Cranfield University

Jacob Cozens

Characterisation of the transcriptional potential of intergenic CCG-repeats

Cranfield Health

MSc Applied Bioinformatics
2010 – 2011

Supervisor: Dr Christopher Jones
Institute for Women's Health, UCL, London

August 2011

Abstract

CCG repeats have been found to have a non random distribution throughout the human genome. Previous analysis of ChIP-seq data from 800 intergenic CCG-repeats shows that they are associated with RNA polymerase II and several histone modifications that characterise transcription initiation including H3K4me3. RNA polymerase II, H3K4me3 and nucleosome distribution profiles have been analysed at these repeat regions using computer programming techniques. These profiles have in part been found to be similar to those observed at transcription start sites and so this raises the potential for CCG repeats to be implicated as alternative sites for transcription initiation.

Table of Contents

List of Figures

Chapter 1. Introduction

Tandem repeats are a regular feature found throughout the human genome. The definition of these tandem repeats is a pattern of at least two nucleotides repeated one after another with no other nucleotides interrupting the sequence. The number of repeats can range from as little as two or three to as much as many thousands of times. The function of tandem repeats for the most part remains unknown, such as those located in intron regions, although some have been found to be intergenic and have specific function which is expressed in a cell while others can be found in non-coding regions but perform a role in terms of transcriptional regulation. The repeat unit length of a tandem repeat is used as a way of classifying them such as Short Tandem Repeats (STRs). STRs, usually located in non-coding regions of the genome, are also referred to microsatellites and can be anywhere between two and ten base pairs long. STRs can be further classified according to the specific repeat pattern size such as Trinucleotide Repeats (TNRs).

1.1 Repeat expansion

Repeat tracts are polymorphic in that their lengths will vary throughout the population. Their lengths can also change within an individual either becoming longer or shorter as the genome is copied during cell division. They are prone to a phenomenon known as triplet expansion which is a form of mutation that happens during DNA replication. Slippage errors can occur because the nature of the trinucleotide repeats allows them to form into complex loop structures on the daughter strand during synthesis such as hairpin loops and quadruplexes. As this would result in the daughter and parent strand being different, the parent strand is 'repaired' by adding extra nucleotides to ensure it is complimentary to the daughter strand. This is illustrated in figure 1.1. This expansion

can have no detrimental effects such as when it is located within intron DNA but if it is located within exon DNA it could have a negative effect on the expression of the gene leading to disease symptoms caused by either gene suppression or non-viable gene products.

Figure 1.1 - Illustration of slippage which leads to repeat expansion

The basic principle of slippage is shown where the repeats on the daughter strand form loops after the replication fork passes. DNA repair mechanisms then add extra nucleotides to the parent strand in order to make it complimentary.

The following research is concerned with a member of this family, notably CCG repeats. There are known to be approximately 5000 (CCG) n≥4 trinucleotide repeats in the human genome, that is CCG motifs repeating more than four times in succession. These have been found to be non-randomly distributed with many being associated with the 5' ends of genes and their CpG islands that regulate them. There are also many CCG-repeats that are not within or near a known gene and little is known of their purpose.

1.2 Trinucleotide repeats and disease

There are numerous examples of diseases caused by triplet expansion, most of which fall outside of the scope of this investigation, and these can exhibit a phenomenon called anticipation. Anticipation describes where genetic disease symptoms occur earlier in life or are expressed to a greater severity in subsequent generations. This is because successive triplet expansions accumulate over the course of many replication processes and once the number reaches a certain point then the disease

phenotype is fully established. Successive generations can effectively expand the repeat sequence further and further and may only express mild symptoms and can be thought of as carriers but at some point the disease will fully present itself once the number of repeats crosses a specific threshold. For example, Fragile X Syndrome (FRAX) which is an X linked disease and the most common heritable cause of intellectual disability and the most common known cause of autism. A CCG repeat tract is located at the 5' end of Fragile X Mental Retardation 1 (FMR1), a gene that codes for a translational regulator, FMRP, which is highly expressed early in embryonic development (Hinds *et al.* 1993). In normal individuals the repeat number is between five and 50 whilst in carriers of FRAX this can be increased to contain 60 to 190 triplet repeats. The symptoms of fragile X syndrome commonly fully appear when the repeat number expands beyond 200. In this instance the severity of the disease associates with the level of methylation of the CpG island near to the CCG repeat tract which in turn leads to the gene being transcriptionally silenced and a deficiency of FMRP. Higher degrees of methylation equate to increased severity of the disease but the mechanism of silencing is still not exactly known (Kumari, Usdin 2010). Fragile X syndrome is so called because of its correlation with the presence of a rare folate sensitive fragile site on the X chromosome known as FRAXA. A folate fragile site is defined as a loci that shows properties of unstable chromatin, they show as breaks or regions which stain poorly in metaphase chromosomes when cells are cultured in a folate deficient environment. Expansions of CCG repeats have been found to be the basis of all the folate-sensitive fragile sites with the mechanism for this being the formation of unusual secondary structures by one or both of the DNA strands.

1.3 Secondary structures

The precise mechanism by which a CCG expansion causes a fragile site is not fully understood. One suggestion is that they are due to the unusual secondary structures that these sequences may

form. Hairpin loops like those shown in figure 1.2 can be formed by a single strand of repeating CCG or the antisense GGC triplets including G-G and C-C bonding together in a non-Watson-Crick manner.

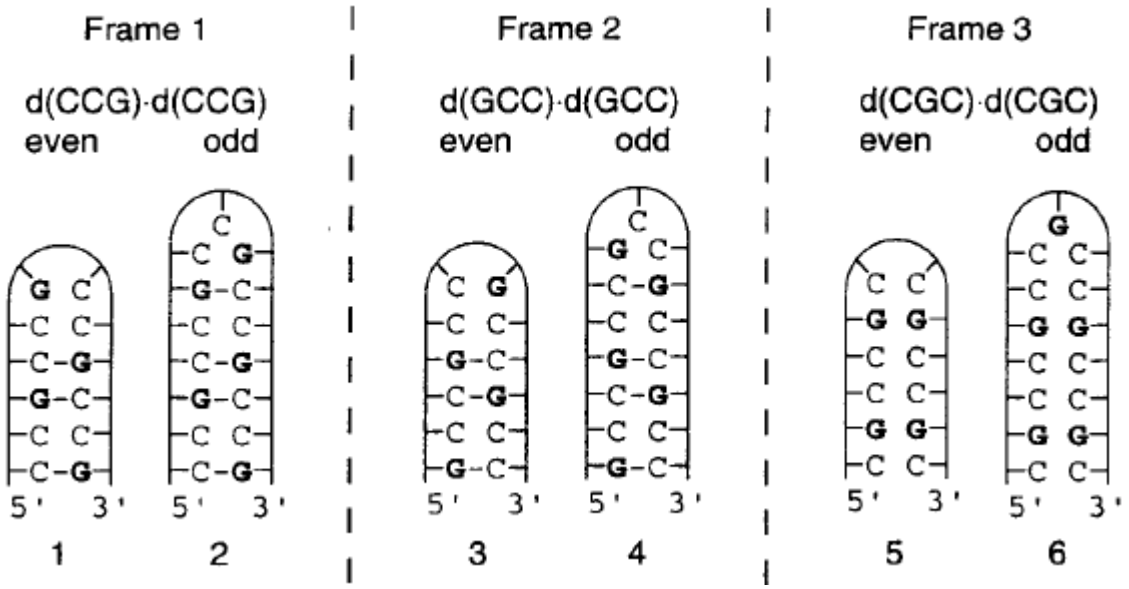

Figure 1.2 – Hairpin loop structures formed by CCG-repeats

The three possible types of hairpin loops formed by CCG repeats. Watson-Crick bonds are displayed but there is also evidence in support of C-C and G-G bonding. (Darlow and Leach, 1998).

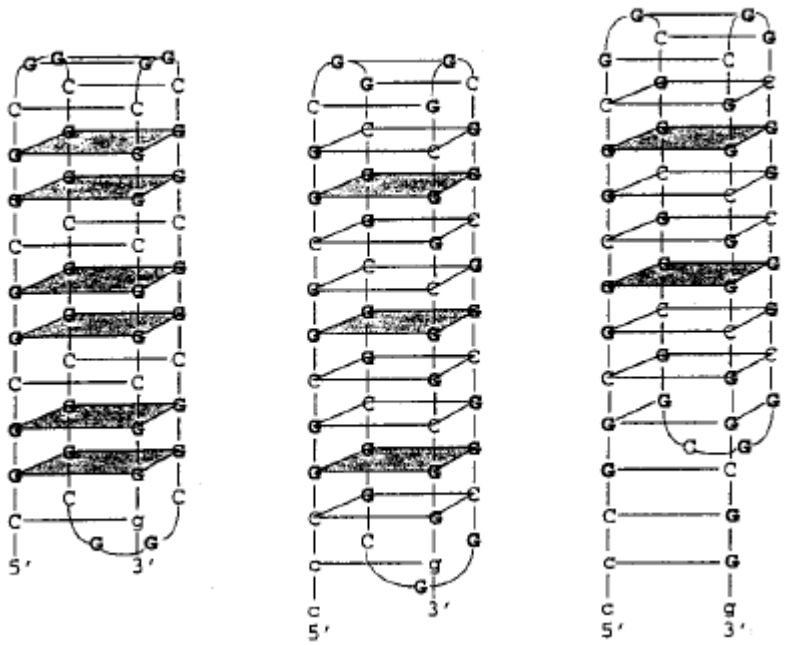

Figure 1.3 – Quadruplexes formed by GGC trinucleotide repeats

Illustration of possible quadruplex formations resulting from a single strand of GGC trinucleotide repeats where G-G bonds are highlighted. The same structures could be possible for CCG repeat tracts. (Darlow and Leach, 1998).

Quadruplexes, figure 1.3, can also be formed by single stranded CCG/GGC repeats where guanine nucleotides can form into G_4-quartets (Darlow, Leach 1998). The result of these unusual secondary structures is that the DNA could be prevented from coiling around the histone core to form nucleosomes and indeed CCG repeat tracts have been shown to be responsible for nucleosome exclusion (Wang *et al.* 1996). This means that sections of DNA containing long CCG repeats do not assemble into nucleosomes as would 'normal' or seemingly 'random' DNA sequences. In fact, nucleosome assembly efficiency decreases with an increase in the length of the repeat tract (Wang *et al.* 1996). Folate sensitive fragile sites could in part be due to nucleosome exclusion. Methylation of CpG islands in eukaryotic cells can cause inhibition of gene expression and this is facilitated by the enzyme DNA methyltransferase which adds a methyl group on to the cytosine residues of 5'-CpG-3' dinucleotides. A methyl-CpG-binding protein is then able to bind to the methylated DNA sequence which in turn causes the inhibition of transcription.

The same process by which nucleosomes are excluded could also lead to a possible explanation for a mechanism by which CCG repeats could have transcriptional potential. If one thinks of CCG repeats being classified as a member of the sequence motif $((G/C)_3NN)n$. This motif has been found to have sequence matches with a group of genes that do not have a 'TATA' box. A TATA box or Goldberg-Hogness box is a common motif found within the promoter region of genes. Genes that do not possess this motif include housekeeping genes, oncogenes and genes which code for transcription factors and growth factors (Wang, Griffith 1996). The exclusion of nucleosomes by the CCG repeats could in these cases be a way in which the promoter region of the gene remains accessible to transcription factors by not allowing the DNA to condense. For example, Sp1, a transcription factor required for the initiation of transcription from genes lacking the TATA box, recognises the consensus sequence 5'-$(G/T)GGGCGGG(G/A)_2(C/T)$-3'. The G/C rich nature of this sequence is not too dissimilar to the motif mentioned above (Wang, Griffith 1996). This may go

some way to providing the basis for an explanation for the non random distribution of CCG repeats associated with the 5' ends of genes and their CpG islands and a mechanism for the association. The uncondensed DNA caused by CCG repeats could make these regions available for the binding of transcription factors and thus be signals for alternative transcription sites.

1.4 Nucleosomes

DNA is for the most part packaged in a condensed form achieved by it coiling around protein core to form nucleosomes. The protein core consists of two copies of each of four histones to make up an octamer. This core provides space for 147 base pairs to wrap around just less than two times. The nucleosomes are separated by stretches of linker DNA with the length of these linker sections varying between species (between 15 and 55 base pairs). The nucleosomes and linker DNA form what is often referred to as “beads on a string”. Further packaging or condensation into an irregular spiral conformation can also occur with approximately six nucleosomes per turn. The chromatin in chromosomal regions which are not being transcribed exists mostly in this condensed form whilst regions that are being transcribed remain in the extended “beads on a string” form.

The histones that make up the octamer core of the nucleosomes each have a flexible tail that is between 11 and 37 amino acids in length. It is these tails that facilitate the chromatin to condense into 30nm fibre. It is thought that this is due to positively charged lysine side chains interacting with the linker DNA (Lodish *et al.* 2003) as well as neighbouring nucleosomes interacting with one another. The histone tails of H3 and H4 are thought to undergo reversible acetylation and deacetylation enabled by enzymes that act upon specific lysine residues in the N-termini of the tails. Lysine carries a positive charge which can form an ionic bond with the negatively charged DNA phosphate groups especially those on the linker DNA which are not already associated with any

histones causing the DNA to condense further. Once acetylated though lysine's positive charge is neutralised and this ionic bonding potential is lost resulting in the DNA remaining in the "beads on a string" form (Lodish *et al.* 2003).

