATE

Section A – to be completed by the student

Student Number	**2228202**		
Programme	MSc Bioinformatics		
Unit Name	Bioinformatics Research Project	Unit Code:	BIOLM0034
Assessment name	**Bioinformatics Project Report**		
Word Count	**5903**		
Do you give permission for you work to be used anonymously in examples given to students in the future? (type Yes or No)			

By submitting this assignment cover sheet, I confirm that I understand and agree with the following statements:

- 'I have not committed plagiarism, cheated or otherwise committed academic misconduct as defined in the University's Assessment Regulations (available at https://www.bristol.ac.uk/media-library/sites/academic-quality/documents/taught-code/annexes/university-examination-regulations.pdf)
- 'I have not submitted this piece, in part or in its entirety, for assessment in another unit assignment (including at other institutions) as outlined in section 4 of the University's Assessment regulations (available at https://www.bristol.ac.uk/media-library/sites/academic-quality/documents/taught-code/annexes/university-examination-regulations.pdf)
- I understand that this piece will be scrutinised by anti-plagiarism software and that I may incur penalties if I am found to have committed plagiarism, as outlined in sections 3 of the University's Examination Regulations (available at https://www.bristol.ac.uk/media-library/sites/academic-quality/documents/taught-code/annexes/university-examination-regulations.pdf)

1 Title

Exploration of the colorectal cancer genome by liquid biopsies and Griffin nucleosome profiling

2 Abstract

Colorectal cancer (CRC) is a prevalent disease with a complex level of transcriptional regulation and molecular sub-types. Liquid biopsies have emerged as a less invasive and more repeatable method to isolate biomarkers compared to tissue biopsies. Circulating tumour DNA (ctDNA) is a biomarker with a wide range of applications in precision medicine. The recently developed bioinformatics framework Griffin can be utilised to infer nucleosome profiles from ctDNA sequenced at low coverages.

This study has applied Griffin to CRC patient liquid biopsy blood samples before radiotherapy. In an exploratory analysis we inferred chromatin accessibility at transcription factor binding sites. Transcription factors with high and low chromatin states and moderate associations with treatment response have been identified. To test Griffin's claim of versatility for transcriptional analysis we also tested its ability to estimate chromatin accessibility at transcription start sites.

Griffin does provide a reproducible framework to apply to CRC data, but other features such as copy number variations and DNA methylation are required for successful tumour sub-typing. Further research is required to determine Griffin's suitability for CRC detection, treatment prediction, and how its susceptibility to batch effects impacts clinical utility.

3 Dedications and acknowledgements

Thank you to Francisca Segers and Adam Chambers for all your support and dedication. Finally, a special thank you to the patients who agreed to take part in this study.

4 Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ...Glen Roarke.................... DATE:...05/09/2023......................

Students must print their name on the examination copy and on the final Library copy.

4 Table of contents

Contents

5 Introduction

5.1 Colorectal cancer is a common disease with complex molecular subtypes

Colorectal cancer (CRC) is the third most prevalent cancer globally with a high rate of metastasis and mortality. The majority of cases are sporadic in nature and incidence is increasing in younger individuals and non-western populations (1). This makes improving clinical outcomes for CRC patients an important global challenge. Current research has demonstrated that CRC tumours present significant intra-tumour heterogeneity, with each tumour estimated to possess small quantities of common mutations and much larger quantities of low frequency mutations (2). There are three main genetic pathways for sub-classification of CRC, with one agreed model separating tumours into three distinct groups. First, the chromosomal instability (CIN) group is identified by mutations in proto-oncogenes and tumour suppressors. Secondly, the microsatellite instability group (MSI) presents genetic hypermutability caused by dysfunctional DNA mis-match repair genes. The final group is CpG Island methylation (CIMP) caused by hypermethylation of DNA regions (3). Other classification models centred on gene regulation of CRC have been suggested adding to the complexity of sub-type classification. Transcriptional profiles provide even more heterogeneity and there is also a consensus molecular subtype (CMS) classification, containing 4 sub-types with prognostic capabilities (4). Intra-tumour heterogeneity leads to sub-clonal populations of tumour cells which compete and co-operate with each other (5). Evidently, new precision medicine techniques are required to better understand the complexity of the CRC mutational and regulatory landscape. This requires research and development of new genetic and epigenetic assays that use genomics and bioinformatic analysis to improve clinical outcomes for CRC patients with divergent cancer sub-types.

5.2 Liquid biopsies and cell free DNA provide a new way to explore the cancer genome and epigenome.

Liquid biopsies (LBs) are a complementary advancement within precision oncology providing a less invasive, more accessible insight into tumours compared to regular tissue biopsies (6). This technology has diagnostic and prognostic capabilities which can be taken much more frequently (7). There are several promising biomarkers within a patients blood plasma, with cell free DNA (cfDNA) being an accessible marker (8). cfDNA are small DNA fragments created during natural programmed cell death processes such as apoptosis and necrosis. For a cancer patient these processes occur on both healthy and pathological cells, with the cancer cells introducing circulating tumour DNA (ctDNA) into the blood stream (9). Experimental evidence has proven that ctDNA can be detected at a high specificity and sensitivity (10). CtDNAs biological features include genetic analysis of single nucleotide polymorphisms and copy number variations, however ctDNA also permits inferences about epigenetic features such as fragmentomics and DNA methylation (11). Cancer cells exhibit different cfDNA fragmentation proportions compared to healthy cells (12). Furthermore, pan-cancer studies have demonstrated that different cancers and disease phenotypes display different fragmentation proportions (12). CRC presents high amounts of ctDNA which justifies applying cfDNA analysis workflows to colorectal sequencing data. Additionally, it has been demonstrated that liquid biopsy and cfDNA analysis detects more tumour heterogeneity compared to traditional tissue biopsies in CRC (13). Consequently, liquid biopsy analysis is becoming increasingly common in a wide range of cancer research.

5.3 Nucleosome profiling gives inferences on gene expression

Nucleosome positioning can be defined as the naturally dynamic genomic location of nucleosomes (14). Nucleosome positioning and chromatin accessibility are important in CRC disease progression because the dysregulation of a complex network of transcription factors determines different cancer phenotypes such as angiogenesis, cell proliferation and metastasis (15). cfDNA fragmentation is determined by the DNA and nucleosome binding relationship, allowing for the exploration of transcriptional regulation using this analyte. DNA wraps around the nucleosome octamer of histone proteins to a length of approximately 147 bp (16). There is also linker DNA present between each nucleosome which varies in DNA length depending on the type of organism and tissue (17). In vitro studies have determined that linker DNAs interaction with histones is important in providing nuclear rigidity and mechanical chromatin function. Histone acetylation is also essential in reducing the interactions between nucleosomes and the decompaction of chromatin (18). Cellular chromatin organisation provides regulatory control of transcription factors (TFs) involved in tumorigenesis and is a contributor to somatic co-mutations in certain driver genes (19). Chromatin differences are largely sub-clonal in nature, and it is important to understand these fundamentals of CRC biology. Recent spatial multi-omics profiling approaches have determined that the epigenome can influence somatic mutations and that transcription factor signals can highlight the occurrence of epigenetic reprogramming (20).

Nucleosome positioning can inhibit and permit the active binding of TFs at promotor sites that regulate the cell cycle. However this is a complex biological process with some studies showing nucleosome positioning at promoters does not inactivate gene expression (21). The term nucleosome profiling is becoming more prominent in epigenetic analyses to make inferences about the regulatory composition of tissues and tumours using cfDNA. The biological feature of cfDNA consists of nucleosomes giving increased protection from nuclease degradation when bound to DNA. At less accessible regions of chromatin where nucleosomes are positioned, a higher coverage is observed when sequencing the cfDNA fragments from blood plasma. Conversely, the more accessible genomic regions are not bound with nucleosomes, have less protection for nuclease degradation and lower resulting coverage profiles. These differences in coverage allow for computational inferences about transcriptional regulation to be made. A recent nucleosome profiling tool called Griffin has been developed making use of the described biological features (22). Griffin implements a novel method for profiling nucleosomes and inferring chromatin accessibility using a fragment-wise GC-correction procedure on ultra-low pass whole genome sequencing (ULP-WGS 0.1x) liquid biopsy data. ULP-WGS has been demonstrated to provide a reliable and cost effective technical method for copy number variant detection (23) and is increasingly being used in cancer genomics to reduce the computational bottleneck of human genome analysis. An important research aim in this study is to apply the innovative tool Griffin to CRC patient data. Advancing the understanding of nucleosome profiling will result in an improved understanding of CRC subgroups such as the chromosome instable (CIN) group and its effect on treatment response.

5.4 Transcriptional start sites and genomic datasets

There are an excellent range of publicly available datasets containing genomic co-ordinates applicable to cancer bioinformatics analysis. The tool Griffin makes use of datasets such as the global transcription regulation database (GTRD) (24). Data sources such as these provide an excellent resource for inputting genomic information into bioinformatics software to make inferences on transcriptional regulation. Transcriptional starts sites (TSS) where RNA-Pol-II is recruited to initiate transcription are a fundamental feature of eukaryotic gene regulation. Correlation analysis studies have demonstrated the usability of the EPD database for analysing nucleosome positioning and its influence on gene regulation of 5 model organisms (25). The nucleosomes are important in transcriptional regulation, and there are highly conserved chromatin structures such as the nucleosome free region (NFR) and nucleosome phasing upstream and downstream of the TSS, sequentially termed -1 and +1 nucleosomes (26).

This study aims to test the claim of Griffin being applicable to all nucleosome transcriptional biology by experimenting with different regulatory databases. The eukaryotic promoter database of experimentally determined RNA-Pol-II promoters could provide an interesting addition to analysing TFs from the GTRD database. Another question is whether the chromatin accessibility of promoters can be determined on a colorectal cancer ULP-WGS 0.1x liquid biopsy dataset? This computational analysis will explore developing a new Griffin configuration for TSS sites to analyse CRC driver genes alongside TFBSs.

5.5 Improving clinical understanding with liquid biopsies and personalised medicine frameworks

The multi-omics interrogation of the cancer genome is essential to transition into the personalised medicine era by classifying the diverse range of patient biomarkers. The molecular sub-typing discussed looks to improve on the diagnostic and clinical challenges in CRC (27). Liquid biopsies have great potential to improve the monitoring of minimal residual disease after treatment. The small quantities of cells left after surgery are very problematic to detect on traditional scans. Studies have shown that ctDNA detection after surgery is a sensitive prognostic biomarker for disease recurrence, and can be used to adjust adjuvant treatments and its duration (28). The combination of sequencing and clinical data presents the opportunity to ask an interesting question about whether a patient’s pre-treatment nucleosome profile can be used to make predictions on post-treatment response to radiotherapy?

During metastasis, time-series analysis of ctDNA is an interesting tool to monitor treatment response and ctDNA quantity correlates with levels of tumour burden (29). Truncal mutations are best suited to tracking treatment response via targeted gene panels, with the genes such as BRAF, TP53, APC and KRAS being investigated (30, 31). cfDNA can also be used to identify genomic drivers for treatment resistance that occur in rare sub-clones presenting intra and inter-heterogeneity. As discussed, epigenetic profiling is an important future consideration in personalised medicine for CRC patients to detect tumour heterogeneity. Population based DNA methylation studies such as GRAIL are an example of the research direction epigenetic assays require to improve clinical utility compared to gene panels (32). Incorporating these promising experimental results into a clinical setting is a significant challenge.

This study aims to process liquid biopsy sequencing data at low coverage to infer chromatin accessibility using nucleosome positioning. TFBSs genomic information was used to create nucleosome profiles at sites of interest. Additionally, a new set of genomic information was processed through Griffin to observe nucleosomes at TSS. A range of multivariate exploratory analysis techniques were used to visualise and identify the TFs with the most extreme chromatin states.

6 Method

6.1 Computing environments

The University of Bristol high performance computing cluster BlueCrystal was used to execute the first two phases of the pipeline, ctDNA pre-processing and Griffin nucleosome profiling. R analysis was undertaken on version 4.2.1.

6.2 Sequencing and samples

The liquid biopsy blood plasma samples were obtained from ASPIRE (IRAS 141548) and SectR cohort study (IRAS 271831). The sequencing method used was NextSeq500 and processed at the Bristol Genomics Facility. Batch 1 read length = 2 x 150bp and batches 2 and 3 = 2 x 75 bp. The patient samples were divided into pre-treatment samples (A) and post-treatment samples (B). Batch 1 and 2 consisted of 25 samples and batch 3 had 33 samples. Patient metadata was obtained from clinical records collated in excel and included information on tumour regression grade (CAP-TRG), age and gender.

6.3 Circulating tumour DNA pre-processing

The raw ctDNA fastq files were pre-processed using a ctDNA data preparation workflow containing the human reference genome (h38) and relevant genome annotations.

The paired end reads were trimmed using fastp and basic read statistics produced (33). The alignment software BWA was used to map the reads against the human genome reference (h38) (34). Next, bam file creation and sorting was conducted using samtools (35). The Genome analysis toolkit (GATK) was used to remove duplicate reads from the bam files and recalibrate base quality scores (36). The GATK recalibrated bam file is used in the Griffin downstream analysis. Table 1 provides a summary of the software, versions and links to documentation.

6.4 Assessing pre-processing performance

The ctDNA preparation pipeline created a range of file outputs which can be used to monitor read pre-processing performance. A pipeline was created to calculate coverage, generate read statistics from fastp outputs and merge the results into a format more useable for data analysis. This data manipulation used pandas and argparse modules to create re-usable pre-processing tools for results generation on the 3 batches of cfDNA sequence data (Table 1).

6.5 Griffin workflow for Transcription factor binding site analysis

The Griffin workflow had three steps, the first is the Griffin genome GC frequency which was completed for the h38 genome by the developers of the software. Next, the GC mappability correction step was executed on the recalibrated bam files, producing a GC bias corrected intermediate file. Lastly, the Griffin nucleosome profiling step was executed on a list of transcription factor binding sites (TFBSs) to generate an inference of quantitative chromatin accessibility. An additional python data consolidation step was required to merge all griffin results into one data frame for further analysis in R (Table 1). Figure 1 shows and overview of the nucleosome pipeline with scripts available at 03_Supplementary_script_HQ_samples_.html.

6.6 Transcription start site analysis using Griffin

A novel configuration was developed to analyse transcription starts sites (TSS) using the Griffin nucleosome profiling framework. A TSS configuration yaml file was sourced from the Griffin development branch (Table 1). Parameters were adjusted in the Griffin snakemake to increase the window size to 2,500 bp up and downstream of the TSS. The EPD database was used to download a text file of TSS co-ordinates (Table 1). Data manipulations were required on the raw EPD file resulting in a file format where each variation in TSS had an individual site file e.g (TSS_1, TSS_2) This modified file had the correct structure to be inputted into Griffins site list. Next, the CRC gene drivers list was used to select the amount of TSS down to 1,116. Only the last stage of Griffin was re-run for the TSS nucleosome profiling stage. The file 06_Supplementary_script_create_TSS_sites provided an overview of how to create TSS site lists.

6.7 Exploratory data analysis of nucleosome coverage

Downstream data analysis used R and the R package Tidyverse. The R package Complex heatmap was used for the visualisation of expression profiles (37). Unsupervised PCA analysis was done using the precomp R package. The variable College of American Pathologists Tumour Regression Grading (CAP-TRG) (38) was used in PCA and hierarchical analysis to observe any clustering patterns with central coverage. The CAP-TRG ranking is,

0) No viable cancer cells complete
1) Single cells or small groups of cells moderate
2) Residual cancer outgrown by fibrosis minimal
3) Minimal or no tumour killed or extensive residual cancer poor

The supplementary scripts contain full lists of R packages and versions.

6.8 PCA analysis and batch effect correction

Batch effect detection, correction and evaluation was approached using the methodology in (39), making use of the mixOmic package. Batch correction was undertaken by the Combat package (40).

7 Results

7.1 Pre-processing of samples identifies variation in sample quality

The processing of samples through the ctDNA preparation step generated pre-processing statistics to monitor the sequencing performance of each batch. The depth of coverage was calculated as approximately 0.64x across all sequencing batches demonstrating the ultra-low pass whole genome sequencing (ULP-WGS) liquid biopsy approach applied to the samples (figure 2.A). In batch 2 and 3 several samples had a low mapping proportion (Figure 2.B). When examining the relationship between read mapping rate and DNA concentration it was found that samples with low quantities of DNA resulted in a low read mapping rate (Figure 2.C). This had identified a technical bias that needs to be considered in downstream analysis.

7.2 Read mapping technical bias influences quantitative measures of chromatin accessibility

Hierarchical clustering of central coverage across all batches indicated that the low read mapping rate caused by low DNA concentration had introduced a technical bias that affected the response variable central coverage. This issue was evident in batch 3 which had several outliers for low read mapping (Figure 3). Consequently, it was justified that certain samples could be excluded from further downstream analysis to maintain consistency. The following exclusion rules were applied across all sequencing batches and are summarised in table 2. Any sample with a read mapping proportion below 0.8 was removed from further analysis (Excluding batch correction and subsequent PCA). Sample As were selected to link to treatment response and not because of technical bias.

7.3 Exploratory analysis of CRC nucleosome profiling data

Griffin and its list of 270 TFs each containing the top 30,000 different transcription factor binding sites (TFBSs) were processed on a liquid biopsy colorectal cancer dataset. The quantity of sites were selected based on prediction performance from the Griffin study (22). The quantitative measures of chromatin accessibility were plotted for TF families of interest and different patient samples. It was necessary to replot this data from the default plots for better interpretation. Figure 4 shows the central coverages for different transcription factor families of interest based on CRC transcriptional literature (15). This figure demonstrated how the chromatin state and regulatory activity varied for CRC transcription factors per patient.

7.4 Patient metadata annotations to chromatin accessibility

The metric CAP-TRG is a measure of tumour regression grade and is an interesting variable to compare with chromatin accessibility profiles. Figure 5 shows the central coverage for 270 TFs from patient samples before radiotherapy treatment. There is no indication of clustering based on the CAP-TRG indicator, with 0 being the best outcome (No viable cancer cells) and 3 being the worst outcome (Minimal or no tumour killed). The extra metadata highlights that most patients are above 50 years old and there is a higher proportion of males. Adding in these features allows for a more detailed exploratory analysis of patient samples and if any associations with central coverage occur.

7.5 Transcription factors with highest and lowest inferred chromatin accessibility

We identified the transcription factors (TFs) with the highest and lowest central coverage values across all batches. By sorting the central coverage matrices and selecting the TFs with the lowest coverage, biological interpretation can be gained from online resources such as NCBI & Uniprot. A list of transcription factors with the lowest coverage in all 3 batches was obtained by cross referencing the three central coverage matrices (Table 3). The same procedure was done for TFs with the highest coverage (Table 3.1). The full set of TFs listed in the inferred high and low chromatin states is provided in the supplementary data 1 & 2 files. ELF1 is a TF identified as low coverage across 2 batches in patient samples before treatment and is visualised in the TF annotations (Figure 5).

7.6 Associations between central coverage and CAP-TRG

Spearman’s correlation was used to search for associations between the central coverage of transcription factors and the ranked treatment response measure CAP-TRG. Any samples with missing CAP-TRG values were excluded from the analysis. Correlations with a p-value lower than 0.05 across all 3 sequencing batches are summarised in Figure 5.1. Moderate positive correlations were observed suggesting that as CAP-TRG indicator increases, so does central coverage for these TFs. These TFs were investigated to determine whether they were tumour suppressors or proto-oncogenes that may impact radiotherapy response. ELK3 has been associated with the RAS pathway which is linked to human cancer, angiogenesis and tumour growth (41). Another significant TF ZNF263 is involved in the regulation of transcription of RNA Pol II and has been linked to apoptosis resistance (42).

7.7 <u>Principal component analysis on Transcription factor binding sites</u>

The outputs of Griffin are multivariate and PCA analysis is an unsupervised method to reduce the number of features into a two-dimensional space. The fraction of variance explained by two principal components (PCs) is approximately 70%. Batch 2 had a higher amount of variance explained by the first two PCs with a value of 80%. Important clinical indicators such as CAP-TRG were added to identify any clustering or patterns. In figure 6 this analysis for PC1 and PC2 has failed to identify any clustering patterns for the clinical indicator CAP-TRG in the high-quality patient samples before treatment. The TF loadings in (Figure 6.1) show how different TFs influence the variance in PCs, which can also be searched in NCBI for biological interpretation. For example, it is interesting that FOXM1 has a small variable component loading of -0.07 for batch 1 and 3. This TF has associations with metastasis and cell proliferation (15).

7.8 <u>Batch effects and correction using ComBat</u>

An investigation into batch effects were undertaken on all pre- and post-treatment samples that met the read mapping threshold of 0.8. Figure 6.2-A demonstrates how across the 3 sequencing batches there is a large amount of variance and clustering. Batches 1 and 3 cluster together more closely than batch 2, with batch 1 being the most uniform. After PCA reduction batch 2 shows high variation between samples, highlighting batch effects occurring on the central coverage variable. ComBat (40) was used to correct the data using both a parametric and non-parametric method for comparison (Figure 6.2, B-C). The initial correction method does result in changes to PCA clustering, with each batch being more central to each other. Figure 6.3 demonstrates how different batch correction methods can change the clustering of patients in different sequencing batches.

7.9 <u>Nucleosome positioning at transcription starts sites.</u>

An output of quantitative chromatin accessibility was generated for central coverage, mean coverage and amplitude. Figure 7 demonstrates how nucleosome detection around the TSS has been achieved at regular intervals around the TSS region of the p53 gene. Patient samples were selected that have read mapping coverage at the p53 TSS. Additionally, using available experimental data within the UNSC genome browser and Nucome, shows that regulatory activity is expected at the same locations (Figure 8). Furthermore, the peaks correspond to 150 bp which is equivalent to the DNA length that wraps around the nucleosome. This TSS mode is still in development and high amounts of variance in normalised coverage have been identified when compared to the TFBS mode. For each sequencing batch high central coverage values have been identified for a wide range of TSS. Figure 9 demonstrates how read mapping proportion results in outliers in central coverage for the low quality samples.

8 Discussion

8.1 Summary

This study analysed colorectal cancer liquid biopsy data to infer the chromatin accessibility for a list of 270 TFs using Griffin. Exploratory analysis of the nucleosome profiles of patients before treatment have identified the top 60 TFs with the highest and lowest inferred chromatin states. A subset of TFs identified consistently in the high and low chromatin groups across all sequencing batches have been consolidated for further review in table 3 and 3.1. A novel mode has been developed to use Griffin to provide nucleosome profiling at transcription start sites, another important gene regulatory mechanism. To summarise, the Griffin TFBSs configuration provided a high breadth of nucleosome analysis across a large number of genomic regions. In contrast, the TSS configuration allowed for a much more precise nucleosome analysis at specific genes and regulatory regions. The transcriptional analysis has provided a list of TFs to further explore for biological relevance in the complex regulation of colorectal cancer. The versatility and customisability claimed by the Griffin framework has been tested on new regulatory sites of interest.

The Griffin framework is still in its early stages and there is a lack of literature on colorectal cancer performance. Quantitative measures of chromatin accessibility have been generated and it would be possible to reproduce these results on new experimental or publicly available sequencing data. This is expected and aligns with the claims from the Griffin publication that a customisable framework has been created to make biological comparisons with any nucleosome transcriptional element. Furthermore, this software was designed to work with ULP-WGS using liquid biopsy data of human cancer patients, which aligns very closely with the sequencing reads in this study.

8.2 Liquid biopsy pre-analytical challenges can impact sample quality

Variation in sample quality has been detected across all three sequencing batches. Low quality samples have been defined with a low DNA concentration responsible for low read mapping rate (Figure 2.C). A limitation of this study is that it has not assessed the level of contamination by leukocyte genomic DNA. Studies have confirmed that this is an important consideration that can influence DNA concentration when samples are exposed to room temperature for extended periods of time (43). There is a need for further analytical techniques to determine ctDNA from genomic DNA to improve accuracy of liquid biopsy analysis and reduce levels of contamination. Additionally, there are biological variables that impact on cfDNA analysis that are difficult to control such as an individual's fluid intake and cfDNA excretion through urine (44). A patients general physical health and levels of exercise also effect cfDNA levels in the blood (45).

8.3 **Identifying biological importance from chromatin inferences**

This bioinformatics analysis provides the opportunity to check the transcription factors for biological relevance and associations with CRC. CTCF, a TF in an accessible chromatin state (Table 3) has been shown to have a tumour promoting role in knockdown studies compared to normal colorectal tissue (46). Another proto-oncogene BCL6 has been identified in an accessible chromatin state. Interestingly, studies indicate that CTCF induces BCL6 chromatin modification at its transcriptionally active locus (47). When examining TFs with the lowest accessibility values more encouraging results are presented (Table 3.1). p63 has a strong association with malignancy and has been identified as a prognostic factor in CRC that correlates with overall survival (48). This biological feature closely aligns with the cancer stage of our patient samples.

A mixture of proto-oncogenes and tumour suppressor genes have been identified in both the high and low chromatin groups (Table 3 & 3.1). Recent studies suggest that proto-oncogenes are upregulated and tumour suppressor genes are downregulated to express a range of cancer phenotypes (15). However, when associating regulation with chromatin states there are added complexities. Upregulated TFs are not always in an open chromatin state and downregulated TFs in a closed chromatin state. The TFs identified in table 3 & 3.1 support this by containing a mixture of proto-oncogenes and tumour suppressors when checked against cancer gene data mining resources (49, 50). This again highlighted the complex regulatory network which relies on more than just the positioning of nucleosomes in cancer gene regulation. The current literature has established ATAC-Seq as an important method for analysing actual chromatin accessibility in response to cancer treatment (51, 52). This needs to be undertaken for identification of causal relationships.

8.4 PCA analysis on tumour regression and sequencing batches is challenging

No distinct clustering of central coverage was observed when undergoing PCA analysis associated with the treatment response measure CAP-TRG (Figure 6). This type of approach is a common reduction method used in multivariate epigenetic profiling studies (53). There are many reasons why this approach could be unsuccessful. A recent study has suggested that the sample size effect is important and careful consideration is required when sub-setting data (54), which was an action taken on these samples for pre-treatment analysis. Furthermore, this study does not attempt to use well-annotated datasets to guide biological interpretation (55). It is also possible that the CAP-TRG and TF variables have a small effect on the respective phenotypic differences being assessed (56). Additional PCA analysis has confirmed that batch effects had occurred in these sequencing samples (Figure 6.2 - A). The phenomenon is very common in low coverage WGS sequencing studies and required further consideration (57). Initial batch effect correction had been attempted using ComBat, with changes in the clustering occurring (Figure 6.2 & 6.3). Further evaluation of correction is required because batch correction has been shown to increase false positive rates in epigenetic studies (58). It will be interesting to analyse future liquid biopsy sequencing batches via these PCA methods to determine similarity between batches.