A paper written by Barski *et al.* investigated the influence of histone modifications on gene expression. They used new next-generation sequencing technologies to map a range of histone lysine and argenine methylations and also the distribution of histone variant H2A.Z and RNA polymerase II throughout the human genome and found that many of the methylations were located at higher levels within the promoter regions of genes. The conclusions drawn were to suggest that covalent modifications to histones and the positioning of nucleosomes relative to the DNA have been found to have an effect on gene expression. By making available or unavailable promoter regions for the binding of RNA polymerases and general transcription factors required for transcription initiation, these processes thus can either switch genes on or off. There are a number of different post translational modifications linked to various genetic processes including transcriptional activation, gene repression and DNA repair. The specific histone modifications H3K4me1, H3K4me2 and H3K4me3 have been found to be present at elevated levels surrounding transcriptional start sites (TSSs). The nomenclature signifies the histone has a single, double or treble methylation of the lysine at position four on the histone H3 in this case. Again in promoter regions, the mechanism for gene silencing is also correlated with H3 methylation where evidence suggests a relationship with high levels of H3K27 methylation at silent promoters (Boyer *et al.* 2006). H3K9 methylations are related to the formation of the tightly packed form of DNA heterochromatin and gene silencing (in the cases of H3K9me2 and H3K9me3) which would result but H3K9me1 has a prevalence in many active promoters surrounding the TSS (Bernstein *et al.* 2007). H3K36me3 levels are also shown to increase in association with TSSs with the peak occurring just after the TSS in active genes (Barski *et al.* 2007). The H2A histone variant H2A.Z

which is highly conserved through evolution has been found to be highly enriched upstream and downstream of the TSS in promoter regions and its binding levels correlate with gene activity in humans (Barski *et al.* 2007). Besides gene expression and repression, histone modifications play a role in DNA repair, in particular acetylation of Histones H3 and H4. The modification H4K16ac and gammaH2A.X are needed for the recruitment of Mdcl which is a DNA repair complex adapter protein to sites of DNA damage (Fullgrabe *et al.* 2007). It has also been demonstrated that H3K56ac is important for chromatin assembly during DNA replication and repair and also plays a role in chromatin stability (Fullgrabe *et al.* 2007).

An investigation by Schones *et al.* focused on the importance of nucleosome positioning in relation to DNA to the regulation of transcription. Up until this the mapping of nucleosomes in the human genome was quite limited. Schones *et al.* produced a genome-wide map of nucleosome positions in active and inactive $CD4^+$ T cells using the same next-gen approach that allowed them complete genomic coverage which was not previously possible. Since the H3K4me3 levels in promoter regions showed multiple small peaks separated by roughly 150 base pairs of DNA it was inferred that these peaks could denote the positions of nucleosomes as this is approximately the number of base pairs which coil to form each nucleosome (Barski *et al.* 2007). So the mapping was done by first using MNase to digest chromatin leaving just the separate sections of DNA included in each single nucleosomes and sequencing the ends of nucleosomes using Solexa sequencing. The resulting reads were then mapped to the human genome so their positions within it could be identified. One of the main discoveries resulting from the mapping was the direct relationship between RNA polymerase II binding and the phasing of nucleosomes relative to transcription start sites. Nucleosomes were found to be well phased relative to transcription start sites in expressed genes but not in those that are unexpressed where only the +1 nucleosome was found to be well positioned. This would suggest that nucleosome positioning plays a pivotal role in activating and

deactivating genes. Silent and active genes are also shown to exhibit differing positioning of the first nucleosome downstream of the start site.

1.5 Objectives

Previous analysis of ChIP-seq data from 800 intergenic CCG-repeats shows that they are associated with RNA polymerase II and several histone modifications that characterise transcription initiation including H3K4me3. This would suggest that CCG-repeats may be signals for alternative transcription start sites and so this forms the basis of this investigation.

This project aims to further characterise the transcriptional potential of these intergenic CCG-repeats using a number of different approaches, in order to better define those CCG-repeats that are likely to be transcriptionally active, transcriptionally silent, or non-functional.

Specific Objectives

- The relative abundance of RNA polymerase II and H3K4me3 at individual CCG-repeats will be determined from published ChIP-seq data. Islands of activity that are significantly elevated over the regional background will be mapped with respect to the CCG-repeats, using published methods.
- CCG-repeats will be grouped based on their potential transcriptional activity, based on their association with overlapping islands of PolII and H3K4me3. Comparison with PolII/H3K4me3 profiles at known transcription start sites will provide estimates of potential transcriptional activity at each CCG-repeat.
- The positioning of nucleosomes at individual CCG-repeats will be determined from published ChIP-seq data. Coordinated positioning of nucleosomes is a feature of active transcription start sites, and comparison of their locations in each of the groups described

above (2) will provide additional evidence as to whether these groups do indeed reflect different levels of transcriptional activity.

Possible objectives, if there is sufficient time

- Interrogate published RNA-seq data to find evidence of transcripts associated with individual CCG-repeats.

Proposed Methodology

- All of the published data described in the "Objectives" are held in a MySQL relational database in the host laboratory. Results will be written to a relational database using SQL. Analysis of CCG-repeat centric data will be performed using the statistical programming environment 'R'.

Chapter 2. Materials and Methods

2.1 Materials

The following details the materials used throughout this study. The project was purely a computing exercise as all the data used had been previously accumulated by other people (more on these resources follows below).

The majority of the work was carried out on a laptop running an Ubuntu 10.10 maverick meerkat operating system. This was used in conjunction with MySQL, version 5.1.49-1ubuntu8.1, to import and query databases. Scripts for analysing and displaying data were all written using R version 2.11.1 using the gEdit text editor along with a terminal for executing them. A couple of R libraries were also used to add extra functionality to the scripts; RmySQL for the incorporation of SQL queries into the scripts and SNOW for running the scripts on a cluster. The clusters in question were two servers, Eve and Sheldon, at the Institute for Women's Health at UCL. These ran on an Apple OSX and OpenSuse operating system respectively and had multiple quad cores which made them ideal for speeding up time consuming computations. Also used throughout the project have been Open Office word processor and Photoshop elements 7.0 for writing and drawing diagrams.

2.2 Resources

A number of resources have been used to carry out this study. Further details of database tables can be found in appendix A;

- A database containing an up to date copy of the human genome, in this case hg18, in order to extract data such as genes and their transcription start sites.

- Text files detailing the relevant subsets of these genes. In this case four separate lists; activated and resting cells and the present and absent genes for each.
- CCG database
- Nucleosome tag database as well as flat files containing nucleosome tag information

2.2.1 UCSC Genes

The gene data used was the same as that used by Schones *et al.* where "expression microarray experiments were performed for both resting and activated T cells using Affymetrix Human Genome U133 Plus 2.0 GeneChip array" (Schones *et al.* 2008). In each case genes were classified as either absent or present. The UCSC genome table browser was used to obtain UCSC ids from the Affymetrix ids and lists were formed for each set which were saved as text files. An R script was used in conjunction with these text files to query the knowgeneold3 table in the hg18 database on the UCSC server. Query results contained all the relevant gene information and this was saved as Rdata files for use in the analysis. It was important to ensure that any findings from this study could be viewed as an extension of the work carried out by Schones *et al.* therefore using the same data was essential.

2.2.2 CCG trinuceotide repeat database

The trinuceotide repeat database was obtained from the Satellog database (Missirlis *et al.* 2005) which contains all pure 1-16 perfect repeat units found in the human genome (Ensembl homo_sapiens_core_19_35b). The version stored on the servers at UCL was upgraded from build v35 to build v36 using the UCSC liftover tool. The gene data came from UCSC (Barski/Schones) but in this instance the definition of the intergenic/intragenic CCG-repeats comes from analysis of

the ensembl v36 database - intergenic means a CCG-repeat that is at least 2kb from any know transcript start site (C. Jones, personal communication). This database consists of multiple tables containing various lists of different trinucleotide repeats. For this study however only the table ccg_repeats was required. This table was made up of over 11 million ccg repeat tags, their start and end points, chromosome, sequence, unit type and length. The mid point of each repeat was later added in order to centre nucleosome counts around this point.

2.2.3 Nucleosome tag database

$CD4^{+}$T cells provided the genetic material for obtaining the nucleosome tag data. Resting and activated T cells were treated separately with the digestion enzyme MNase in order to separate individual nucleosomes. Agarose gels were used to isolate each mononucleosomal DNA fragment which were then sequenced using an Illumina 1G Genome Analyser (Barski *et al.* 2007). These fragments were then mapped to the human genome to ascertain their location within it. The nucleosome tag database was broken down in to multiple tables according to chromosome and whether the tags are found on either the genome in resting or activated cells. Information includes tag id, start, end and strand.

2.3 Methods

2.3.1 Gene subset data frames

The gene information from knowGeneOld3 was extracted using four gene lists described in Schones *et al.* 2008. These were simple text files and detailed the following subsets of genes: Present genes in resting cells, absent genes from resting cells, present genes in activated cells and

absent genes from activated cells. Firstly the text file was read into R and assigned as an object and then a mySQL connection was established to the hg18 database at UCSC using the R library RmySQL. A small R script was written to loop through each entry within the gene lists and retrieve information from the database for each gene and build a new R object containing gene information such as id, chromosome, strand, start and end for each subset. The four gene subsets were then saved as Rdata files.

2.3.2 Functions

The ultimate aim of the main function was to count the number of nucleosome tags around each gene's transcription start site in order to gain insight into the distribution of nucleosomes in these areas. The primary goal was to devise a method similar to that used by Schones *et al.* with the aim being to initially be able to recreate their results and so ensure that when the study was expanded to CCG repeats then subsequent findings would be consistent with these results.

Nucleosome tags were counted within a +/-1000 base pair range around each TSS. A sliding window of five base pairs was used in order to help smooth the data and make any phasing in the plots easier to observe. In terms of the method this meant that when a nucleosome tag was observed then the five positions including and around the tag position were scored. The basic flow of the R script written to carry out this counting is illustrated as a flow chart in figure 2.1 and the script itself is presented here along with other R scripts used as appendix B.

A main function was used to initialise the data frames that would contain the nucleosome count data and provide a framework to call two further functions which retrieve the tags within the +/-1000 base pair range for each TSS and count the tags and their positions within this range. Here also the R libraries were sourced and a connection to the relevant database established. All of the genes are analysed using a “for” loop inside which the chromosome, strand and base pair position of the TSS

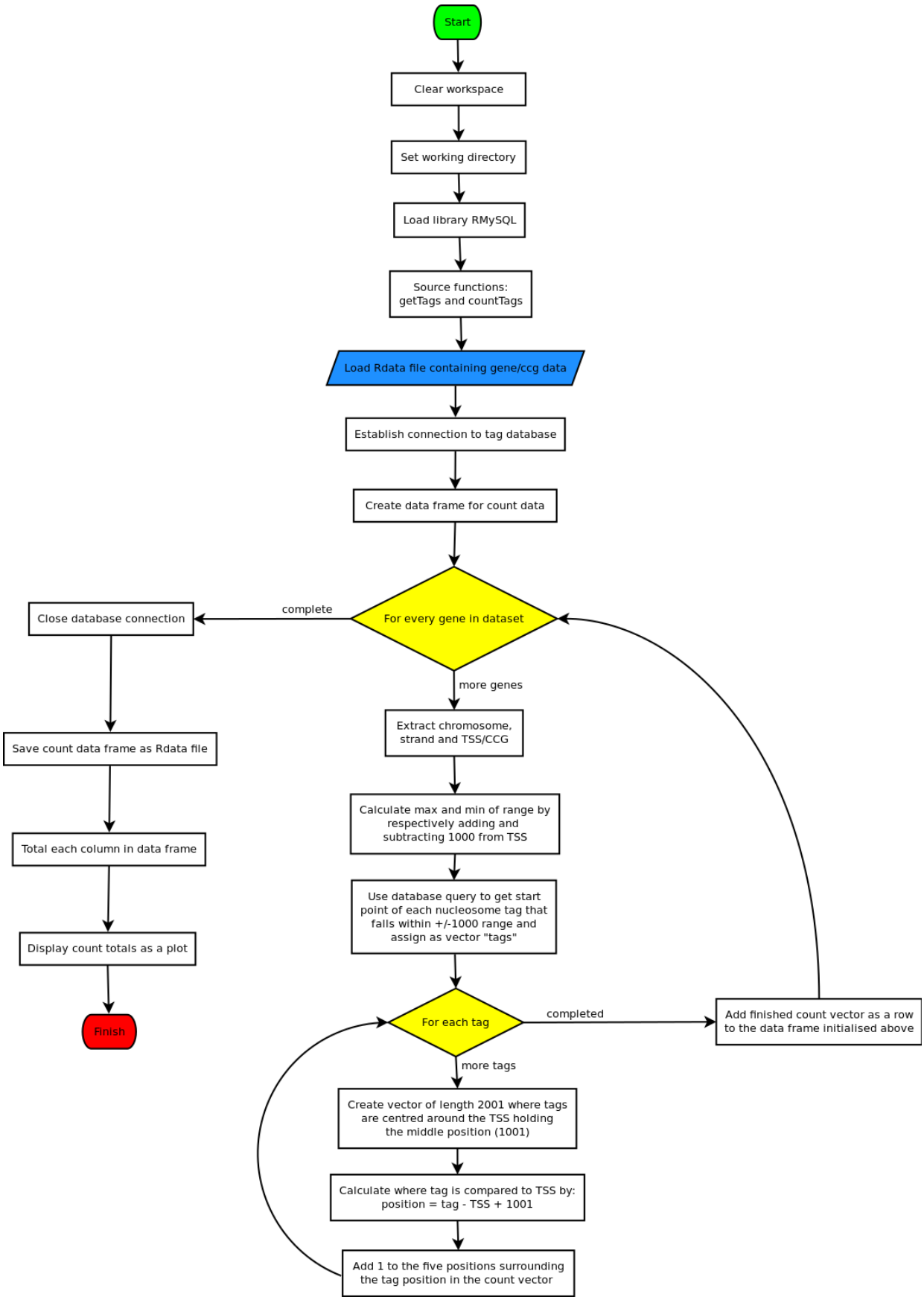

Figure 2.1 – Analysis function work flow

A flow chart to illustrate how the main function to standardise and count nucleosome tags works. The two loops represent the two sub functions 'getTags' and 'countTags'. The same function was also used to count H3K4me3 and polymerase II tags by changing the database query in the first loop; 'getTags'.

are assigned as new variables. Then an “if” statement separates each gene according to which strand it is on.

Within the if statement first the function 'getTags' is called and to this is passed the chromosome, TSS, strand and database connection. GetTags first sets the maximum and minimum (+/-1000bp) either side side of the TSS and database query is used to retrieve all of the tags within this range from the nucleosome_tags database. This list represents all of the nucleosome tags that are associated with a particular transcription start site and is returned to the main function.

In turn the tag list is passed onto another function which counts the tags. A vector of length 2001 is initialised and this is used to store the tag count for each gene. The vector represents the +/-1000 base pair range with the TSS being in the middle. Each tag is standardised to this vector by:

$$\text{standardised tag} = \text{tag start} - \text{TSS} + 1001$$

As an example: if the TSS = 1234 and the tag start = 1500 then the standardised tag position would be 1267. This can also be viewed as the tag being 266 base pairs upstream of the TSS. By subtracting the TSS from the tag start the distance between the two is measured and the 1001 must be added because in this instance the TSS occupies the middle of the vector, position 1001 so each standardised tag value now holds a value between 1 and 2001. One is added to the vector in the five positions around each standardised tag to simulate the progress of a five base pair sliding window. The sliding window would loop through each position in the vector and count how many standardised tags were present there before adding this number to the vector. This was found to be computationally intensive and so the method described above, although seemingly more complicated to imagine, was used to speed up the process.

The completed vector is then returned to the main function and is added as a row to the data frame initialised at the start of the main function. The process outlined above is also run in parallel for the

tags counted on the anti-sense strand resulting in a second data frame. Once all the genes have been analysed in this way the database connection is terminated and the data frames saved.
To visualise this data each column in the data frame is totalled to make two vectors of length 2001 which are then plotted on the same axis.

2.3.3 Parallel Processing of Data Analysis

In order to decrease the time taken to run each script the R package SNOW was used. SNOW stands for Simple Network of Workstations and is a method for running R jobs in parallel. This is achieved by linking up a cluster of multiple computers with each sharing the work load that would otherwise normally be carried out on a single CPU core. In the case of this study the cluster simply comprised of two quad core processors making a total of eight nodes. When using SNOW the dataset being analysed had to split into multiple equal parts according to how many nodes were being used. SNOW works by having a single master process and one or more slave processes. The command is executed by the master process and this farms out to the slave processes their portion of the overall computation. This communication is one way however so the function had to include a save statement as the output from each slave process would not be returned to the master. Within in the function was also a command to assign the first entry in each subset of the data in question (rep_id for the CCG list or gene_id for the gene lists) and this variable was used in the save file command to ensure that the saved file from each slave process would be unique as if all had the same name they would overwrite each other. A small script was devised to automatically subset the data being analysed according to how many nodes were to be used in the cluster and this can be found in the appendix A.

Chapter 3. Results

3.1 Nucleosome phasing

In order to analyse the distribution of nucleosomes in relation to the position of CCG repeats in a similar vein to that done by Schones *et al.* for transcription start sites it was necessary to calculate the middle of each CCG repeat. This was done by averaging the values of start and end of the repeat.

3.1.1 Nucleosome phasing at the TSS

The initial goal for the analysis was the reproduction of the plots found in the Schones paper. When writing the script for analysing nucleosome distribution at first the same data was used, namely activated (present) genes found in resting T cells. The validity of any subsequent findings relating to CCGs lay in the accurate reproduction of the Schones analysis. It was also useful to have a point to aim for to ensure that the R script being written was correct. Figure 3.1 shows the phasing of nucleosome levels around the transcription start site of activated genes in resting T cells. The sense strand is shown in red whilst the antisense strand is in green. The nucleosome phasing can be inferred from the distance between the peaks. There are a few peaks upstream of the TSS with more peaks with regular phasing downstream. The gaps between these peaks would seem to correlate with the 146 base pairs wrapped around a nucleosome. The nucleosome tags represent the start position of each nucleosome and so it is fair to assume that nucleosomes are phased with respect to the position of the TSS itself.

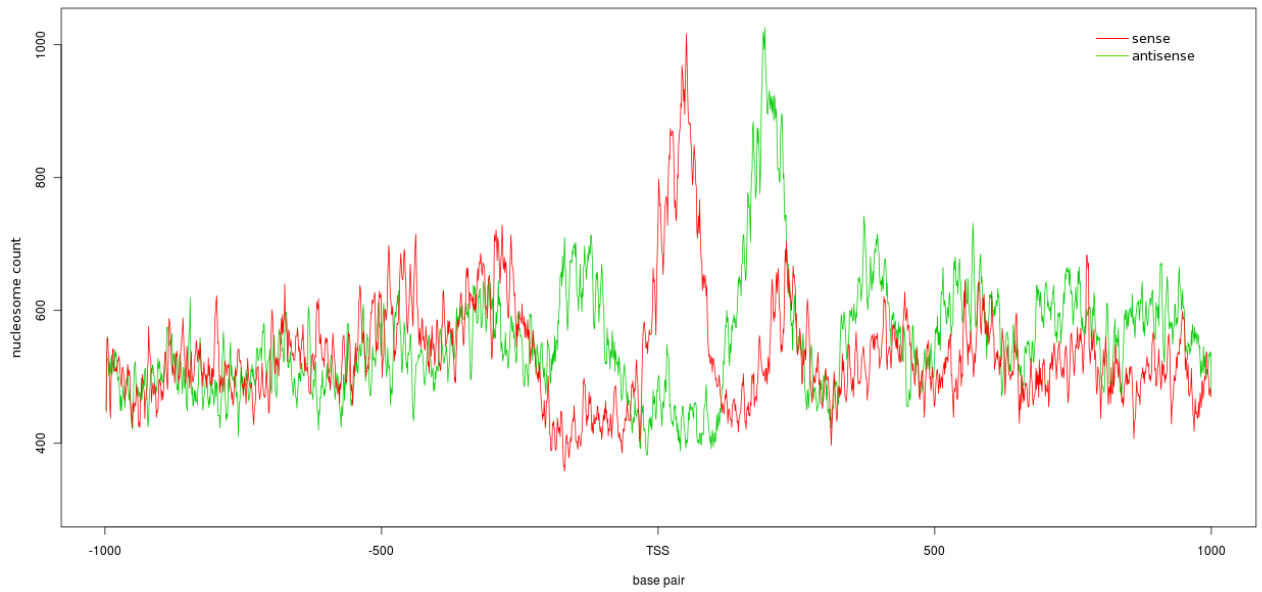

Figure 3.1 – Distribution of nucleosome tags relative to TSSs

Recreation of figure featured in Schones *et al.* 2008. The number of nucleosomes found around transcription start sites. Phasing of nucleosomes can clearly be seen as the peaks of each plot. Sense strand is red and antisense is green.

3.1.2 Nucleosome distribution at CCG-repeats

Once the function had been tested on transcription start sites it was used on the CCG-repeat data. Using R coupled with MySQL queries subsets of the CCG-repeat database were created as data frames according to the number of trinucleotide units (n) comprising the repeats. Nucleosome tags were counted with respect to the mid point of each CCG-repeat for n=3, n=4, n=5, and n=6. The resulting plots are shown in figures 3.2 – 3.5. The frequency of counts has been normalised by dividing each count by the number of repeats in the set.

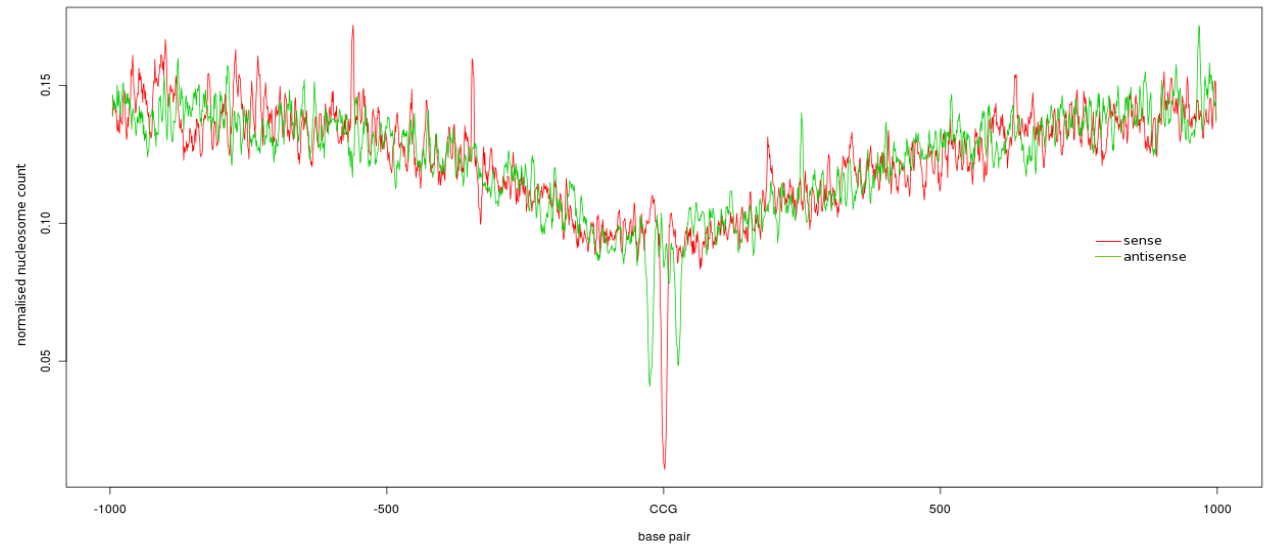

Figure 3.2 – Distribution of nucleosome tags relative to CCG-repeats (n=3)

Distribution of nucleosomes around CCG-repeats where the number of repeat units is three (n=3).

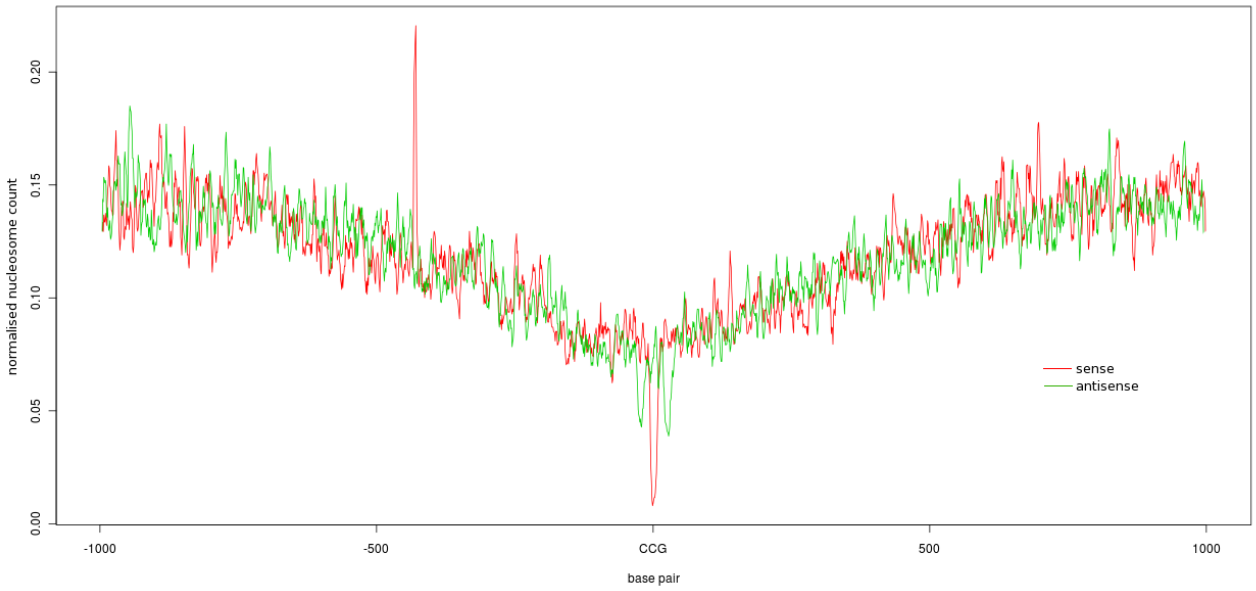

Figure 3.3 – Distribution of nucleosome tags relative to CCG-repeats (n=4)