8.5 The importance of healthcare meta data and Griffin's susceptibility to batch effects

The hierarchical clustering analysis of nucleosome profiling alongside patient metadata such as age, gender and CAP-TRG has not revealed any clustering patterns (Figure 5). The visualisations do provide further insight into the unbalanced dispersion of age and gender across each sequencing batch. Cancer is a disease of variation and it is important to consider personal factors and biomedical treatment response measures. There is a complex set of biological features that contributes to cancer disease progression. It is agreed that personal attributes such as age, gender and ethnicity are very important in cancer disease progression (32).

There is a higher quantity of males in all samples on visual inspection (figure 4), and disparities in CRC aggressiveness and tumour location exist between different genders and sexes (59). An important contrast to the Griffin study (22) is that this sample of the CRC population was arguably closer aligned to real world healthcare data that is sporadic in nature and has data quality issues. Publicly available experimental datasets used in studies (22, 60) are likely to be thoroughly cleaned with any outliers or missing data removed. It has been acknowledged that Griffin is susceptible to batch effects occurring between different cfDNA sequencing workflows, lowering compatibility of Griffin across different workflows. This greatly reduces its applicability in a healthcare setting because data collection and sequencing is often un-standardised (61).

8.6 Challenges associating pre-treatment nucleosome profiles to treatment response and tumour sub-typing

Correlation analysis between TF central coverage and CAP-TRG for pre-treatment samples identified moderate positive correlations (Figure 5.1). The TF NFYA has the most significant p-value and recent studies have associated NFYA as an inhibitor of E-cadherin which then promotes CRC metastasis, the mediating co-transportation complexes of this pathway, S100A2/KPNA2 are a potential therapeutic target (62). However, putting these results into context is important, it is reasonable to expect common TFs that are experimentally validated therapeutic targets such as TP53 (63) to have significant correlation with treatment response and not extremely high p-values of 0.64. Important research is on-going to understand pre-treatment differential patient chromatin states and ctDNA levels in an attempt to predict radiotherapy treatment response (64). Additionally, detecting chromatin heterogeneity and understanding how it contributes to survival is an important factor (65).

Furthermore, no predictive molecular sub-typing based on treatment response measures were achieved in this analysis. The techniques such as PCA analysis and hierarchical clustering have been used in a preliminary attempt to understand this complex data. It is arguable that this work can contribute to future tumour sub-typing classification approaches used in Griffin. The foundations of the Griffin software were to create logistic regression models and apply the detective (66) and predictive (67) capability to external datasets. It is reasonable to suggest that Griffin can be applied to different CRC datasets. For example, it would be interesting to observe how Griffin performs on CRC data that has confirmed sub-typing in place pathologically and whether it can predict these sub-types on the sequencing batches used in this study. It is important to note the heterogeneity of different cancer types because the molecular sub-typing of CRC has different models which are not as well categorised as other cancers (3, 4). There is limited evidence whether Griffin can contribute to improving treatment response but tumour sub-typing is an important factor in achieving this.

Analysing transcription start sites was difficult to validate

Progress has been made to develop a new configuration of Griffin which provides the nucleosome positioning analysis at transcription start sites (TSS). However, this new approach is in its infancy and has mainly demonstrated how new datasets can be inputted into Griffin. Further work is required to understand the response variable central coverage and its increased variation compared to the TFBSs configuration. Studies have shown how gene regulation at TSS is important in the dysregulation of cancer, with histone modifications resulting in transcriptional repression of cancers (68). Additionally, research on the positions of nucleosomes at the TSS of known tumour suppressors is on-going (69). Further studies examining specific genes of interest in colorectal adenomas suggested silenced genes progress through different stages, with nucleosome positioning up and downstream of the TSS being an important component (70). Evidently, there is justification to develop the Griffin TSS configuration and it is likely why a development script has been provided by the publisher of the software. To conclude, as discussed in (22), TSS analysis at specific genomic locations of interest is better suited to samples with a higher depth of sequencing coverage. The Ulz et al study which Griffin is based on used a significantly higher mean coverage of 14.96× (60). There is a real need to determine if CRC sequencing coverage needs to be higher compared to other cancer types.

8.7 Limited access to controls and the need for complementary omics data

There are several limitations within this investigation which need to be improved to further validate the results. First, it was not feasible to obtain a control sample during this run of Griffin analysis which would allow for better interpretation of the TF chromatin states. It would be interesting to apply artificial cfDNA controls used in similar cfDNA studies on paediatric solid tumours (71). Second, the sample sizes are relatively small, and it is unlikely they are fully representative of the population of CRC samples. Lastly, this is a single omic approach that does not consider factors such as DNA methylation or histone modifications. As evidenced previously, it is agreed that CRC has a DNA CpG Island Methylation (CIMP) subtype caused by hypermethylation (3), which influences gene regulation. To successfully predict treatment response from current chromatin states these extra molecular subtypes need to be incorporated into the analysis.

8.8 The implications of accessing the genome through blood liquid biopsies

This study has a wide range of applications because it has been shown to provide inferences about chromatin accessibility for colorectal cancer patients. Within an academic setting it has provided a list of TFs to further investigate. Practically, this work provided an excellent complement to wet lab research and in the future could help to drive hypotheses in new directions by mining large lists of TFs or target defined sites of interest. Wider implications have examined the potential of nucleosome profiling and liquid biopsies in precision medicine. It is important to highlight how liquid biopsies and cell free DNA have improved the accessibility of epigenetic studies analysing the human genome by reducing the need for large, high-cost sequencing projects. Nonetheless, the advancement of this technology does create ethical considerations around who has access to analyse patients' blood using this methodology. For example, is it appropriate to investigate neurodivergent phenotypes where classification as a disease is being debated against a history of ableism in genetic research? (72, 73) Elsewhere, regulations need to be developed around identifying predispositions to genetic diseases when the individual has no obvious pathology. Blood testing is a common diagnostic tool, and it is feasible that extra sequencing will detect unexpected disease phenotypes if incorporated into clinical practice. Finally, false positive rates must be minimised to avoid causing distress to patients, currently a negative liquid biopsy sample does not equate to being cancer free..

8.9 The future need for multi-omic analysis

Future research directions could include analysis of colorectal cancer data with existing sub-typing identified through pathologically defined study groups. This would allow for cancer detections and tumour sub-typing machine learning models to be created as in (22) and re-applied to this CRC liquid biopsy dataset. As mentioned, control samples need to be analysed as a next step to determine non-pathological chromatin states. While Griffin was designed for cancer analysis using low coverage data, it would be interesting to examine other model organisms with more precise genetic manipulation capabilities to study knockouts of highly conserved transcriptional mechanisms (74). There is a whole community of developers creating cfDNA analytical tools and it would be interesting to apply techniques for CNV detection or DNA methylation (75).

8.10 Conclusion

To conclude, Griffin provides a new framework to explore the cancer genome and its complex network of transcriptional regulatory activity. Further research is required to understand the role chromatin accessibility has in colorectal cancer tumour progression and its effect on radiotherapy treatment. As further liquid biopsy samples are sequenced it will be interesting to observe emerging chromatin accessibility trends on a pan-cancer scale.

9 Data availability

Summary of available data in table 4.

10 References

1. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nature reviews Gastroenterology & hepatology. 2019;16(12):713-32.
2. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J. The genomic landscapes of human breast and colorectal cancers. Science. 2007;318(5853):1108-13.
3. Bogaert J, Prenen H. Molecular genetics of colorectal cancer. Annals of gastroenterology. 2014;27(1):9.
4. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, et al. The consensus molecular subtypes of colorectal cancer. Nature Medicine. 2015;21(11):1350-6.
5. Polyak K, Marusyk A. Clonal cooperation. Nature. 2014;508(7494):52-3.
6. Chae YK, Davis AA, Jain S, Santa-Maria C, Flaum L, Beaubier N, Platanias LC, Gradishar W, Giles FJ, Cristofanilli M. Concordance of Genomic Alterations by Next-Generation Sequencing in Tumor Tissue versus Circulating Tumor DNA in Breast CancerConcordance of Tissue and Liquid Biopsies in Breast Cancer. Molecular cancer therapeutics. 2017;16(7):1412-20.
7. Heidrich I, Ačkar L, Mossahebi Mohammadi P, Pantel K. Liquid biopsies: Potential and challenges. International journal of cancer. 2021;148(3):528-45.
8. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, Liu CL, Neal JW, Wakelee HA, Merritt RE. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nature medicine. 2014;20(5):548-54.
9. Rostami A, Lambie M, Yu CW, Stambolic V, Waldron JN, Bratman SV. Senescence, Necrosis, and Apoptosis Govern Circulating Cell-free DNA Release Kinetics. Cell Rep. 2020;31(13):107830.
10. Page K, Guttery DS, Fernandez-Garcia D, Hills A, Hastings RK, Luo J, Goddard K, Shahin V, Woodley-Barker L, Rosales BM, et al. Next Generation Sequencing of Circulating Cell-Free DNA for Evaluating Mutations and Gene Amplification in Metastatic Breast Cancer. Clinical Chemistry. 2017;63(2):532-41.
11. Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science. 2021;372(6538).
12. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, Mair R, Goranova T, Marass F, Heider K. Enhanced detection of circulating tumor DNA by fragment size analysis. Science translational medicine. 2018;10(466):eaat4921.

13. Parikh AR, Leshchiner I, Elagina L, Goyal L, Levovitz C, Siravegna G, Livitz D, Rhrissorrakrai K, Martin EE, Van Seventer EE. Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. Nature medicine. 2019;25(9):1415-21.

14. Struhl K, Segal E. Determinants of nucleosome positioning. Nature Structural & Molecular Biology. 2013;20(3):267-73.

15. Merhi M, Ahmad F, Taib N, Inchakalody V, Uddin S, Shablak A, Dermime S, editors. The complex network of transcription factors, immune checkpoint inhibitors and stemness features in colorectal cancer: A recent update. Seminars in Cancer Biology; 2023: Elsevier.

16. Lowary P, Widom J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. Journal of molecular biology. 1998;276(1):19-42.

17. Arya G, Maitra A, Grigoryev SA. A Structural Perspective on the Where, How, Why, and What of Nucleosome Positioning. Journal of Biomolecular Structure and Dynamics. 2010;27(6):803-20.

18. Shimamoto Y, Tamura S, Masumoto H, Maeshima K. Nucleosome–nucleosome interactions via histone tails and linker DNA regulate nuclear rigidity. Molecular biology of the cell. 2017;28(11):1580-9.

19. Shi Y, Su X-B, He K-Y, Wu B-H, Zhang B-Y, Han Z-G. Chromatin accessibility contributes to simultaneous mutations of cancer genes. Scientific Reports. 2016;6(1):1-12.

20. Heide T, Househam J, Cresswell GD, Spiteri I, Lynn C, Mossner M, Kimberley C, Fernandez-Mateos J, Chen B, Zapata L. The co-evolution of the genome and epigenome in colorectal cancer. Nature. 2022:1-11.

21. Deniz Ö, Flores O, Aldea M, Soler-López M, Orozco M. Nucleosome architecture throughout the cell cycle. Scientific Reports. 2016;6(1):19729.

22. Doebley A-L, Ko M, Liao H, Cruikshank AE, Santos K, Kikawa C, Hiatt JB, Patton RD, De Sarkar N, Collier KA. A framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA. Nature Communications. 2022;13(1):1-18.

23. Mazzonetto PC, Villela D, da Costa SS, Krepischi AC, Milanezi F, Migliavacca MP, Pierry PM, Bonaldi A, Almeida LGD, de Souza CA. Low-pass whole genome sequencing is a reliable and cost-effective approach for copy number variant analysis in the clinical setting. medRxiv. 2023:2023.05. 26.23290606.

24. Yevshin I, Sharipov R, Kolmykov S, Kondrakhin Y, Kolpakov F. GTRD: a database on gene transcription regulation—2019 update. Nucleic acids research. 2019;47(D1):D100-D5.

25. Dreos R, Ambrosini G, Bucher P. Influence of rotational nucleosome positioning on transcription start site selection in animal promoters. PLoS computational biology. 2016;12(10):e1005144.

26. Kubik S, Bruzzone MJ, Shore D. Establishing nucleosome architecture and stability at promoters: Roles of pioneer transcription factors and the RSC chromatin remodeler. Bioessays. 2017;39(5):1600237.

27. Morris VK, Strickler JH. Use of Circulating Cell-Free DNA to Guide Precision Medicine in Patients with Colorectal Cancer. Annual Review of Medicine. 2021;72(1):399-413.

28. Tie J, Wang Y, Tomasetti C, Li L, Springer S, Kinde I, Silliman N, Tacey M, Wong H-L, Christie M. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. Science translational medicine. 2016;8(346):346ra92-ra92.

29. Hsu H-C, Lapke N, Wang C-W, Lin P-Y, You JF, Yeh CY, Tsai W-S, Hung HY, Chiang S-F, Chen H-C. Targeted sequencing of circulating tumor DNA to monitor genetic variants and therapeutic response in metastatic colorectal cancer. Molecular cancer therapeutics. 2018;17(10):2238-47.