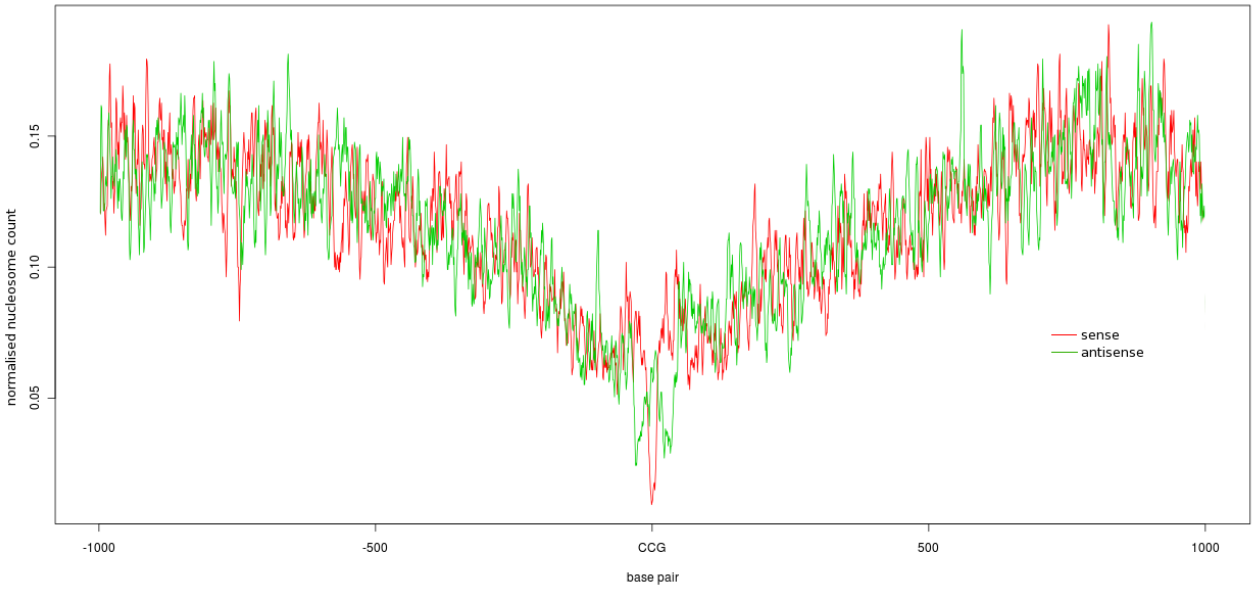

Figure 3.4 – Distribution of nucleosomes tags relative to CCG-repeats (n=5)

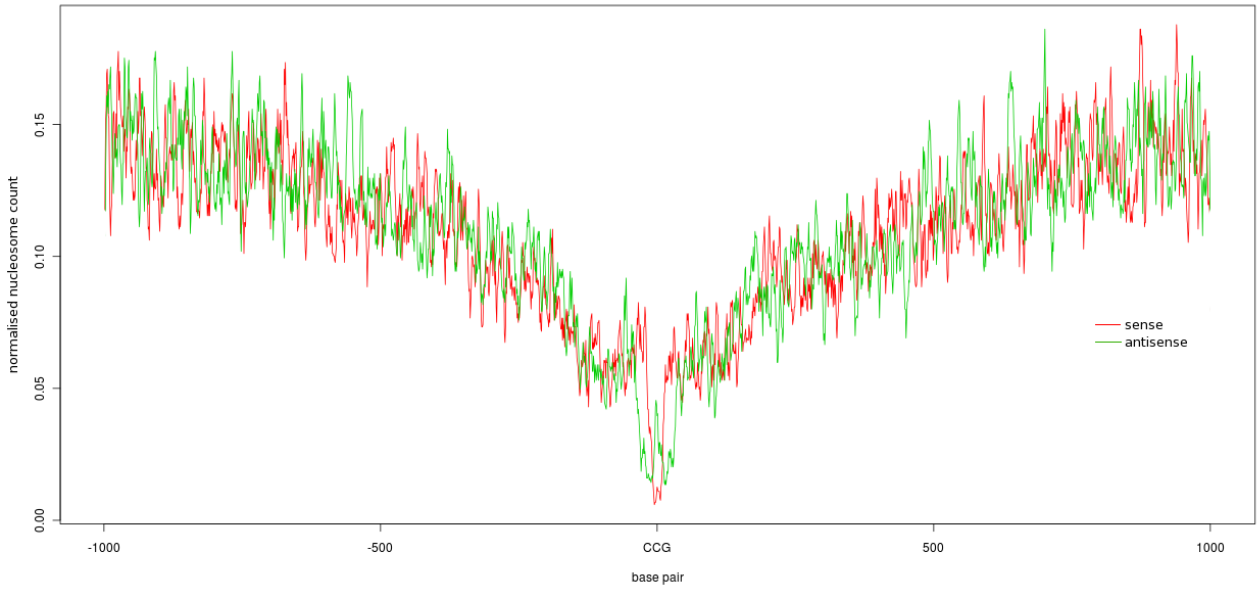

Figure 3.5 – Distribution of nucleosomes relative to CCG-repeats (n=6)

The nucleosome distribution plots above clearly show that nucleosomes are being excluded at CCG-repeats. Furthermore the extent of the exclusion is increased with the number of repeat units. The relative frequency of nucleosome tags at each base pair would appear to be about 0.14 in areas away from the CCG-repeats. This tag frequency falls in all cases at about 700 base pairs away from the repeat while the gradient increases with increased number of repeat units. In all cases almost no nucleosome tags are observed at the repeat itself. Interestingly there is a noticeable difference between the profile of the sense and antisense strand close to the repeat. Nucleosome tags are seen at their lowest number on the repeat for the sense strand (red) whilst for the antisense strand (green) there is actually a small peak at this point in comparison to dips either side. This 'W' shape is not present for the sense strand.

The next step of the analysis involved looking into the apparent difference between the sense and antisense strands. The nucleosome tags were counted in relation to the repeats found on each strand separately. Prior to this the term CCG-repeats has been used to refer to CCG-repeats as well as the reverse compliment GGC-repeats. In the following plots these have been separated.

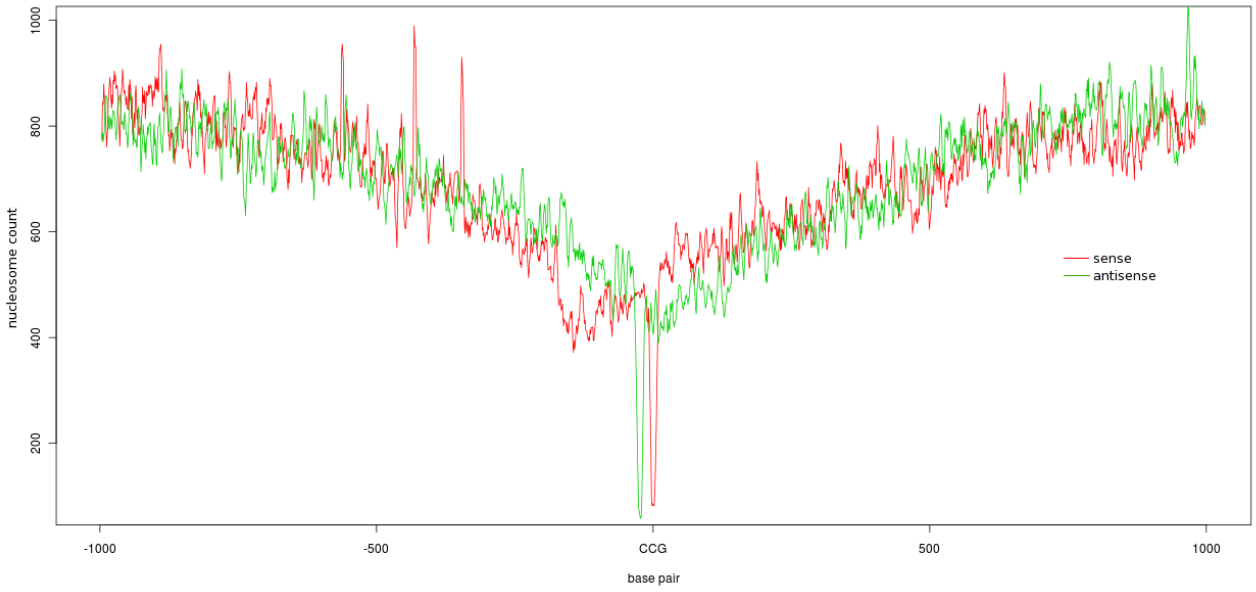

Figure 3.6 – Distribution of nucleosomes in relation to CCG-repeats; those found on the sense strand

The middle of the plot illustrates an interesting feature in that the nucleosomes are excluded just prior to the CCG-repeat (roughly 20-30 base pairs upstream) whilst nucleosomes on the sense

strand are excluded most at the repeat itself. A similar profile is observed when plotting the GGC-repeats in figure 3.6. Here also is shown more clearly a pattern around the troughs in that the sense and antisense plots mirror each other immediately around the repeat and have roughly the same distributions further away from the repeat. Upstream of the repeat more nucleosome tags are found on the antisense strand and downstream of the repeat the opposite is true.

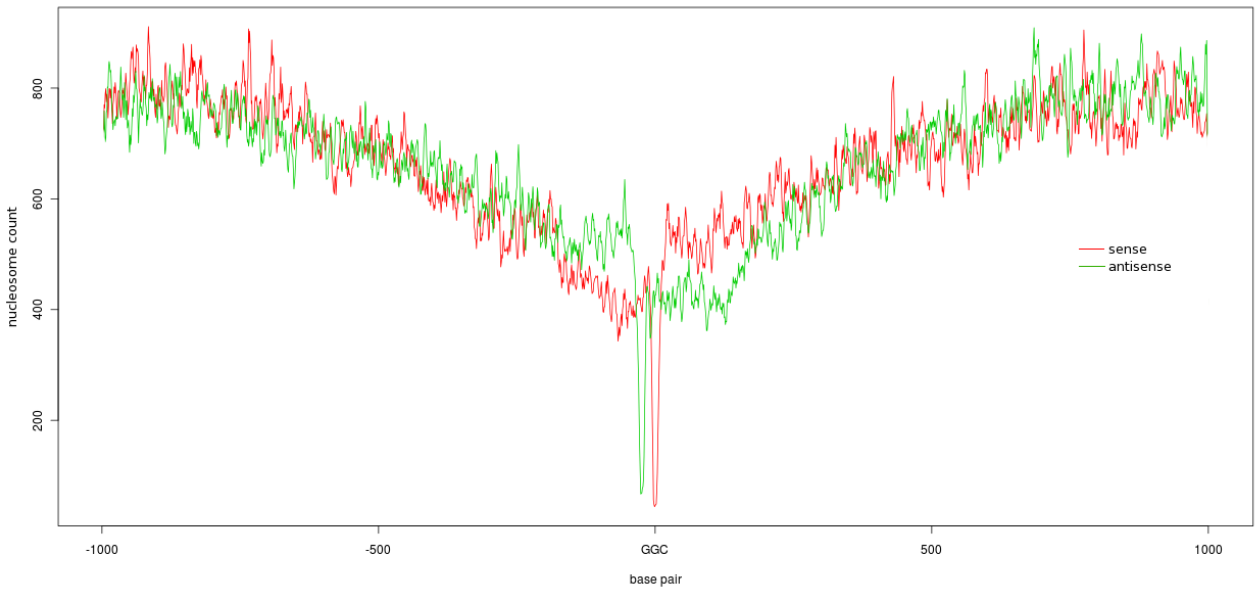

Figure 3.6 – Distribution of nucleosomes in relation to GGC-repeats; those found on the antisense strand

Quadruplexes play an important role in nucleosome exclusion as has been discussed in the introduction so a subset of the CCG-repeat data set was also created to include only those repeats that were not within +/- one kilobase pairs of a quadruplex. Figure 3.7 shows the resulting distribution and confirms that CCG-repeats themselves actively exclude nucleosome formation.

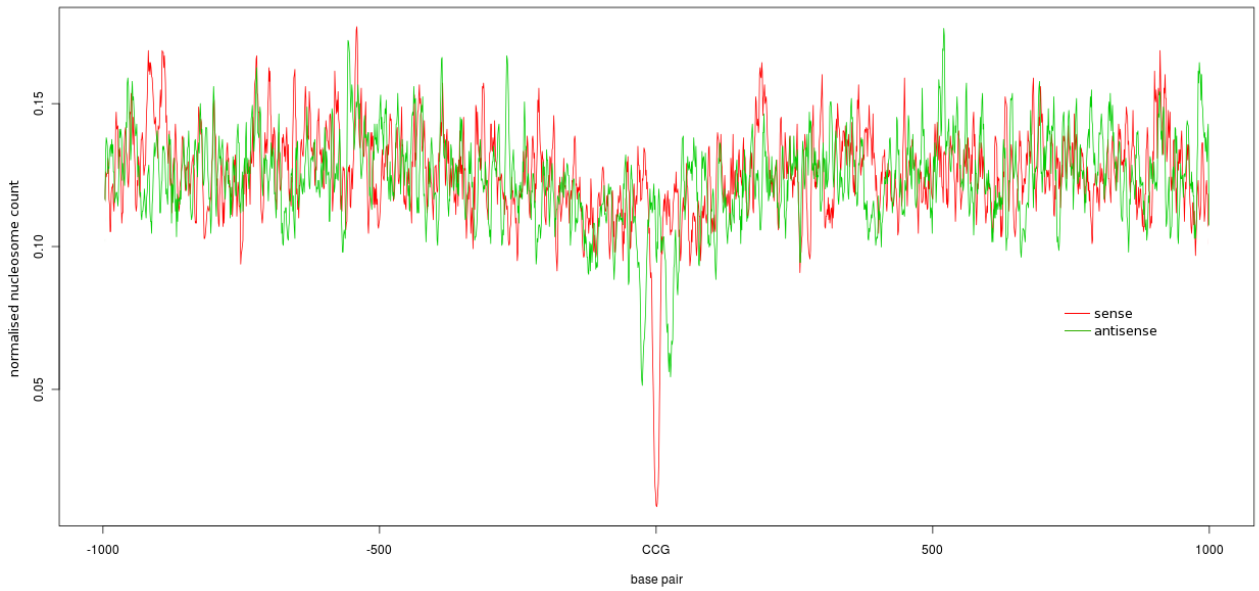

Figure 3.7 – Distribution of nucleosome tags relative to CCG-repeats not near quadruplexes

Distribution of nucleosome around CCG-repeats that are not within 1kb of a quadruplex. This plot illustrates how the repeats themselves exclude nucleosome formation rather than just the quadruplexes that they could in theory be forming also.