30. Corcoran RB, André T, Atreya CE, Schellens JH, Yoshino T, Bendell JC, Hollebecque A, McRee AJ, Siena S, Middleton G. Combined BRAF, EGFR, and MEK Inhibition in Patients with BRAFV600E-Mutant Colorectal CancerBRAF/EGFR/MEK Inhibition in BRAFV600E Colorectal Cancer. Cancer discovery. 2018;8(4):428-43.

31. Network CGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330.

32. Tang WW, Yimer H, Tummala M, Shao S, Chung G, Clement J, Chu BC, Hubbell E, Kurtzman KN, Swanton C. Performance of a targeted methylation-based multi-cancer early detection test by race and ethnicity. Preventive Medicine. 2023;167:107384.

33. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884-i90.

34. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.

35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

36. Van der Auwera GA, O'Connor BD. Genomics in the cloud: using Docker, GATK, and WDL in Terra: O'Reilly Media; 2020.

37. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847-9.

38. Chen HY, Feng LL, Li M, Ju HQ, Ding Y, Lan M, Song SM, Han WD, Yu L, Wei MB, et al. College of American Pathologists Tumor Regression Grading System for Long-Term Outcome in Patients with Locally Advanced Rectal Cancer. Oncologist. 2021;26(5):e780-e93.

39. Wang Y, LêCao K-A. Managing batch effects in microbiome data. Briefings in bioinformatics. 2020;21(6):1954-70.

40. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR genomics and bioinformatics. 2020;2(3):lqaa078.

41. Wasylyk C, Zheng H, Castell C, Debussche L, Multon M-C, Wasylyk B. Inhibition of the Ras-Net (Elk-3) pathway by a novel pyrazole that affects microtubules. Cancer research. 2008;68(5):1275-83.

42. Cui J, Liu J, Fan L, Zhu Y, Zhou B, Wang Y, Hua W, Wei W, Sun G. A zinc finger family protein, ZNF263, promotes hepatocellular carcinoma resistance to apoptosis via activation of ER stress-dependent autophagy. Translational oncology. 2020;13(12):100851.

43. Song P, Wu LR, Yan YH, Zhang JX, Chu T, Kwong LN, Patel AA, Zhang DY. Limitations and opportunities of technologies for the analysis of cell-free DNA in cancer diagnostics. Nature Biomedical Engineering. 2022;6(3):232-45.

44. Burnham P, Dadhania D, Heyang M, Chen F, Westblade LF, Suthanthiran M, Lee JR, De Vlaminck I. Urinary cell-free DNA is a versatile analyte for monitoring infections of the urinary tract. Nature communications. 2018;9(1):2412.

45. Tug S, Helmig S, Ricarda Deichmann E, Schmeier-Jürchott A, Wagner E, Zimmermann T, Radsak M, Giacca M, Simon P. Exercise-induced increases in cell free DNA in human plasma originate predominantly from cells of the haematopoietic lineage. Exercise immunology review. 2015;21.

46. Lai Q, Li Q, He C, Fang Y, Lin S, Cai J, Ding J, Zhong Q, Zhang Y, Wu C, et al. CTCF promotes colorectal cancer cell proliferation and chemotherapy resistance to 5-FU via the P53-Hedgehog axis. Aging (Albany NY). 2020;12(16):16270-93.

47. Batlle-López A, Cortiguera MG, Rosa-Garrido M, Blanco R, del Cerro E, Torrano V, Wagner SD, Delgado MD. Novel CTCF binding at a site in exon1A of BCL6 is associated with active histone marks and a transcriptionally active locus. Oncogene. 2015;34(2):246-56.

48. Guo H-Q, Huang G-L, Liu O-F, Liu Y-Y, Yao Z-H, Yao S-N, Zhao Y, Liu T, Pu X-X, Lin T-Y. p63 Expression is a prognostic factor in colorectal cancer. The International journal of biological markers. 2012;27(3):212-8.

49. Liu Y, Sun J, Zhao M. ONGene: A literature-based database for human oncogenes. J Genet Genomics. 2017;44(2):119-21.

50. Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic acids research. 2016;44(D1):D1023-D31.

51. Yang CM, Kang M-K, Jung W-J, Joo J-S, Kim Y-J, Choi Y, Kim H-P. p53 expression confers sensitivity to 5-fluorouracil via distinct chromatin accessibility dynamics in human colorectal cancer. Oncology Letters. 2021;21(3):1-.

52. Cooper M, Ray A, Bhattacharya A, Dhasarathy A, Takaku M. ATAC-seq Optimization for Cancer Epigenetics Research. JoVE (Journal of Visualized Experiments). 2022(184):e64242.

53. Hsu YL, Huang PY, Chen DT. Sparse principal component analysis in cancer research. Transl Cancer Res. 2014;3(3):182-90.

54. Lenz M, Müller F-J, Zenke M, Schuppert A. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. Scientific Reports. 2016;6(1):25696.

55. Lenz M, Schuldt BM, Müller F-J, Schuppert A. PhysioSpace: relating gene expression experiments from heterogeneous sources using shared physiological processes. PLoS One. 2013;8(10):e77627.

56. Schneckener S, Arden NS, Schuppert A. Quantifying stability in gene list ranking across microarray derived clinical biomarkers. BMC medical genomics. 2011;4(1):1-11.

57. Lou RN, Therkildsen NO. Batch effects in population genomic studies with low‐coverage whole genome sequencing data: Causes, detection and mitigation. Molecular Ecology Resources. 2022;22(5):1678-92.

58. Zindler T, Frieling H, Neyazi A, Bleich S, Friedel E. Simulating ComBat: how batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies. BMC bioinformatics. 2020;21:1-15.

59. Kim S-E, Paik HY, Yoon H, Lee JE, Kim N, Sung M-K. Sex-and gender-specific disparities in colorectal cancer risk. World journal of gastroenterology: WJG. 2015;21(17):5167.

60. Ulz P, Perakis S, Zhou Q, Moser T, Belic J, Lazzeri I, Wölfler A, Zebisch A, Gerger A, Pristauz G. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nature communications. 2019;10(1):4666.

61. Sweet LE, Moulaison HL. Electronic health records data and metadata: challenges for big data in the United States. Big data. 2013;1(4):245-51.

62. Han F, Zhang L, Liao S, Zhang Y, Qian L, Hou F, Gong J, Lai M, Zhang H. The interaction between S100A2 and KPNA2 mediates NFYA nuclear import and is a novel therapeutic target for colorectal cancer metastasis. Oncogene. 2022;41(5):657-70.

63. Li H, Zhang J, Tong JHM, Chan AWH, Yu J, Kang W, To KF. Targeting the oncogenic p53 mutants in colorectal cancer and other solid tumors. International journal of molecular sciences. 2019;20(23):5999.

64. Schou J, Larsen F, Sørensen B, Abrantes R, Boysen A, Johansen J, Jensen B, Nielsen D, Spindler K. Circulating cell-free DNA as predictor of treatment failure after neoadjuvant chemo-radiotherapy before surgery in patients with locally advanced rectal cancer. Annals of Oncology. 2018;29(3):610-5.

65. Kleppe A, Albregtsen F, Vlatkovic L, Pradhan M, Nielsen B, Hveem TS, Askautrud HA, Kristensen GB, Nesbakken A, Trovik J. Chromatin organisation and cancer prognosis: a pan-cancer study. The Lancet Oncology. 2018;19(3):356-69.

66. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, Jensen SØ, Medina JE, Hruban C, White JR. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570(7761):385-9.

67. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, Gydush G, Reed SC, Rotem D, Rhoades J, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature Communications. 2017;8(1):1324.

68. Ando M, Saito Y, Xu G, Bui NQ, Medetgul-Ernar K, Pu M, Fisch K, Ren S, Sakai A, Fukusumi T, et al. Chromatin dysregulation and DNA methylation at transcription start sites associated with transcriptional repression in cancers. Nature Communications. 2019;10(1):2188.

69. Farman FU, Iqbal M, Azam M, Saeed M. Nucleosomes positioning around transcriptional start site of tumor suppressor (Rbl2/p130) gene in breast cancer. Molecular Biology Reports. 2018;45(2):185-94.

70. Hesson LB, Sloane MA, Wong JW, Nunez AC, Srivastava S, Ng B, Hawkins NJ, Bourke MJ, Ward RL. Altered promoter nucleosome positioning is an early event in gene silencing. Epigenetics. 2014;9(10):1422-30.

71. Stankunaite R, George SL, Gallagher L, Jamal S, Shaikh R, Yuan L, Hughes D, Proszek PZ, Carter P, Pietka G. Circulating tumour DNA sequencing to determine therapeutic response and identify tumour heterogeneity in patients with paediatric solid tumours. European Journal of Cancer. 2022;162:209-20.

72. Chiapperino L, Hens K. How to talk about autism: reconciling genomics and neurodiversity. Nature Medicine. 2023;29(7):1607-8.

73. Green S, Prainsack B, Sabatello M. Precision medicine and the problem of structural injustice. Medicine Health Care and Philosophy. 2023.

74. Raveh-Sadka T, Levo M, Shabi U, Shany B, Keren L, Lotan-Pompan M, Zeevi D, Sharon E, Weinberger A, Segal E. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nature genetics. 2012;44(7):743-50.

75. Huang C-C, Du M, Wang L. Bioinformatics analysis for circulating cell-free DNA in cancer. Cancers. 2019;11(6):805.

76. Chen X, Yang H, Liu G, Zhang Y. NUCOME: A comprehensive database of nucleosome organization referenced landscapes in mammalian genomes. BMC Bioinformatics. 2021;22(1):321.

11 Supplementary material

Supplementary data and scripts are available in the associated supplementary.zip and will be added to github repository . GlenRoarke/Bioinformatics_Project: A repository for code used in my research project analysing circulating tumour DNA (github.com)

Supplementary	Description
01_Supplementary_script_preprocessing_stats.html	Generating visualisation of the ctDNA preprocessing
02_Supplementary_script_coverage_ALL_samples.html	Creating visualisations of all sample and batch effects (figure 2)
03_Supplementary_script_HQ_samples_.html	Central coverage and high quality samples nucleosome profiles (figure 3 & 4, tables 2 & 2.2)
04_Supplementary_script_PCA_analysis_CAP-TRG.html	PCA analysis (Figure 5)
05_Supplementary_script_batch_correction.html	Visualisaton of batch effects and correction attempt (figures 6.1- 6.2)
06_Supplementary_script_create_TSS_sites.html	Creates the new sites files for Griffin
07_supplementary_Script_TSS_analysis.html	TSS plots Figure 7
08_supplementary_Script_correlation_cap-trg.html	Spearmans rank correlation of TFs against CAP-TRG (clinical indicator)
09_supplementary_script_batch_corrected_heatmaps.R	Heatmaps of Combat corrections from 05_supplementary_script.
Supplementary_data_1_lowest_central_coverage_TFs.csv	More accessible chromatin.
Supplementary_data_2_highest_central_coverage_TFs.csv	Less accessible chromatin.
08_Supplementary_correlation_analysis.csv	CAP-TRG and TF correlations
Supplementary RData files	For R data frames and objects.
Figure 10 & 11	Biology overview of TFBS and TSS.

12 Figures and Tables

12.1 Tables

Software	Version	Link
fastp	v0.21.0	OpenGene/fastp: An ultra-fast all-in-one FASTQ preprocessor (QC/adapters/trimming/filtering/splitting/merging...) (github.com)
BWA	v0.7.17	Burrows-Wheeler Aligner (sourceforge.net)
Samtools	v1.12	Samtools (htslib.org)
GATK	v4.2.3.0	GATK (broadinstitute.org)
Griffin	v0.1.0	adoebley/Griffin: A flexible framework for nucleosome profiling of cell-free DNA (github.com)
Griffin TSS development	v1.0	Griffin/snakemakes/griffin_nucleosome_profiling/config/config_TSS.yaml at development · adoebley/Griffin (github.com)
Preprocessing tools	v1.0	Bioinformatics_Project/preprocessing_scripts at main · GlenRoarke/Bioinformatics_Project (github.com)
Merge Griffin outputs	V1.0	Bioinformatics_Project/griffin_analysis/merge_outputs at main · GlenRoarke/Bioinformatics_Project (github.com)

Table 1 – A summary of software

Sequencing batch	Total samples	Final samples
batch 1	25	15
batch 2	25	9
batch 3	33	10

Table 2 – Summary of high quality samples after exclusion rules are applied. Only samples with a read mapping proportion over 0.8 were selected. Sample As were investigated in more detail to assess pre-radiotherapy treatment chromatin profiles.