3.1.3 Nucleosome profiles for genes near to a CCG-repeat

To gain further insight into the effect of CCG-repeats on transcription initiation a subset of the gene data was created to include only those genes that are within +/-150 base pairs of a CCG-repeat. Plotting nucleosome counts around these areas with respect to the TSS should produce a profile similar to that seen in figure 3.1 and indeed figure 3.8 below does illustrate phasing downstream of the TSS. Although the smaller sample size probably accounts for a more noisy profile the peaks downstream of the TSS appear to be at regular intervals of approximately 146 base pairs.

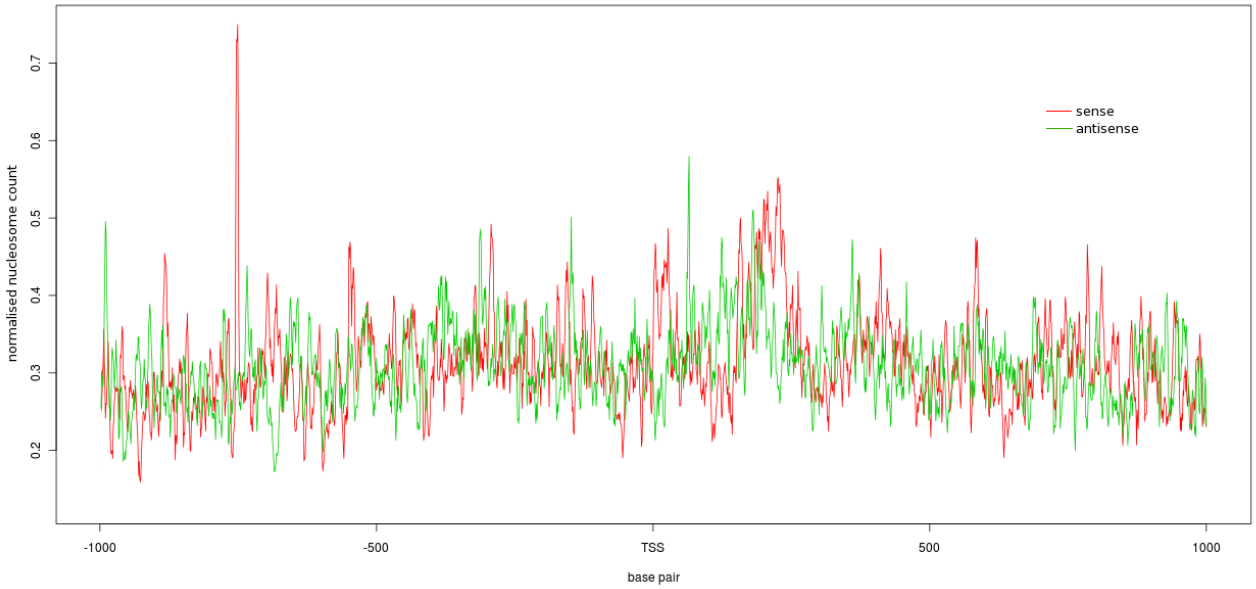

Figure 3.8 – Distribution of nucleosome tags for genes that are near CCG-repeats (A)

Subset of the active genes where a CCG-repeat is found within +/-150 base pairs. Nucleosome distribution profile centred around the TSS.

The same data was plotted again but this time with respect to the mid point of the CCG-repeat concerned. The profile appeared to be significantly different from those nucleosome profiles of CCG-repeats in general featured above. Exclusion of nucleosomes still occurs at the repeat itself but in addition there is a hint of the plot peaking at regular intervals of around 146 base pairs downstream especially noticeable in the sense strand. This adds weight to the hypothesis of CCG-repeats possessing transcriptional potential as one of the traits of transcription start sites (phasing of nucleosomes) is also present to an extent in the case of CCG-repeats. Again, the graph is fairly noisy due to the relatively small sample size.

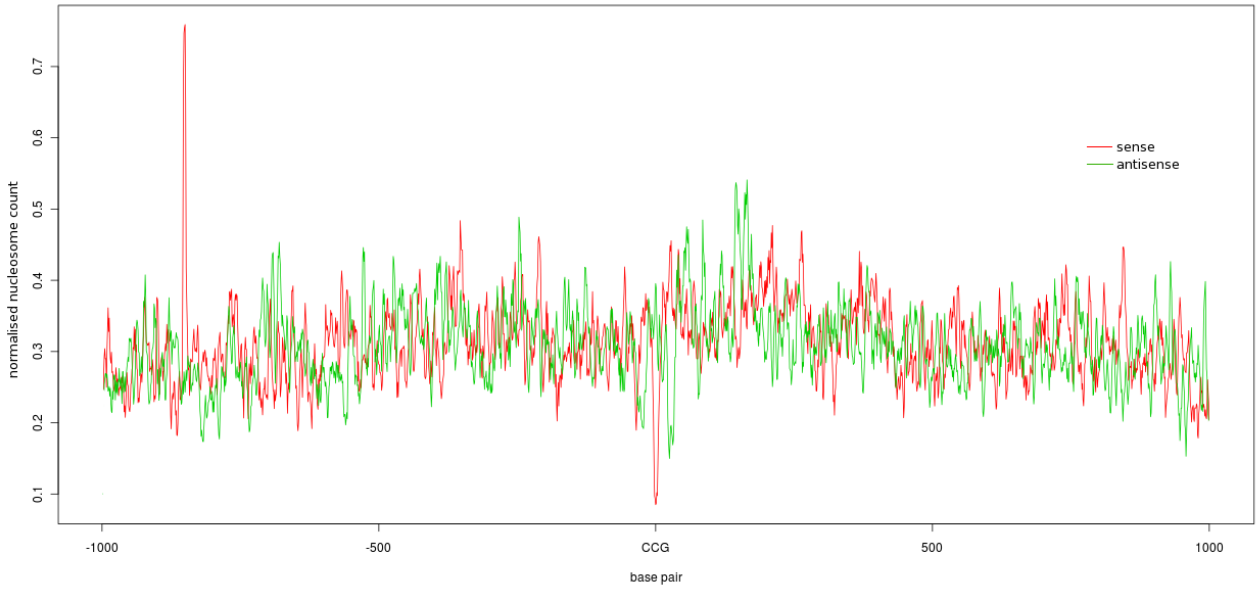

Figure 3.9 – Distribution of nucleosome tags for genes that are near CCG-repeats (B)

Subset of the active genes where a CCG-repeat is found within +/-150 base pairs. Nucleosome distribution profile centred around the CCG repeat.

3.2 H3K4me3 histone modifications

Another of the features characterising transcription start sites identified by Schones *et al.* was the pattern of histone modifications. Certain histone modifications are associated with transcription start sites with H3K4me3 being the most significant of these. For this reason H3K4me3 tags were counted in the +/-1000 base pair region around the transcription start sites of genes that have a CCG-repeat within +/-150 base pairs. The analysis was then carried out with tags being counted with respect to the repeats.

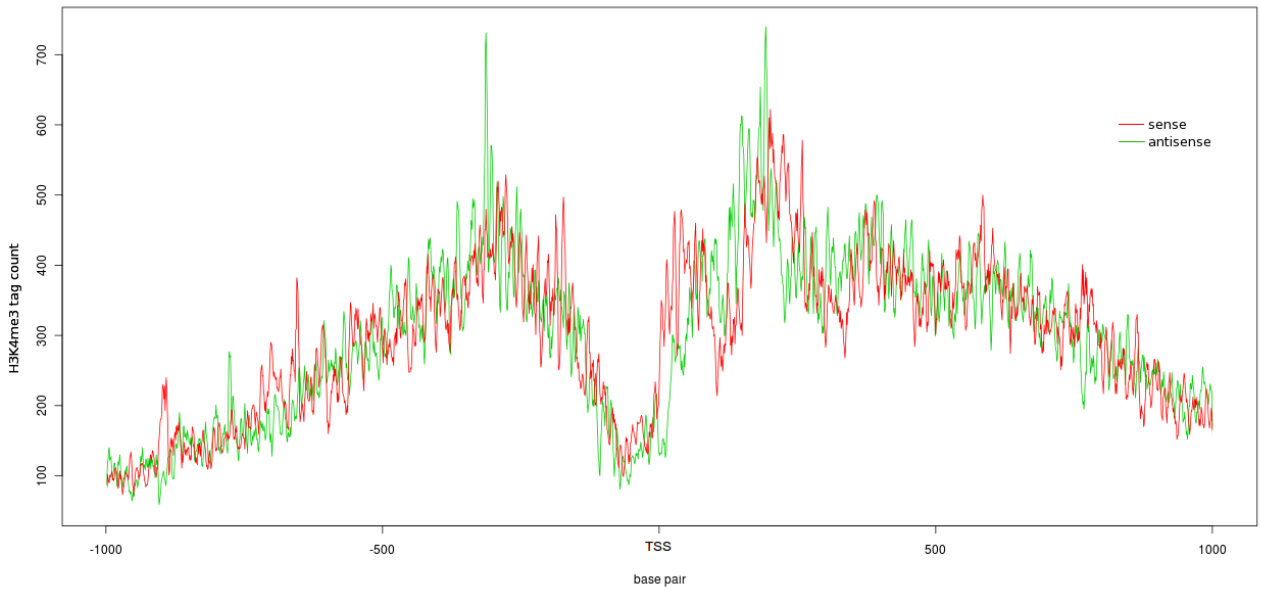

Figure 3.10 – Distribution of H3K4me3 tags relative to TSS

H3K4me3 tags around the TSS of genes where a CCG-repeat can be found within +/-150 base pairs.

The figure above shows how the levels of the histone modification H3K4me3 change at transcription start sites. There is a steady increase upstream of the TSS with levels falling sharply approximately 250 base pairs upstream. At the TSS itself the level H3K4me3 again starts to rise, this time steeply, before tailing off downstream of the TSS.

The profile for H3K4me3 tags for the same data set but centred around CCG-repeats is shown in figure 3.11. It has similar profile to that of figure 3.10. As with the nucleosomes there is a relative absence of tags found at the repeat but some of the more subtle characteristics are also present such as four peaks downstream of the TSS or CCG.

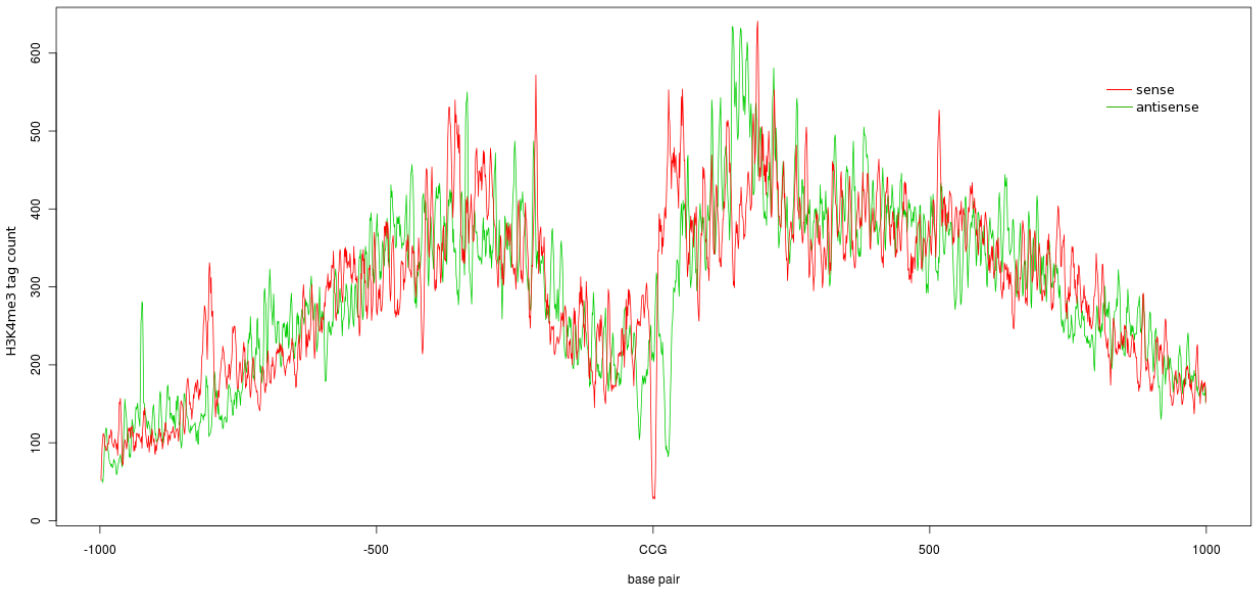

Figure 3.11 – Distribution of H3K4me3 tags relative to CCG-repeats

H3K4me3 tags relative CCG-repeats that are within +/-150 base pairs of active genes.