ENTREZID	SYMBOL	GENENAME	UNIPROT
604	BCL6	BCL6 transcription repressor	P41182
10664	CTCF	CCCTC-binding factor	P49711
2305	FOXM1	forkhead box M1	A8K591
2305	FOXM1	forkhead box M1	Q53Y49
3659	IRF1	interferon regulatory factor 1	P10914
3662	IRF4	interferon regulatory factor 4	Q15306
83855	KLF16	KLF transcription factor 16	Q9BXK1
4778	NFE2	nuclear factor, erythroid 2	Q16621
5993	RFX5	regulatory factor X5	P48382
6688	SPI1	Spi-1 proto-oncogene	P17947
6722	SRF	serum response factor	B4DU24
6772	STAT1	signal transducer and activator of transcription 1	P42224
30009	TBX21	T-box transcription factor 21	Q9UL17
57621	ZBTB2	zinc finger and BTB domain containing 2	Q8N680
80345	ZSCAN16	zinc finger and SCAN domain containing 16	Q9H4T2

Table 3 – Lowest coverage transcription factors in all batches

An example subset of the TFs inferred to have the highest chromatin accessibility or lowest central coverage across all three batches. Transcription factor structural variation is included via uniprot ids. Full data is available in files Supplementary_data_1

ENTREZID	SYMBOL	GENENAME	UNIPROT
429	ASCL1	achaete-scute family bHLH transcription factor 1	P50553
2034	EPAS1	endothelial PAS domain protein 1	B3KW07
2034	EPAS1	endothelial PAS domain protein 1	Q99814
2099	ESR1	estrogen receptor 1	G4XH65
2099	ESR1	estrogen receptor 1	P03372
2100	ESR2	estrogen receptor 2	F1D8N3
2100	ESR2	estrogen receptor 2	Q92731
2116	ETV2	ETS variant transcription factor 2	K7ERX2
8928	FOXH1	forkhead box H1	O75593
148979	GLIS1	GLIS family zinc finger 1	Q8NBF1
9464	HAND2	heart and neural crest derivatives expressed 2	P61296
4654	MYOD1	myogenic differentiation 1	P15172
4656	MYOG	myogenin	P15173
8626	TP63	tumor protein p63	Q9H3D4
8626	TP63	tumor protein p63	A0A0S2Z4N5
7161	TP73	tumor protein p73	O15350
7490	WT1	WT1 transcription factor	Q6PI38
7494	XBP1	X-box binding protein 1	P17861
7546	ZIC2	Zic family member 2	O95409
7705	ZNF146	zinc finger protein 146	Q15072

Table 3.1 - Highest coverage transcription factors in all batches

Subset of the TFs inferred to have the lowest chromatin accessibility and highest central coverage across all three batches. Transcription factor structural variation is included via uniprot ids. Supplementary_data_2

Data	
ctDNA fastq	Healthcare data not publicly available.
Eukaryotic promoter database	https://ccg.epfl.ch/mga/hg38/epd/Hs_EPDnew_006_hg38.sga.gz
CRC driver gene list	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9684080/bin/41586_2022_5202_MOESM8_ESM.xlsx

Table 4 - Data availability

A summary of new datasets used in this study.

12.2 Figures

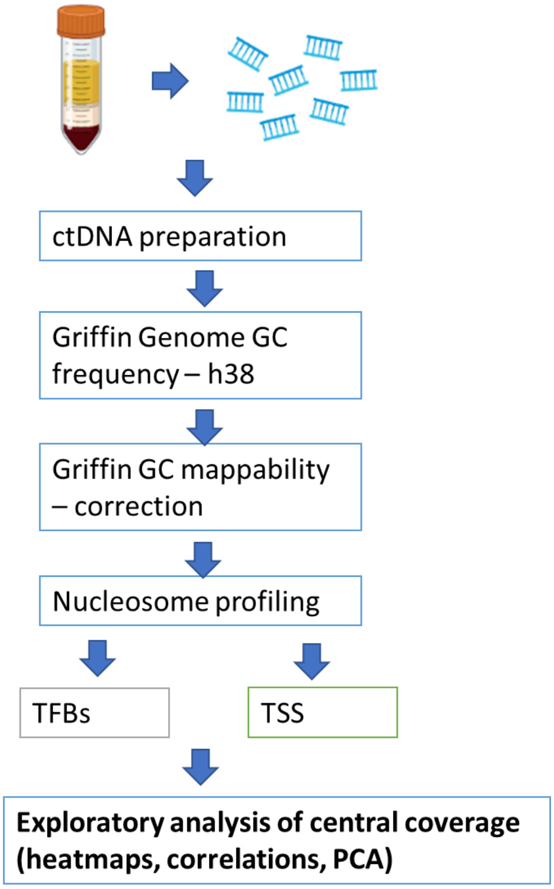
ctDNA preparation
Griffin Genome GC frequency – h38
Griffin GC mappability – correction
Nucleosome profiling
TFBs
TSS
Exploratory analysis of central coverage (heatmaps, correlations, PCA)

Figure 1 ctDNA nucleosome profiling pipeline

Diagram of ctDNA pipeline. Biomarkers are extracted from the blood plasma and sequenced. Raw sequencing reads then undergo the ctDNA preparation step to create a BAM file. The Griffin software corrects for GC bias on the h38 reference genome and then each cfDNA fragment. Either TFBSs or TSS (new development) can be used to generate quantitative measures of chromatin accessibility. R was used in the final data analysis stages for heat map visualisations, PCA and correlation analysis.

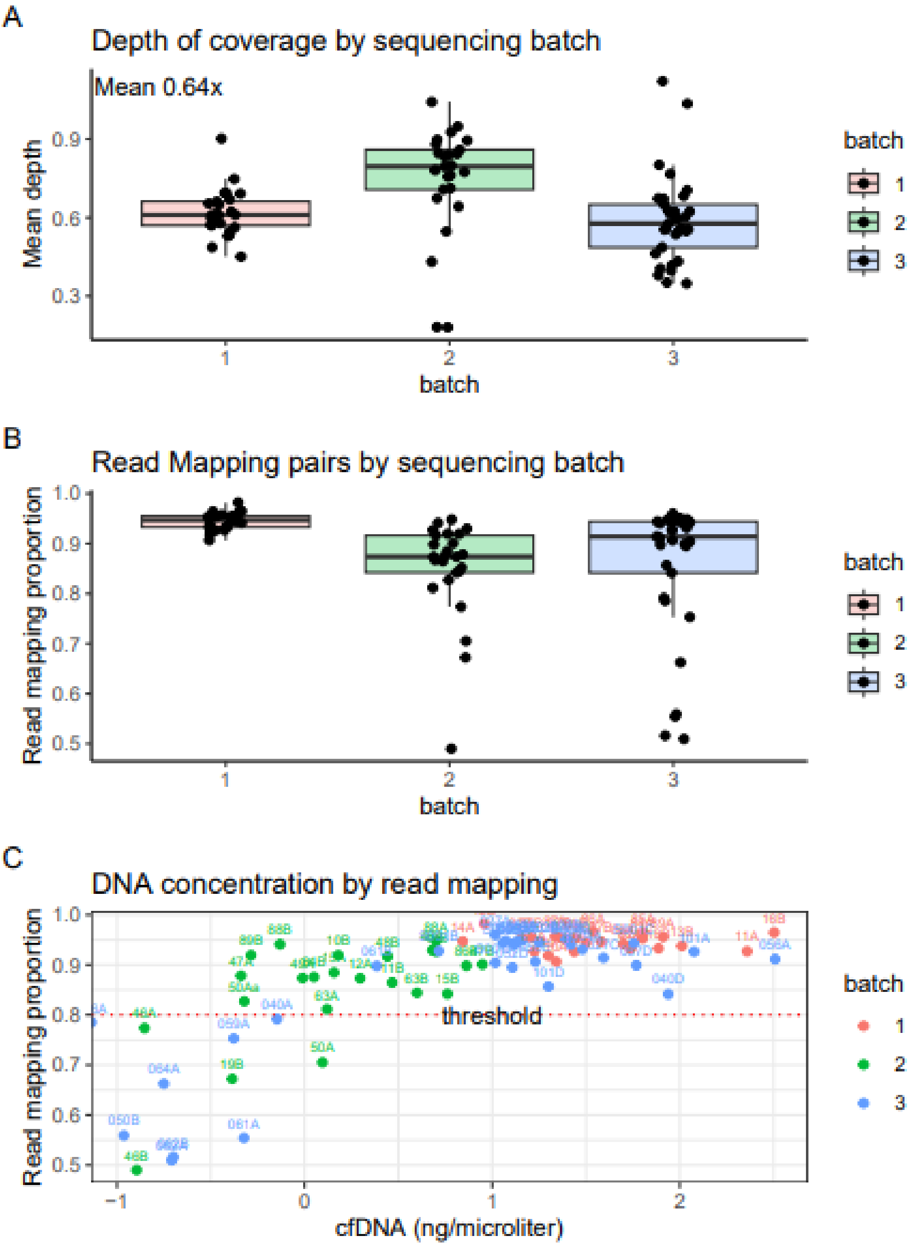

Figure 2 – ctDNA pre-processing analysis identifies samples quality issues.

A – Depth of coverage categorised by sequencing batch (ULP-WGS).

B – Read mapping proportion categorised by sequencing batch .

C – Read mapping proportion positively correlated with logged cfDNA concentration.

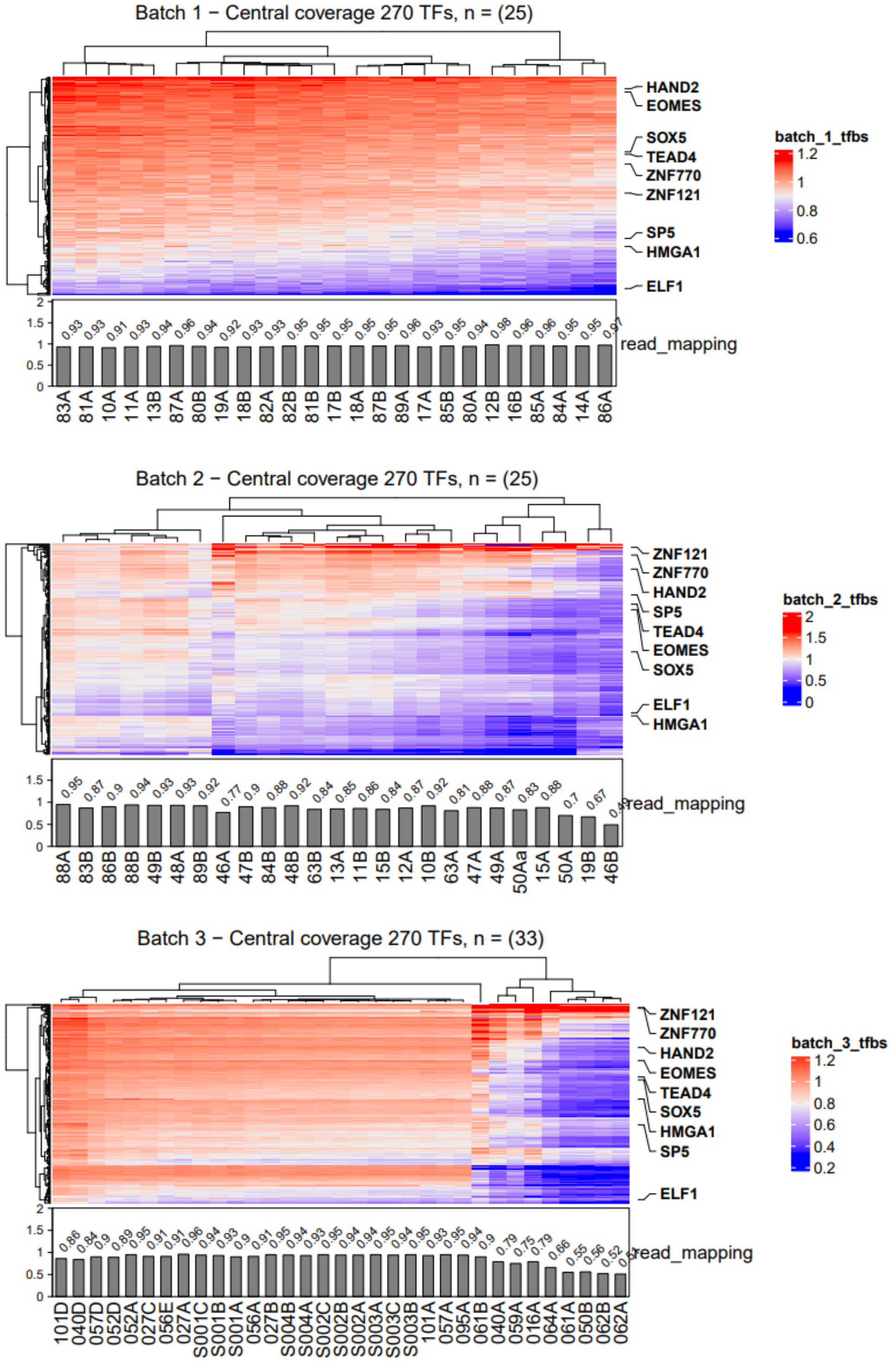

Figure 3 – Griffin Nucleosome profiling across batches 1-3

Heatmaps visualising all 270 TFs in Griffin sites lists annotated with the read mapping proportion of each patient sample against the h38 reference. The default hierarchal clustering mode of ecludician distance was used.The annotated TFs on the right are selected for every 30th TF in the list of 270 to improve clarity, they are not significant.