3.3 RNA Polymerase II

RNA polymerase II binding was also found to correlate with nucleosome phasing around transcription start sites by Schones *et al.* As would be expected polymerase II levels are at their highest at the TSS and this can be seen clearly in figure 3.12.

In this instance Schones *et al.* used a different size of sliding window to count the polymerase II tags in order to create a less noisy plot, however it was not noted what size this sliding window was. By altering this variable and rerunning the function it was possible to approximate the profile in the Schones paper and use this as a basis for doing the same but for CCG-repeats. The resulting plot is shown in figure 3.13. The polymerase II levels do peak at the repeat which characterises transcription initiation but the profile is quite different suggesting that CCG-repeats are not acting in exactly the same way as transcription start sites.

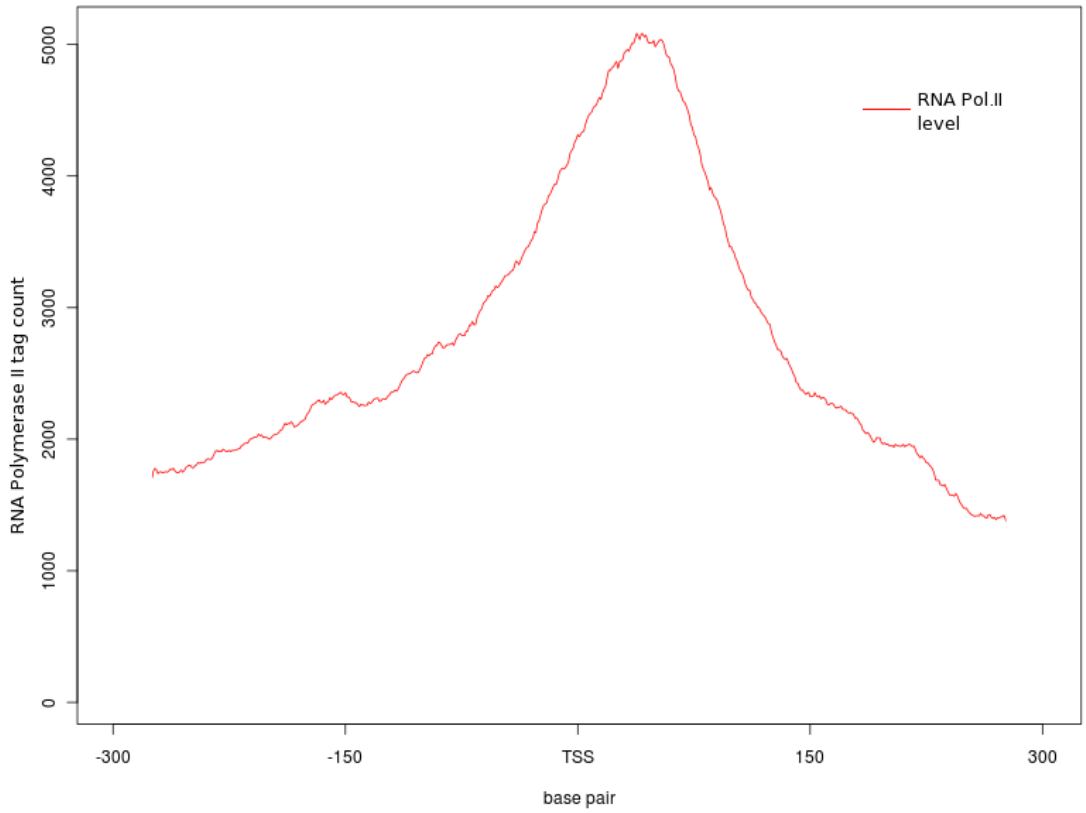

Figure 3.12 – Polymerase II levels around transcription start sites

Distribution shows increased RNA polymerase II found at transcription start sites

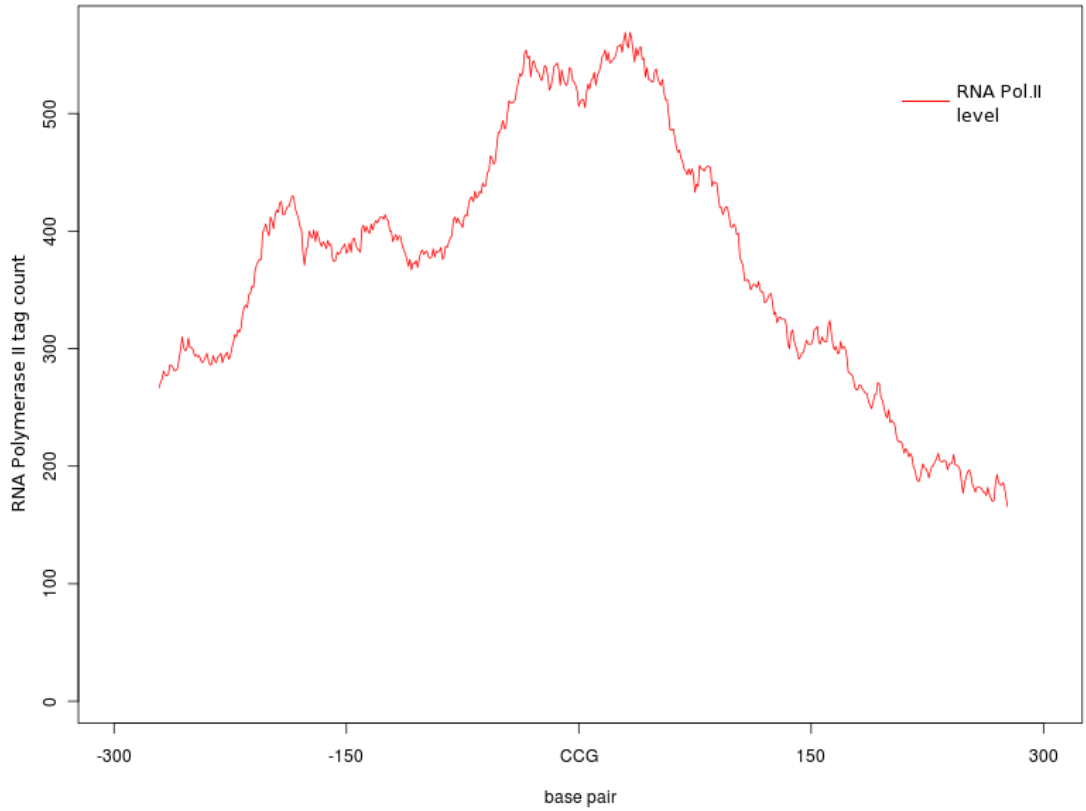

Figure 3.13 – Polymerase II levels around CCG-repeats

This plot suggests increased RNA polymerase II at CCG-repeats but the profile is more complex than that found at the TSSs

Finally, to try and gain further insight into the correlation of polymerase II and nucleosome phasing, the active CCG-repeats were divided according to whether they exhibited stalled or elongating polymerase II. Only CCG-repeats that are not near to known genes were used in this part of the analysis in order to separate any findings from the influence that the genes themselves may have. It was important to characterise the CCG-repeats only and not have genes effecting the analysis. This was done using a method like the one used by Schones *et al.* The stalling index was calculated for each repeat and this is defined by Schones as being the ratio of the promoter polymerase II level over the average gene body level. The promoter polymerase II level being the sum of all polymerase II signals in a one kilobase region surrounding the the TSS. The average gene body level was calculated for one kilobase windows throughout the gene body (Schones *et al.* 2008).

CCG-repeats are of course different to genes in that they are far shorter in length and so windows of one kilobase were used ranging from -2kb to +2kb with respect to the CCG-repeat start instead of the 1kb windows throughout the gene body. The promoter polymerase II level in the case of the CCG-repeats was defined as the total polymerase II tag counts in the range of +/-500 base pairs around the CCG-repeat.

CCG-repeats that had a stalling index of ten or more were classified as being stalled, that is significantly more polymerase II can be found at the CCG-repeat. A stalling index of between one and three and having an average body polymerase II level of at least five were classified as exhibiting elongating polymerase II.

Nucleosome distributions were investigated for these two subsets of CCG-repeats and these can be seen in figures 3.14 and 3.15.

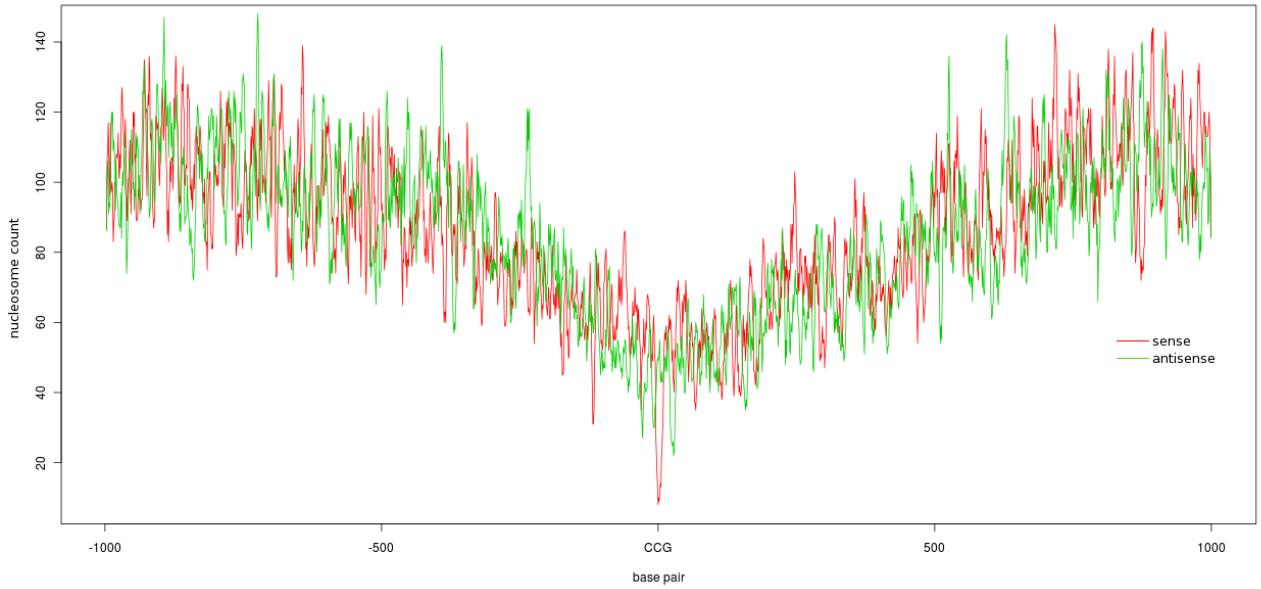

Figure 3.14 – Nucleosome distribution pattern around CCG-repeats with elongating polymerase II

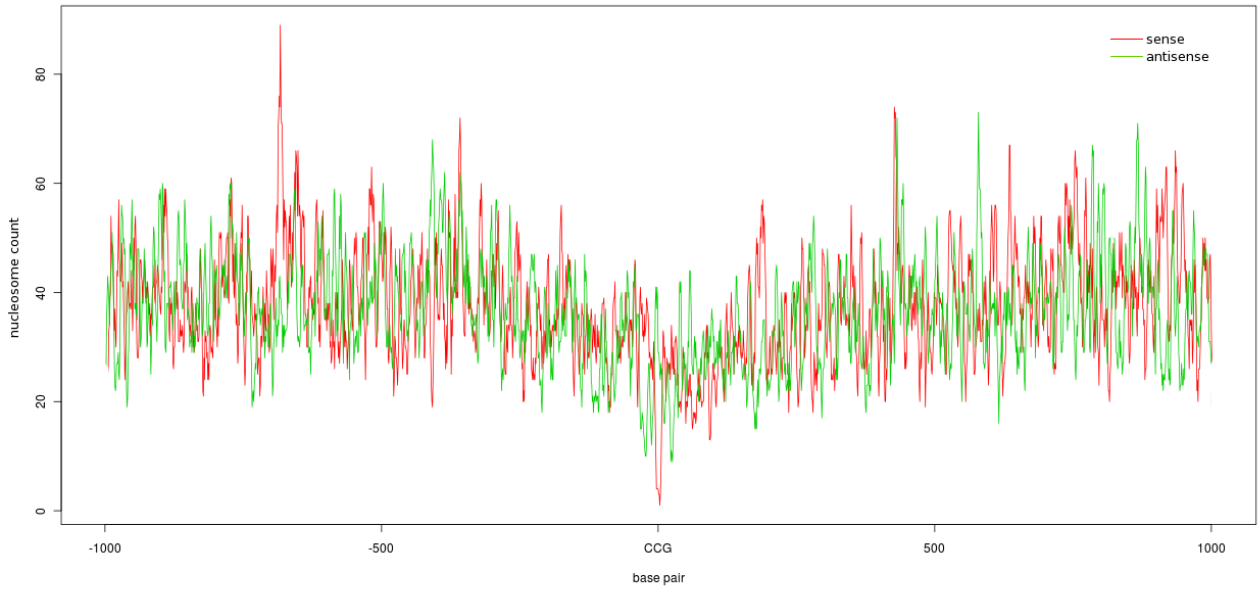

Figure 3.15 – Nucleosome distribution pattern around CCG-repeats with stalled polymerase II

Chapter 4: Discussion

The characterisation of the transcriptional potential of intergenic CCG-repeats was the aim of this study with the hypothesis being that these regions of the genome could act as alternative transcription start sites. This was tested by analysing certain features of CCG-repeats and comparing these features to those of transcription start sites. The hypothesis was tested by reproducing and expanding upon certain analyses used in a paper by Schones *et al.* where the phasing of nucleosome distribution, the levels of histone modifications and the levels of RNA polymerase II were used to characterise transcription start sites. Profiles for these variables were plotted and those of CCG-repeats were analysed and also compared to those of transcription start sites to assess any similarities and whether or not these could provide any evidence of functionality.

4.1 CCG-repeats exclude nucleosomes

The counting of nucleosomes around CCG-repeats produced the most visibly striking profiles and gave strong evidence for the exclusion of nucleosomes by the repeats. This was expected as CCG-repeats were already known to prevent the formation of nucleosomes but the plots nevertheless illustrated this well for a genome wide context. Also the degree to which nucleosomes are excluded according to the length of the repeat has been shown with a negative correlation between length of repeat and the amount of nucleosome formation. Not only did CCG-repeats exclude nucleosomes to a greater degree at the repeat itself when they increased in size but the area of exclusion widened beyond that of the repeat.

This on its own can be used as evidence for transcription initiation potential as relatively large transcriptional machinery is required for transcription activation the assembly of which is

incompatible with the nucleosome structure (Lorch *et al.* 1987).

It was of interest to note how CCG-repeats on the sense and antisense strands excluded nucleosome differently. When plotting the sense and antisense repeats separately it is shown that nucleosome tags found on the antisense strand initially show less exclusion but then reach their lowest point slightly upstream of the repeat before recovering to an intermediate level and then following the expected pattern as seen in the previous plots. The tags on the sense strand do almost the exact opposite with the highest amount of exclusion being on the repeat.

As discussed in the introduction CCG-repeats can form secondary structures including quadruplexes. For this reason it was important to assess the impact of CCG-repeats that were not associated with quadruplexes. The resultant profile shows the expected exclusion of nucleosomes at the repeats in a way that is similar to that seen for the plot for trios of repeats (where n=3). This would probably suggest that nucleosome exclusion is a result of a combination of quadruplexes as well as other factors; possibly hairpin loops. A likely explanation for the profile observed is that the subset of CCG-repeats that are not near quadruplexes was for the most part made of short repeat sequences (n=3). As the repeat sequences increase in length they become more likely to form quadruplexes, thereby excluding themselves from the group of repeats not near to quadruplexes, and this is why the degree of exclusion of nucleosomes for this group is underwhelming.

4.2 CCG-repeats near genes

In order to further assess the transcriptional potential of CCG-repeats in terms of nucleosome distribution gene data and repeat data was combined to create a subset of genes that were within +/- 150 base pairs of a CCG-repeat. This was done to try and see if the nucleosome profile of these areas showed phasing around the CCG-repeat. The profile centred around the transcription start sites of these genes followed the expected pattern albeit more noisy with less well defined peaks.

This can be assumed mainly to be due to the smaller dataset. Phasing can be seen as a regular series of peaks downstream of the TSS roughly 146 base pairs apart suggesting coordinated positioning of nucleosomes.

The main feature of the same dataset but with nucleosome counts being relative to the CCG-repeats is once again the exclusion at the repeat itself. Elsewhere the profile is less well defined than before but does not appear to be random. There is a hint of a series of peaks downstream of the CCG-repeats but it is more difficult to make out. It would therefore seem logical to infer that there is some coordinated positioning of nucleosomes in these areas but it is less.

The relatively weak phasing here could point to CCG-repeats in this instance having a relatively small amount of transcription activation potential and thus be linked with genes that are expressed at low levels. It would be therefore be interesting to profile these genes further too see what their expression levels are and what gene products they encode.

4.3 CCG-repeats share a H3K4me3 profile with TSSs

The histone modification H3K4me3 is shown in elevated levels surrounding transcription start sites and plays a role in the formation of nucleosomes (Barski *et al.* 2007). This information was plotted by counting the H3K4me3 tags around transcription start sites of the genes near CCG-repeats and plotted to produce a distinct profile. The same tags were counted relative to the CCG-repeats to see if the two profiles were similar. The two profiles were indeed remarkably similar; elevated levels of H3K4me3 both upstream and downstream of the TSS or repeat with a series of regularly spaced peaks downstream. Since H3K4me3 is implicated in the condensation of DNA it can be inferred that in both cases there is a likeness in the mechanism of coordinated nucleosome positioning.

4.4 RNA polymerase II levels at CCG-repeats

RNA polymerase II tags were counted around transcription start sites to produce a figure like that by Schones *et al.* Since presence of polymerase II signifies transcription it would be expected to find elevated levels at transcription start sites which indeed you do as illustrated. Following this on to CCG-repeats that were near genes the same process was carried out to see if polymerase II levels increased there. The basic profile of the resulting plot did indicate increased polymerase II levels at CCG-repeats strengthening the hypothesis of transcription activation but the profile was more complex than that for transcription start sites. Looking at the plot it would seem that there small sub-peaks or 'shoulders' to the general shape. This could suggest that these points are where the actual transcription start sites are in relation to the CCG-repeats and so the increased levels there are showing up on the CCG-repeats plot also. Further analysis of the distributions of transcription start sites around CCG-repeats and vice-versa could shed light on this and I would expect this distribution to have a likeness to the CCG-polymerase II plot.

4.5 Variations in nucleosome positioning with polymerase II binding

Schones *et al.* showed how nucleosome phasing and positioning is surrounding transcription start sites is dependent on the level of polymerase II binding. The trend that they found was that greater binding of polymerase II increased nucleosome phasing. The technique that Schones *et al.* used to identify genes as exhibiting stalled or elongating polymerase II was adapted for classifying CCG-repeats in the same manner. Should the CCG-repeats have analogous characteristics in this respect then it would have been expected to see some evidence of nucleosome phasing in each case (although to a lesser degree as already seen for the whole CCG-repeat dataset compared to that for transcription start sites). In actuality the only definitive feature of each plot is the exclusion of nucleosomes at the repeat. It could be argued that there is a slight hint of phasing downstream in the

elongating plot but this is only vaguely recognisable. The amount of noise due to the small dataset prevents any meaningful information to be gleaned from either plot.

Chapter 5: Conclusions

5.1 Objectives

The objective of this study was to attempt to characterise the transcriptional potential of CCG-repeats. This objective was approached by investigating a combination of three features associated with transcription start sites; histone modifications, notably H3K4me3, RNA polymerase II abundance and coordinated positioning of nucleosomes.

- **The relative abundance of RNA polymerase II and H3K4me3 at individual CCG-repeats will be determined from published ChIP-seq data. Islands of activity that are significantly elevated over the regional background will be mapped with respect to the CCG-repeats, using published methods.**

  The relative abundances of both RNA polymerase II and H3K4me3 were successfully mapped with respect to the CCG-repeats. The main feature of the polymerase II profile was shared with that of the same analysis done for transcription start sites in that the amount of polymerase II increased from a background level peaking at the CCG-repeat. It was also clear, however, that this plot had other features such as a smaller peak roughly 200 base pairs upstream of the repeat and that the main peak itself was actually two peaks with a dip in between the CCG-repeat.

  Again, the H3K4me3 profiles for transcription start sites and CCG-repeats were very similar in shape. The only real difference being the relative lack of tags found at the CCG-repeat compared to the to the TSS. If H3K4me3 is associated with transcription start sites then the similar profile found for CCG-repeats suggests an underlying similarity in functionality of these regions. Overall it can be concluded that with an additional increase in polymerase II

abundance at the CCG-repeats the hypothesis of these repeats sharing transcriptional properties with transcription start sites is reinforced although the differences suggest that they are certainly not one and the same.

- **CCG-repeats will be grouped based on their potential transcriptional activity, based on their association with overlapping islands of PolII and H3K4me3. Comparison with PolII/H3K4me3 profiles at known transcription start sites will provide estimates of potential transcriptional activity at each CCG-repeat.**

  It was possible to group CCG-repeats according to overlapping islands of polymerase II although this was not done for H3K4me3. Whilst repeats were grouped according to whether they had elongating or stalled polymerase II the resulting profiles were largely inconclusive. The use of a smoothing function in this case would perhaps have resolved the data enough for patterns to emerge.

- **The positioning of nucleosomes at individual CCG-repeats will be determined from published ChIP-seq data. Coordinated positioning of nucleosomes is a feature of active transcription start sites, and comparison of their locations in each of the groups described above (2) will provide additional evidence as to whether these groups do indeed reflect different levels of transcriptional activity.**

  Profiling of nucleosome distribution at CCG-repeats was successfully completed. The nature and level of nucleosome exclusion was clearly demonstrated for increasing lengths of repeats to illustrate how increased repeat length correlates with increased nucleosome exclusion. The role that quadruplexes play in nucleosome exclusion was also investigated successfully confirming that quadruplexes form the main mechanism for exclusion. There was a small amount of evidence found for the coordinated positioning of nucleosomes downstream of CCG-repeats near genes although further steps could have been taken to make this clearer.

The evidence obtained from the various lines of investigation do support the hypothesis that CCG-repeats have transcriptional potential. The role that they play is more complex than that of transcription start sites and it may be that the repeats function along side these in certain cases and for a certain family of genes.

5.2 Improvements

One of the main limitations of this study came to fore during the analysis; the relatively small size of the dataset used meant that certain profiles were often difficult to see properly due to the level of noise generated by the analysis. This was compounded by subsetting the data further. Since the graphical figures relied on interpretation by eye to discern patterns this became a problem as any patterns that were present were partially masked. This issue could be at least partially resolved by employing a smoothing function into the plotting script or by using software that provides various options for this. In hindsight, a simple moving average would not have been difficult to use and this could have greatly improved the visualisation of some of the data.

5.3 Future Work

This study has produced some interesting and significant findings providing insight into the function of CCG-repeats and so it would certainly be worthwhile to supplement this with further research.

Immediate lines of investigation which relate to areas already covered should include profiling of the genes that are located near to CCG-repeats. Clearly not all genes are near CCG-repeats so if the repeats are significant then this may infer that there is some reason for some genes to be associated them. If CCG-repeats do possess some transcription activation potential then the mechanism for

transcription of nearby genes may depend on this and so genes near to CCG-repeats may share some common characteristics such as patterns of expression levels suggesting functions relating to a shared process or processes. A more specific function of CCG-repeats could be elucidated should associated genes be found to have mutual traits.

Investigation into the distribution of genes with respect to CCG-repeats within a certain range could also help to explain whether there is a specific link between genes and CCG-repeats. The plot of RNA polymerase II levels at CCG-repeats featured numerous sub-peaks in addition to the main trend of polymerase II levels being increased at CCG-repeats. This could be interpreted purely as an quirk of polymerase levels at the repeats but it is more likely to be the increased polymerase II levels of transcription start sites. Furthermore the occurrence of sub-peaks indicates that the position of the CCG-repeats and transcription start sites to one another is not completely random but at specific intervals. An organised distribution in this instance would further add weight to the hypothesis of CCG-repeats having a function in transcription activation.

Further comparisons between the histone modifications found at transcription start sites and CCG-repeats would also be useful in order to characterise the transcriptional potential of CCG-repeats further. The levels of histone modifications that have already been found to silence genes could be assessed at CCG-repeats that are near to genes that are not being expressed. This would be beneficial in seeing if CCG-repeats have more control over gene expression than simply providing a means for certain cases of transcription activation.

References

- Barski, A. *et al.*, 2007. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4), pp.823-37.
- Berger, S.L., 2002. Histone modifications in transcriptional regulation. *Current Opinion in Genetics & Development*, 12(2), pp.142-148.
- Bernstein, B.E., Meissner, A. & Lander, E.S., 2007. The Mammalian Epigenome. *Cell*, 128(4), pp.669-681.
- Boyer, L.A. *et al.*, 2006. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, 441(7091), pp.349-353.
- Darlow, J.M. & Leach, D.R.F., 1998. Secondary structures in d(CGG) and d(CCG) repeat tracts. *Journal of Molecular Biology*, 275, pp.3-16.
- Datta, S. *et al.*, 2011. Nucleosomal occupancy and CGG repeat expansion: a comparative analysis of triplet repeat region from mouse and human fragile X mental retardation gene 1. *Chromosome research*, 19(4), pp.445-55.
- Fuks, F., 2005. DNA methylation and histone modifications: teaming up to silence genes. *Current Opinion in Genetics & Development*, 15(5), pp.490-5.
- Füllgrabe, J., Kavanagh, E. & Joseph, B., 2011. Histone onco-modifications. *Oncogene*, 30, pp.3391–3403.
- Hinds, H.L. *et al.*, 1993. Tissue specific expression of FMR-1 provides evidence for a functional role in fragile X syndrome. *Nature Genetics*, 3(1), pp.36-43.
- Kumari, D. & Usdin, K., 2009. Chromatin remodeling in the noncoding repeat expansion diseases. *The Journal of Biological Chemistry*, 284(12), pp.7413-7.
- Kumari, D. & Usdin, K., 2010. The distribution of repressive histone modifications on silenced FMR1 alleles provides clues to the mechanism of gene silencing in fragile X syndrome. *Human Molecular Genetics*, 19(23), pp.4634-42.
- Lodish, H., Darnell, J.E. & *etc*, 2003. *Molecular Cell Biology* 5th ed., W.H.Freeman & Co Ltd.
- Lorch, Y., LaPointe, J.W. & Kornberg, R.D., 1987. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell*, 49(2), pp.203-210.
- Lukusa, T. & Fryns, J.P., 2008. Human chromosome fragility. *Biochimica et Biophysica Acta*, 1779(1), pp.3-16.
- Matchan, A., 2007. Analysis of CCG-repeats in the human genome: Do translocation genes have a unique profile?, Unpublished MSc Thesis, Cranfield University, Cranfield.