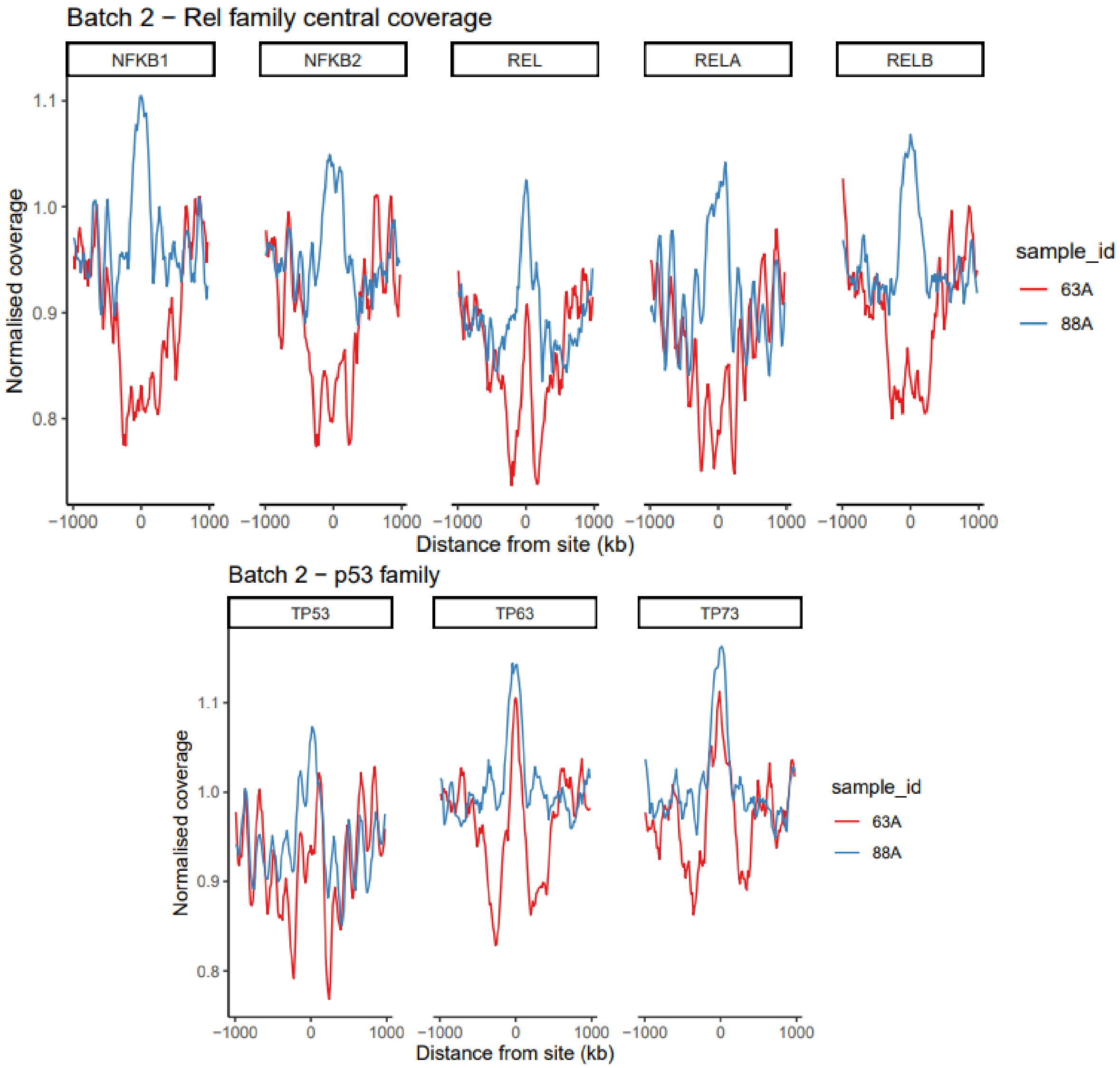

Figure 4 – Exploring proto-oncogene and tumour suppressor chromatin accessibly before treatment.

Inferences about chromatin accessibility for CRC transcription factors such as the Rel and P53 families are shown for patient 63 and 88. A low central coverage value reflects a more accessible or open region of chromatin. Normalised central coverage is plotted in a 1000bp+/ window.

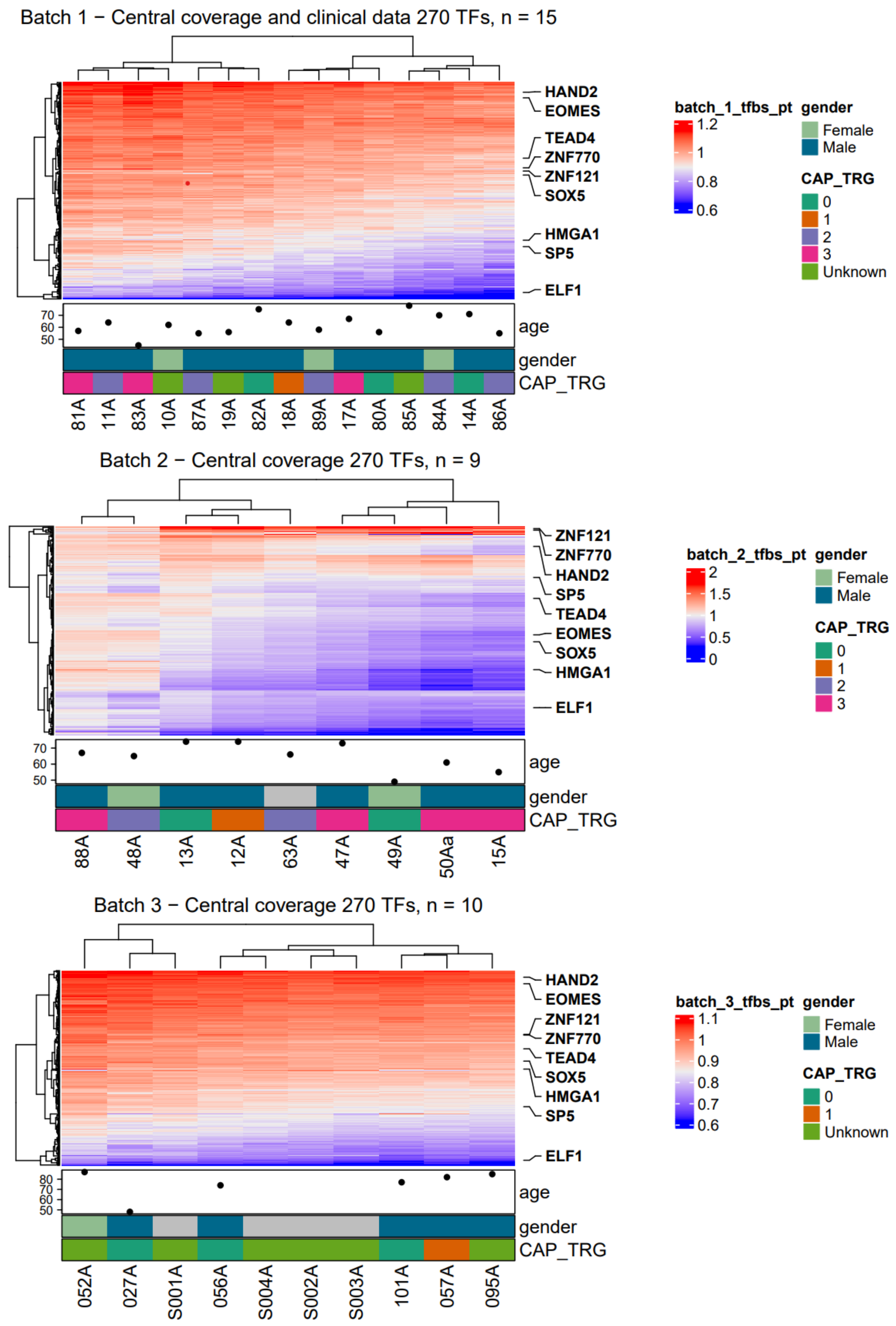

Figure 5 – Central Coverage and patient metadata before treatment.

Patient metadata has been added to central coverage profiles with age, gender and CAP-TRG important variables to consider. For gender the grey colour represents a null value, for CAP-TRG null values are green.

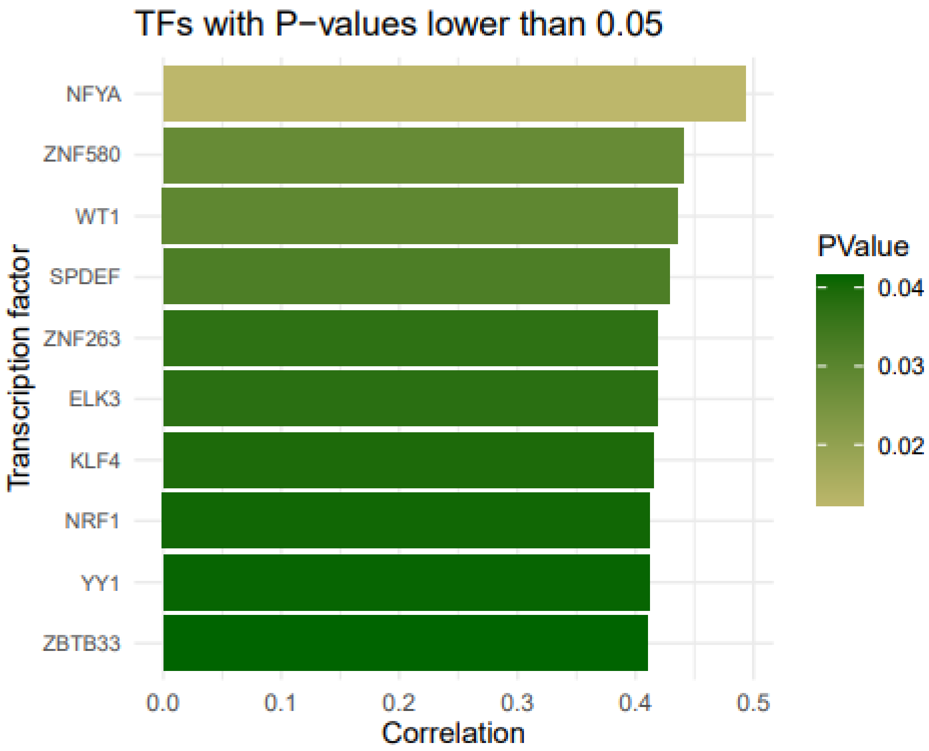

Figure 5.1 – Correlation analysis of CAP-TRG and central coverage variables. Moderate positive correlations have been identified between CAP-TRG and TF central coverage. TFs with statistically significant p-values (< 0.05) are summarised across all three sequencing batches.

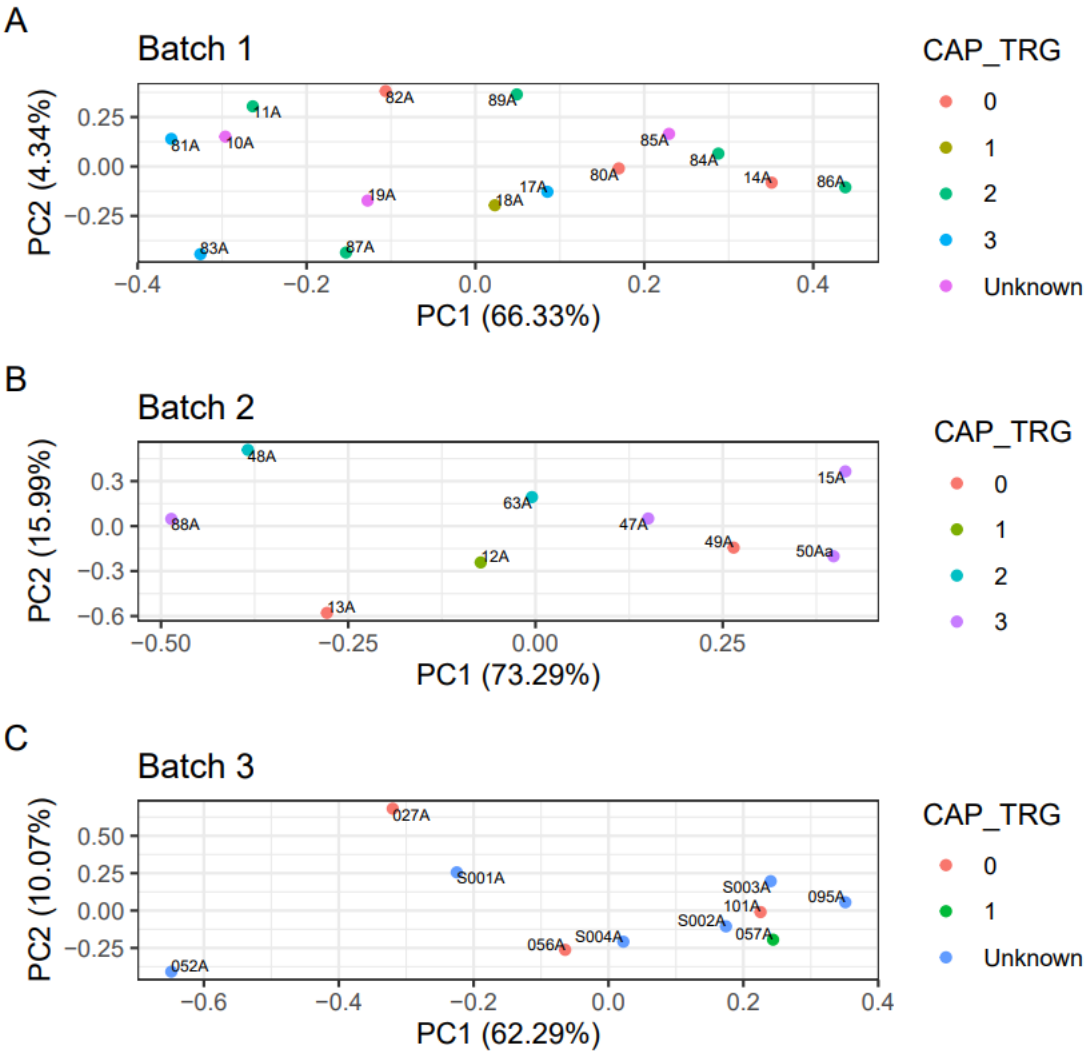

Figure 6 - PCA analysis of CAP-TRG indicator

Ranked treatment response indicator CAP-TRG was added to the PCA analysis across each sequencing batch represented by the colour legend. 0 is a positive treatment response and 3 is a worse response. There are a large quantity of missing values in batch 3 due to data quality issues.