- Missirlis, P.I. *et al.*, 2005. Satellog: a database for the identification and prioritization of satellite repeats in disease association studies. *BMC Bioinformatics*, 6, p.145.

- Radman-Livaja, M. & Rando, O.J., 2010. Nucleosome positioning: how is it established, and why does it matter? *Developmental Biology*, 339(2), pp.258-66.

- Schones, D.E. *et al.*, 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5), pp.887-98.

- Strahl, B.D. & Allis, C D, 2000. The language of covalent histone modifications. *Nature*, 403(6765), pp.41-5.

- Sutherland, G.R. & Richards, R.I., 1995. Simple tandem DNA repeats and human genetic disease. *Proceedings of the National Academy of Sciences*, 92, pp.3636-3641.

- Tabolacci, E. *et al.*, 2005. Differential epigenetic modifications in the FMR1 gene of the fragile X syndrome after reactivating pharmacological treatments. *European Journal of Human Genetics*, 13(5), pp.641-8.

- Tassone, F. *et al.*, 2011. Differential usage of transcriptional start sites and polyadenylation sites in FMR1 premutation alleles. *Nucleic acids research*.

- Taverna, S.D. *et al.*, 2007. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nature Structural & Molecular Biology*, 14(11), pp.1025-40.

- Wang, Y.H. *et al.*, 1996. Long CCG Triplet Repeat Blocks Exclude Nucleosomes: A Possible Mechanism for the Nature of Fragile Sites in Chromosomes. *Journal of Molecular Biology*, 263(4), pp.511-6.

- Wang, Y.H. & Griffith, J., 1996a. Methylation of expanded CCG triplet repeat DNA from fragile X syndrome patients enhances nucleosome exclusion. *The Journal of Biological Chemistry*, 271(38), pp.22937–22940.

- Wang, Y.H. & Griffith, J., 1996b. The [(G or C)3NN]n motif: a common DNA repeat that excludes nucleosomes. *Proceedings of the National Academy of Sciences*, 93(17), pp.8863-7.

- Wells, Robert D, 2009. Mutation spectra in fragile X syndrome induced by deletions of CGG*CCG repeats. *The Journal of Biological Chemistry*, 284(12), pp.7407-11.

Appendix A

List and outline of database tables used

Table: nucleosome_tags

Description:			
All of the nucleosome tag data was stored in tables like this. Tags were separated into separate tables according to chromosome and whether they were found in resting or activated cells.			
Columns:			
Field	**Type**	**Key**	**Description**
tag_id	int(10)	primary	Auto-increment ID
start	in(9)		Base pair location of start of nucleosome tag
end	int(9)		Base pair location of end of nucleosome tag
strand	tinyint(1)		Strand that tag is found on (sense = 1, antisense = -1)
Example:			
tag_id	**start**	**end**	**strand**
1	21722086	21722109	-1

Table: ccg_repeats

Description:

The Satellog CCG-repeats was stored in the following format. In the case of the gene_context intergenic intergenic means a CCG-repeat that is at least 2kb from any know transcript start site.

Columns:

Field	Type	Key	Description
rep_id	int(7)	primary	
chr	char(2)		Chromosome repeat is located on
start	int(9)		Base pair location of start of repeat relative to chromosome
end	int(9)		Base pair location of end of repeat relative to chromosome
unit	char(3)		Repeat unit *e.g*. CCG, GGC, GCC, *etc*.
seq	varchar(100)		The whole repeat sequence
length	tinyint(2)		Number of whole repeats
strand	tinyint(1)		Strand repeat is found on
gene_context	enum		Whether gene is intragenic or intergenic
mid_point	decimal(10)		Mid point between start and end of repeat

Example:

rep_id	chr	start	end	unit	seq	length	strand	gene_context	mid_point
1502	19	21953	21962	CCG	CCGCCGCCGC	3	-1	intergenic	21957.5

Table: H3K4me3_tags

Description:			
Table contained in the nucleosome_tag database detailing each H3K4me3 tag location.			
Columns:			
Field	**Type**	**Key**	**Description**
tag_id	int(10)	primary	Auto-increment ID
chr	char(2)		Chromosome that tag is found on
start	int(9)		Base pair location of start of tag on the relative chromosome
end	int(9)		Base pair location of end of tag on the relative chromosome
strand	tinyint(1)		Strand tag is found on

Example:

tag_id	chr	start	end	strand
1	14	70487263	70487286	1

Table: polII_tags:

Description:			
Table contained in the nucleosome_tag database detailing each RNA polymerase II tag location.			
Columns:			
Field	**Type**	**Key**	**Description**
tag_id	int(11)	primary	Auto-increment ID
chr	char(2)		Chromosome that tag is found on
start	int(11)		Base pair location of start of tag on the relative chromosome
end	int(11)		Base pair location of end of tag on the relative chromosome
strand	tinyint(1)		Strand tag is found on

Example:

tag_id	chr	start	end	strand
4	10	11117281	11117304	1

Appendix B

Download gene data from UCSC

```
library(RMySQL)

# read in text files of present and absent genes to make 2 data frames
X <- read.table("resting-present-gene-list.txt", sep="\n", header=FALSE)[,1]
Z <- read.table("resting-absent-gene-list.txt", sep="\n", header=FALSE)[,1]

# connect to UCSC database
m <- dbDriver("MySQL")
con <- dbConnect(m, host = "genome-mysql.cse.ucsc.edu", user = "genome", password = "", dbname = "hg18")

# create data frame for resting present genes info (X)
resting.genesP <- data.frame()

# loop each gene in X (present genes)
for (gene in X){
        print(gene)
        # query ucsc database
        sql <- paste("select * from knownGeneOld3 where name=",gene,"", sep="'")
        # assign outcome as Y
        Y <- dbGetQuery(con, sql)
        # add returned data (Y) to data frame in R
        resting.genesP <- rbind(resting.genesP, Y)
}

resting.genesA <-data.frame()
for (gene in Z){
        print(gene)
        sql <- paste("select * from knownGeneOld3 where name=",gene,"", sep="'")
        Y <- dbGetQuery(con, sql)
        resting.genesA <- rbind(resting.genesA, Y)
}

# close connection to ucsc
dbDisconnect(con)

# save the two data frames so they can be loaded another time
save(resting.genesP, resting.genesA, file = "old3restingGenes.Rdata")
```

Adding TSS to gene data

```
# remove columns that aren't needed
resting.genesA <- resting.genesA[,-c(6:12)]

# add transcription start site to genes data frame
startSites <- c()

for (a in 1:length(row.names(resting.genesA))){
        print(a)
```

```
		strand <- resting.genesA$strand[a]
		if(strand == "+"){
			TSS <- resting.genesA$txStart[a]
			startSites <- rbind(startSites, TSS)
		}else if(strand == "-"){
			TSS <- resting.genesA$txEnd[a]
			startSites <- rbind(startSites, TSS)
		}
}

resting.genesA = transform(resting.genesA, TSS = startSites)
```

Main function

```
rm(list=ls())

# set the working directory
setwd("/home/jake/Documents/Thesis/data/5.08.11")

# function to pass to cluster
my.function <- function(data){

		# initialise data frame to contain nucleosome count data
		plusAllCounts <- data.frame()
		minusAllCounts <- data.frame()

		# set the working directory for my.function
		setwd("/home/jake/Documents/Thesis/data/5.08.11")

		# source libraries and functions
		library("RMySQL")
		source("getTags.R")
		source("countTags.R")

		# MySQL connection details
		m <- dbDriver("MySQL")
		con <- dbConnect(m, host = "localhost", user = "root", password = "emu", dbname = "nucleosome_tags") #
		connect to db via MySQL

		# make vector of first gene so when saving each sub list it has a unique name
		first.entry <- data[,1][1]

		# loop through each gene/repeat in the set 'data'
		for (a in 1:length(row.names(data))){
			chr <- data$chr[a]
			strand <- data$strand[a]
			point <- data[,6][a]

			# sense strand
			if(strand == "1"){
				plusTags <- getTags(chr, point, '1', con)
				minusTags <- getTags(chr, point, '-1', con)
				plusCounts <- countTags(plusTags, point)
				minusCounts <- countTags(minusTags, point)

			# antisense strand
			}else if(strand == "-1"){
```

```
                plusTags <- getTags(chr, point, '1', con)
                minusTags <- getTags(chr, point, '-1', con)
                plusCounts <- rev(countTags(plusTags, point))
                minusCounts <- rev(countTags(minusTags, point))
            }
            plusAllCounts <- rbind(plusAllCounts, plusCounts)
            minusAllCounts <- rbind(minusAllCounts, minusCounts)
        }
        dbDisconnect(con)
        save(plusAllCounts, minusAllCounts, file = paste(first.entry,"_AllCounts.Rdata",sep=""))
}

# load Rdata file containing data frames with ccg information (id, start, end, chromosome, point, etc)
load(file="restingGenesAllccgs.Rdata")

# subset of genes where TSS has nearby ccg repeat
data1 <- subset(resting.genesP, resting.genesP$mid.points != "NA")

# set parameters for cluster and to subset dataset for each node
ccgList <- list()
node.num <- 8
ccg.num <- length(row.names(data))
split.num <- ccg.num%/%node.num

for(i in 1:(node.num-1)){
        ccgList[[i]] <- data[(((i-1)*split.num)+1):(i*split.num),]
}
ccgList[[node.num]] <- data[(((node.num-1)*split.num)+1):length(row.names(data)),]

# create the cluster using SNOW
library(snow)
cluster <- makeCluster(node.num, type="SOCK")

# run function on cluster
clusterApply(cluster, ccgList, my.function)

# stop cluster when finished
stopCluster(cluster)
```

Get tags function

```
# function to get all nearby tags (nucleosome, polymerase II or H3K4me3)
getTags <- function(chr, mid_point, strand, con){

        readStart <- mid_point-1002
        readEnd <- mid_point+1002

        tags <- dbGetQuery(con, paste(
                "select start from nucleosome_tags where chr=",chr,
                " and strand=",strand,
                " and start>=",readStart,
                " and start<=",readEnd,
                sep=""
        ))
        if(length(tags)>0){
                return(tags[,1])
        }else{
```

```
		return(c())
	}
}
```

Count tags function

```
# function to tally the tags and their positions with respect to each gene/CCG
countTags <- function(tags, mid_point){	# inputs: plus/minusTags, mid_point

	tagCount <- rep(0,2001)

	for (n in tags){
		i <- n-mid_point+1001
		if (i == 1){
			tagCount[(1):(3)] <- tagCount[(1):(3)]+1
		}else if(i == 2){
			tagCount[(1):(4)] <- tagCount[(1):(4)]+1
		}else if(i == 2000){
			tagCount[(1998):(2001)] <- tagCount[(1998):(2001)]+1
		}else if(i == 2001){
			tagCount[(1999):(2001)] <- tagCount[(1999):(2001)]+1
		}else if(i > 2 & i < 2000){
			tagCount[(i-2):(i+2)] <- tagCount[(i-2):(i+2)]+1
		}
	}
	return(tagCount)
}
```

Load output files

```
rm(list=ls())

allFiles <- list.files()
files <- grep("uc00", allFiles)

plusAll <- data.frame()
minusAll <- data.frame()

for(i in 1:length(files)){
	print(i)
	load(file=allFiles[files[i]])
	names(minusAllCounts) <- paste("X",c(1:2001),sep="")
	names(plusAllCounts) <- paste("X",c(1:2001),sep="")
	minusAll <- rbind(minusAll, minusAllCounts)
	plusAll <- rbind(plusAll, plusAllCounts)
	rm(minusAllCounts)
	rm(plusAllCounts)
}
```

Totalling nucleosome counts

```
# add up the nucleosome tag count at each bp position +/-1000 either side of the TSSs
plusTotals <- rep(0,2001)
for (j in 1:length(plusTotals)){
        plusTotals[j] <- sum(plusAll[,j])
}

minusTotals <- rep(0,2001)
for (j in 1:length(minusTotals)){
        minusTotals[j] <- sum(minusAll[,j])
}
```

Creating plots

```
# user inputs file name for saving and title for plot
cat ("enter file name to save plot as (eg plot.png):\n")
fileName <- scan ("", what=character(0), nlines=1)
cat ("enter plot title:\n")
title <- scan ("", what=character(0), nlines=1)

# for positioning x axis labels
AT <- c(0, 500, 1000, 1500, 2000)

# vector of labels for x axis
X <- c(-1000, -500, "CCG", 500, 1000)

# set dimensions of graphic file
png(filename=fileName, width=1260, height=640)

# plot vectors
matplot(minusTotals, type = "l", col = "green3", ylab="frequency", xlab="base pair", xaxt="n", main = title)
matplot(plusTotals, type = "l", col = "red", add = TRUE)

# set custom x axis
axis(1, at = AT, labels = X, tick = TRUE)
dev.off()
```