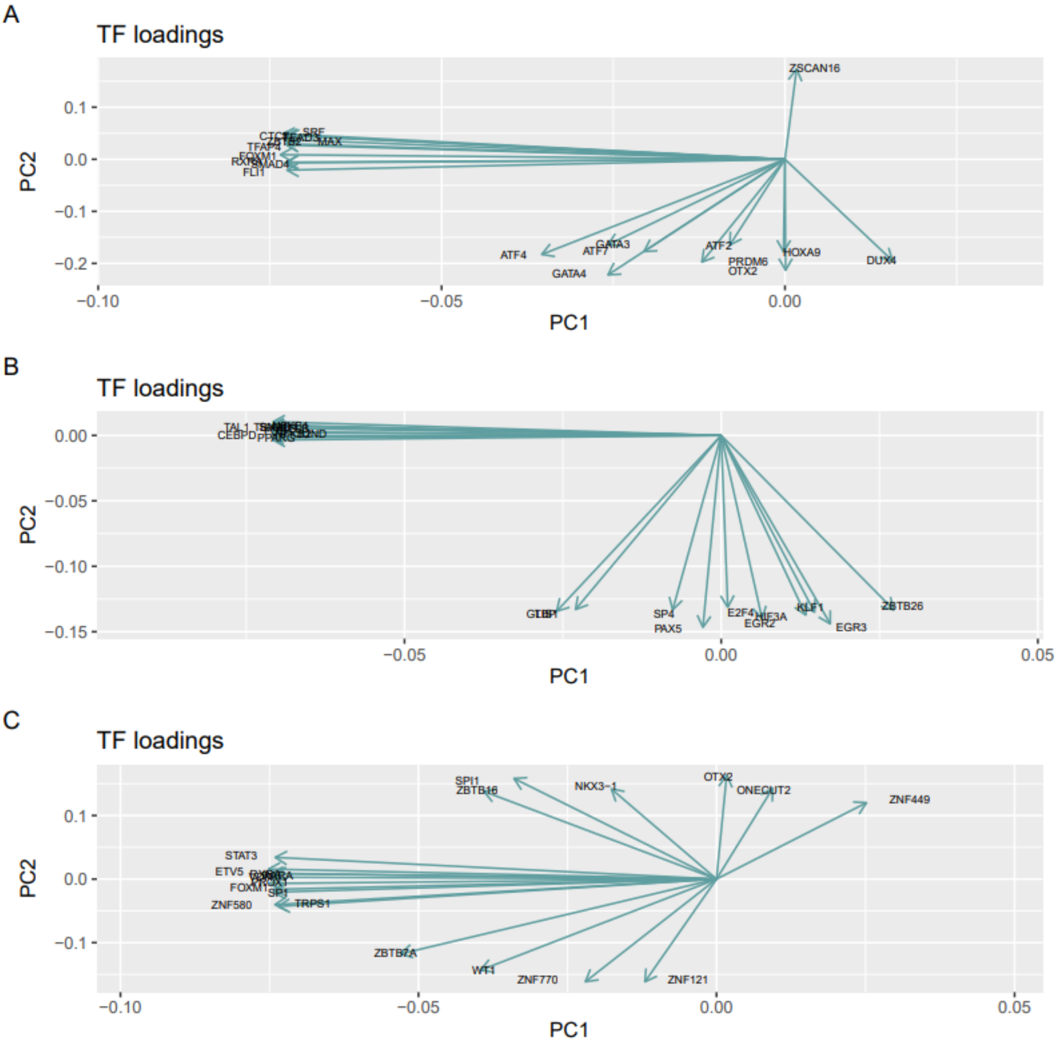

Figure 6.1 - Transcription factor loading for each sequencing batch.

Batches 1-3 are equal to A-C in sequential order. TFs are responsible different levels of principal component variation. These loading are connected to the same PCA analysis in figure 8.

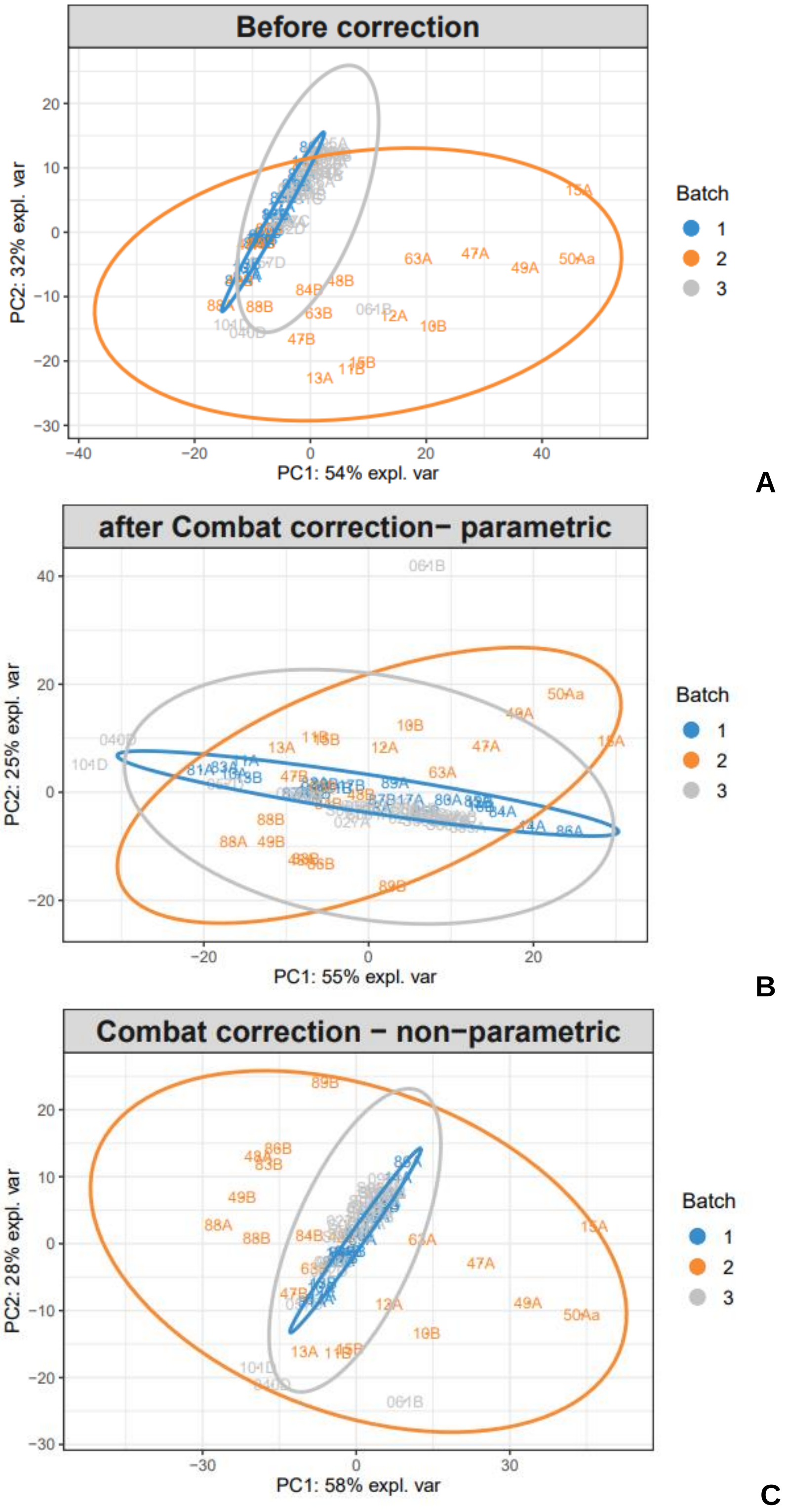

Figure 6.2 – Evaluation of batch effect correction method using Combat

First two principle components before and after batch correction. All three batches do overlap slightly more after correction. B is parametric and C is non-parametric correction.

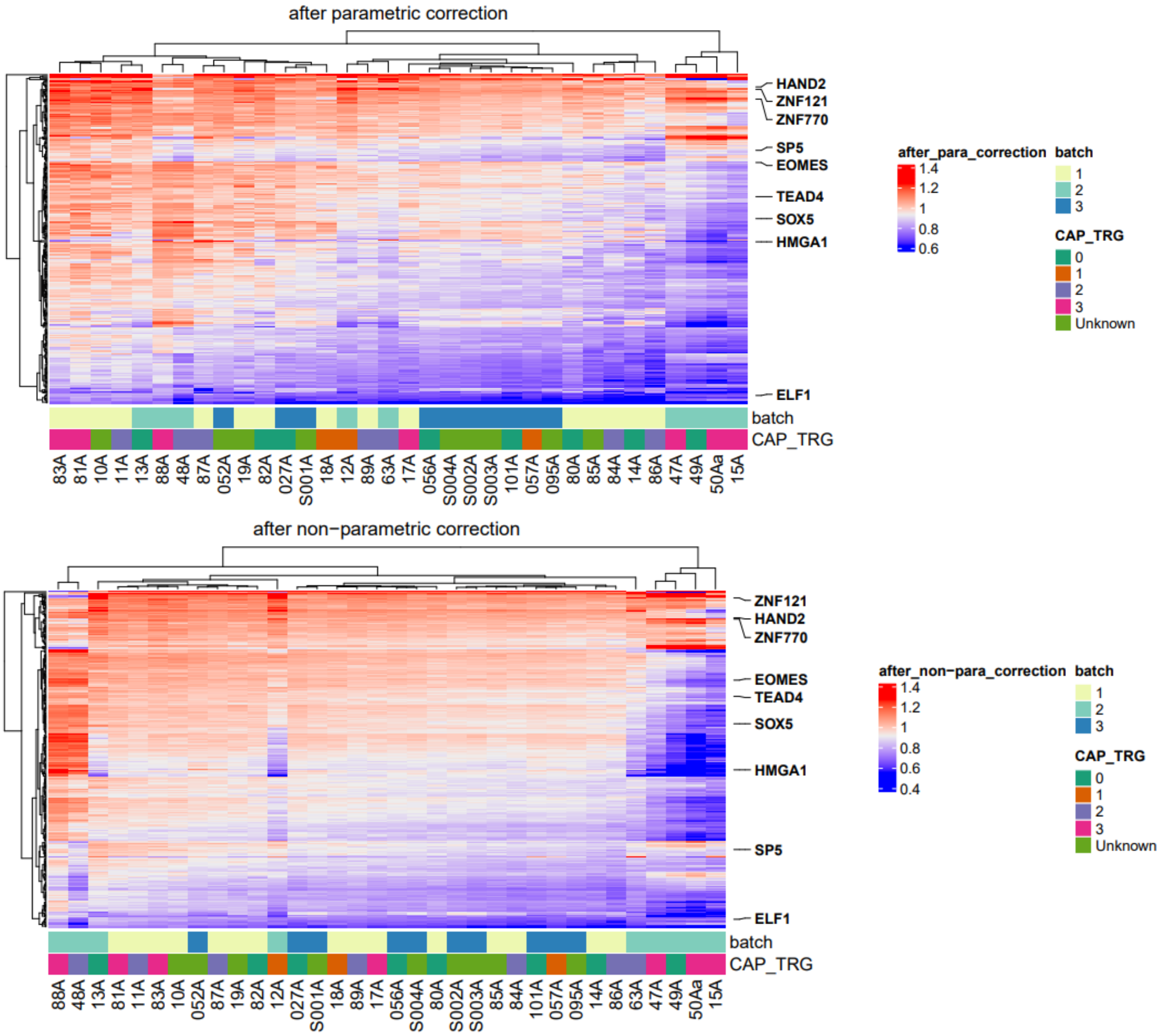

Figure 6.3 – Heatmap after parametric and non-parametric Combat correction

All high quality samples in batches 1-3 are displayed. Chromatin accessibility similarity can be observed across all batches for different correction types.

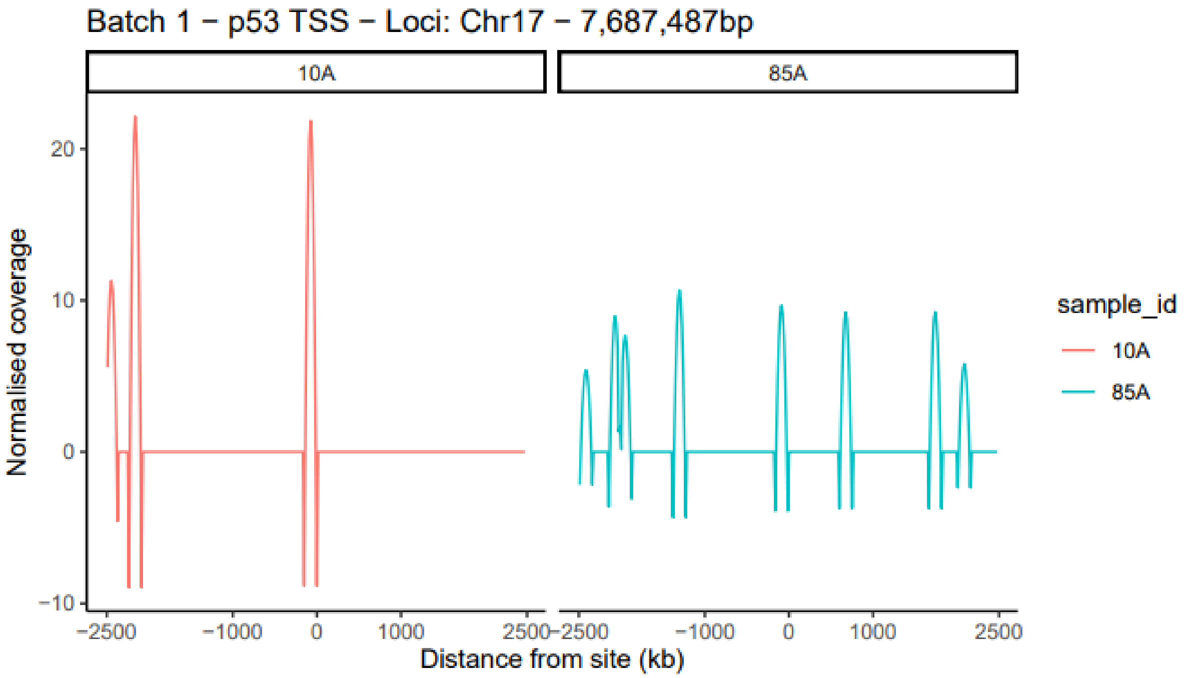

Figure 7 – Nucleosome positioning at the p53 gene.

Selection of the p53 transcription factor was used to observe nucleosome positioning at the TSS. Nucleosome peaks have been detected in a phased structure that is conserved in eukaryotes. Patient samples were selected based on sequencing coverage occurring at the TSS.

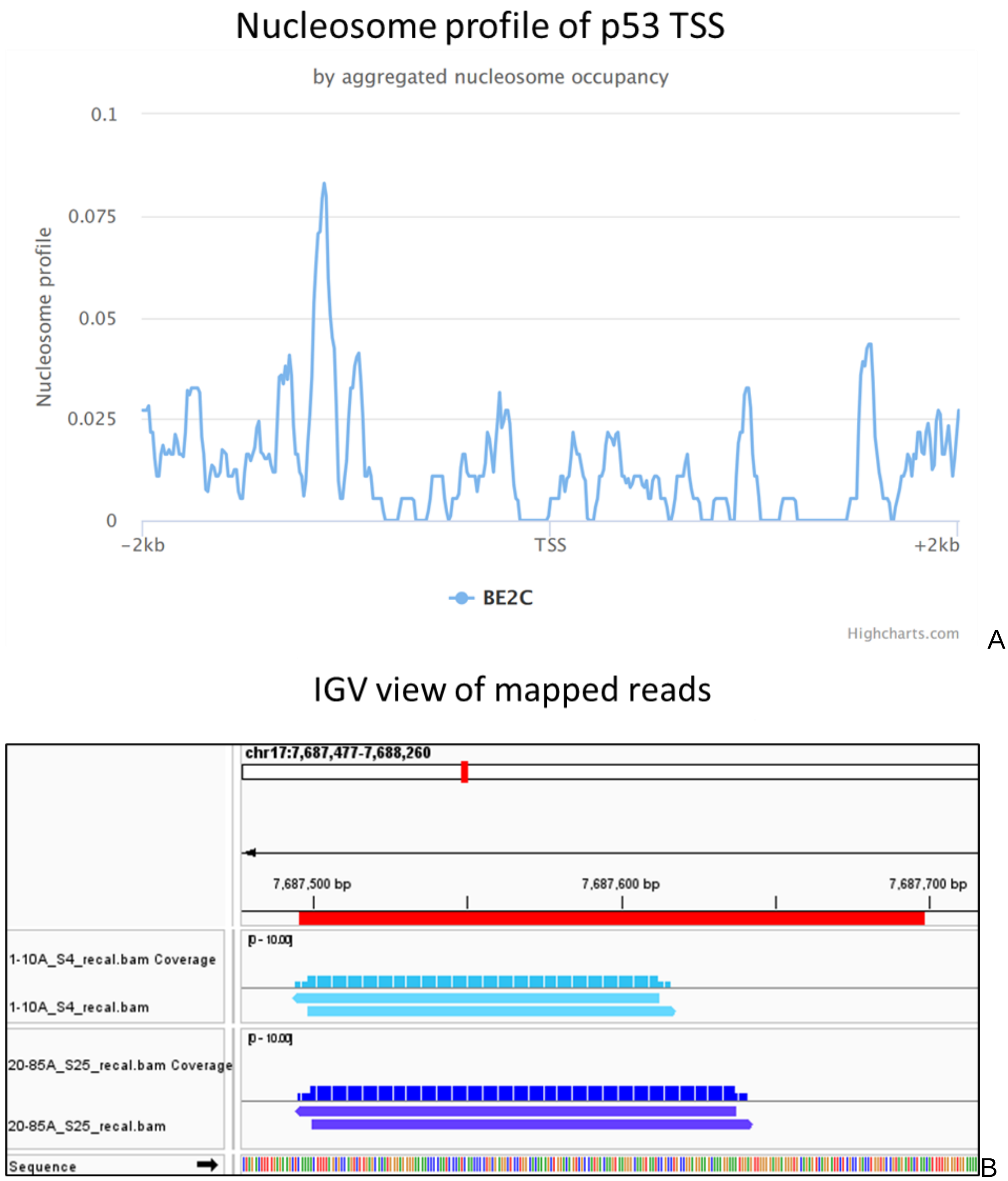

Figure 8 – p53 Validation of TSS nucleosome profiling

A –Nucome provides experimentally validated nucleosome peaks for any genomic region. There is a nucleosome free region present for p53.

B – Searching for samples with coverage mapped at the TSS for the p53 gene (76).

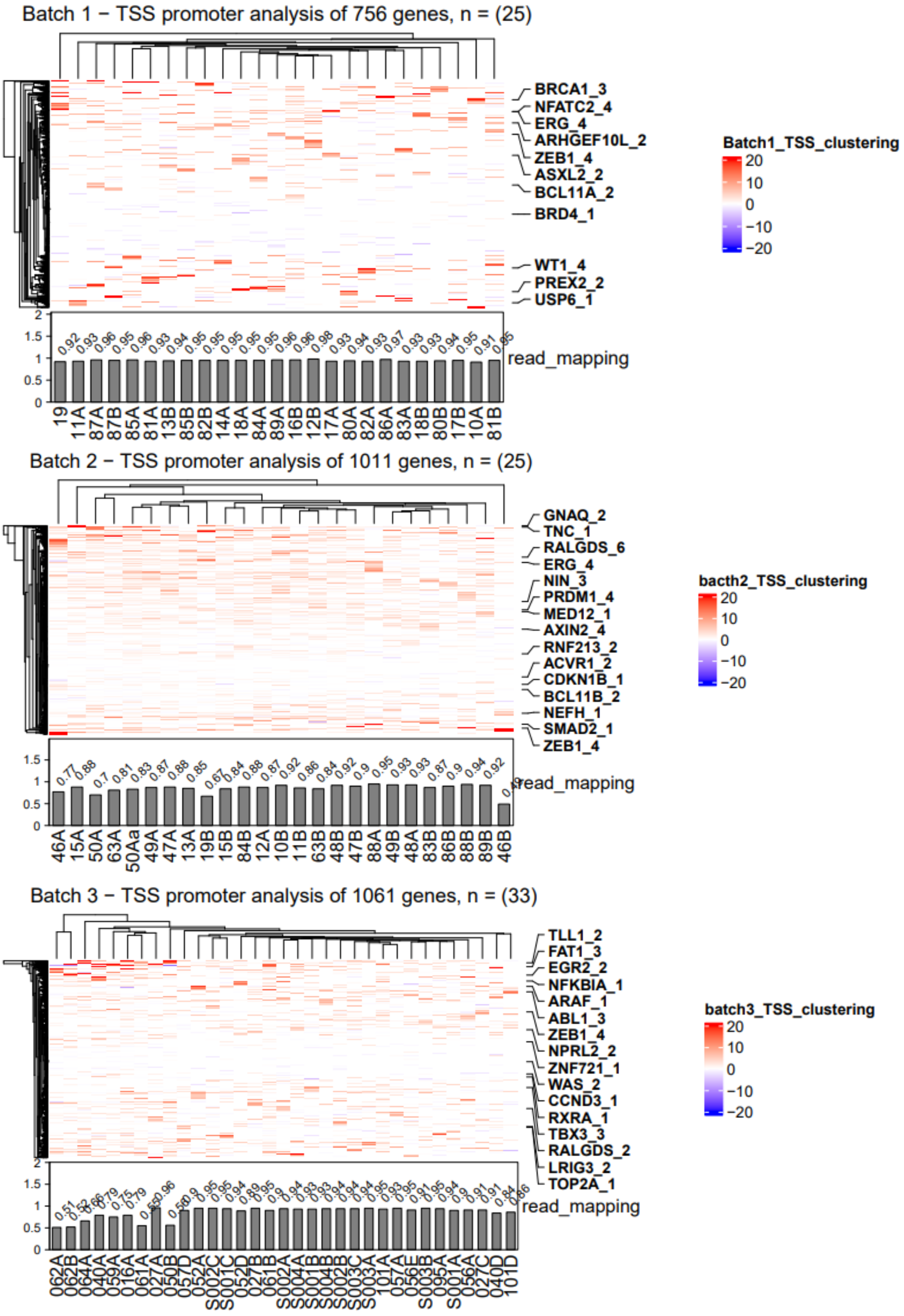

Figure 9 – Nucleosome profiling at colorectal cancer genes.

An analysis of 1,116 TSS sites associated with CRC confirms a wide range of nucleosome peaks. Any rows with zero coverage across all samples have been filtered out from the analysis. Read mapping proportion has been annotated below.

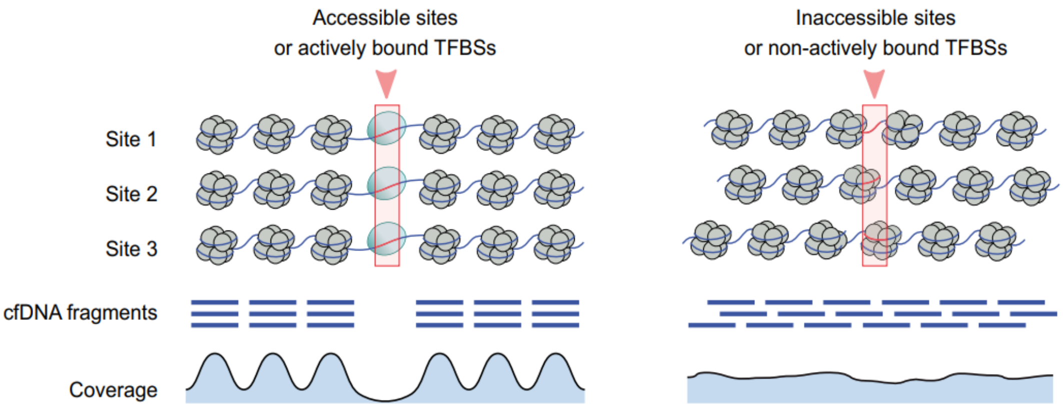

Figure 10 – Overview of nucleosome positioning and its effect on coverage(22).

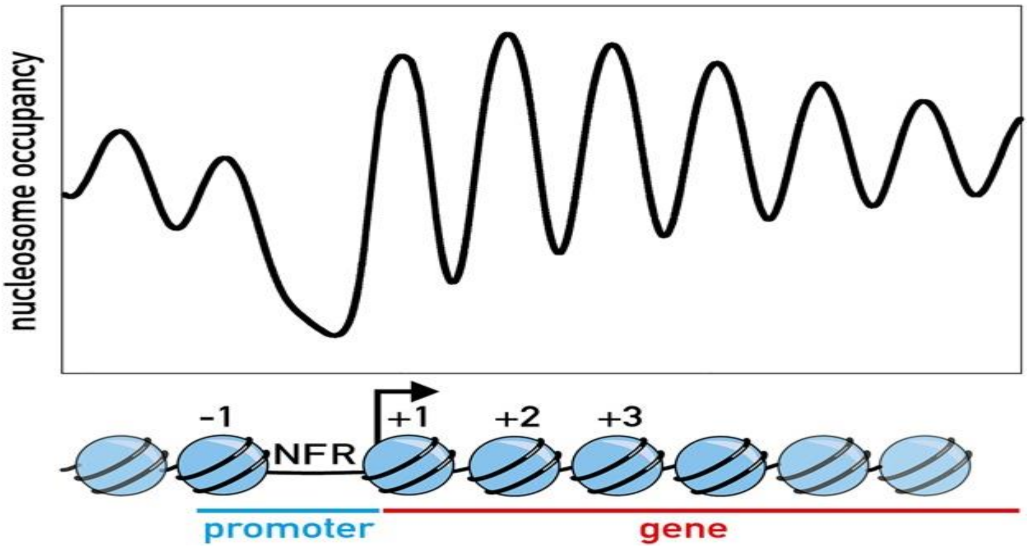

Figure 11 – Highly conserved nucleosome structure at TSS sites (26)