The Genomic Architecture of Nucleolar Organiser Regions on the Short Arms of Human Acrocentric Chromosomes

Sofia Nazaré de Pereira Barreira

A thesis submitted to the

School of Mathematics, Statistics and Applied Mathematics
National University of Ireland, Galway

In fulfilment of the requirements for the degree of
Doctor of Philosophy

Under the supervision of
Professor Cathal Seoighe
Professor Brian McStay

September 2015

Contents

List of Figures

List of Tables

Abstract

Nucleolar Organiser Regions (NORs) are comprised of ribosomal gene (rDNA) arrays and adjacent sequences. Nucleoli, the sites of ribosome biogenesis and key regulators of cellular growth and proliferation, form around NORs. In humans, NORs are positioned on the short arms of the five acrocentric chromosomes (13, 14, 15, 21 and 22). These chromosome arms are not included in the human reference genome and have only recently started to be mapped and characterised.

This thesis has focussed on contributing to the characterisation and extension of these underexplored genomic regions. Previous work had suggested that as many as one third of rDNA repeats are rearranged. These could impact on nucleolar and ribosomal formation and protein synthesis. By performing next generation sequencing on DNA extracted from purified nucleoli, I demonstrated that there is no evidence for rearranged rDNA repeats in human cell lines. This conclusion was emphasised by a detailed analysis of more recent long read DNA sequence data sets. The second objective of this thesis was to describe the spatial organisation of sequences distal to the clusters of rDNA repeats. These sequences exhibit a euchromatic-like chromatin organisation, are transcriptionally active and appear to function as an anchor for the linked rDNA array during interphase. In the post genomic age, much effort now focuses on describing the chromatin status and 3D organization of the genome in a variety of human cell types and it is common practice to make the raw sequencing data from these genome-wide studies publicly available. Exploiting Hi-C data sets

designed to capture genome organisation revealed the existence of a transcription dependent stem-loop structure encompassing over 200 kb of NOR distal sequence that may play a role in NOR regulation. The third objective was to extend the sequences distal to NORs and characterise them. Using a combination of nucleolar sequencing reads and Hi-C data, this region was extended by 180 kb. Analysis of data from the ENCODE project suggests that this region is transcriptionally active and marks the beginning of interchromosomal variability on the short arms of acrocentric chromosomes.

These results provide a platform for investigating the role of NORs in nucleolar formation and maintenance and serve as a starting point for the identification and characterisation of the unknown regions of the p-arms of acrocentric chromosomes.

Acknowledgements

I would like to thank my two supervisors Professor Cathal Seoighe and Professor Brian McStay, for their guidance, advice and patience. Their knowledge and generosity was truly inspiring.

My colleagues and friends, Chelly, Michael, Mayo, Alice, Peter, Alan, Simone, Liam, Aisling, Hazel, Ngoc, Yaxuan, Barbara, Teri, Joseph, Thong, Ioanna, Paul K. and Paul G., Suraya and Martin, thank you for everything but most importantly, thank you for the laughs. Chelly, Michael, Mayo and Alice, I still don't know how you put up with me all these years.

I would also like to thank Professors Kevin Sullivan, Andrew Flaus and Uri Frank, for the support and guidance.

To my Irish family, Chelly, John, Austin, Maura and Brendan, thank you for letting me be a part of your lives.

A huge thank you to Dr. Christine Schnitzler and Dr. Andy Baxevanis for making me believe in myself.

A big thank you to my Cranfield friends Teresia Karlsson, Julia Feichtinger and Tommaso Oggian. Your emails and text messages always arrived when I most needed.

E o agradecimento mais importante, para os meus pais e irmã. O vosso apoio, amor e carinho são sem dúvida a única razão para eu triunfar na vida.

1 Introduction

1.1 Human Reference Genome

The human reference genome assembly is a collection of the nucleotide sequences of the human genome (Lander et al., 2001; Wright et al., 2001). It contains more than 3 billion base pairs that have been assigned to all human chromosomes (Lander et al., 2001; Wright et al., 2001). Numerous strategies were employed to achieve the final draft, which has been steadily updated over the years. Many of the sequence gaps have been bridged in the current version, GRCh38 (Cunningham et al., 2015; Miga et al., 2014).

1.1.1 Creation of the human reference genome

The sequencing and assembly of the human reference genome was carried out by an international research collaboration known as the Human Genome Project (Adekoya et al., 2001). The majority of the DNA sequencing was initially carried out by Sanger sequencing (Anderson, 1981). Sanger sequencing is a chain termination technique that involves synthesis of new strands from template DNA using dideoxynucleotides (Sanger et al., 1977). Lanes on polyacrylamide gels are used to order the new dideoxynucleotide-terminated strands and construct a consensus of the template DNA (Sanger et al.,

1977). The introduction of pyrosequencing (454) greatly decreased the cost and production time and increased the yield of DNA sequencing (Prober et al., 1987; Ronaghi et al., 1998). 454 sequencing technology works by placing small beads in a water-in-oil emulsion. DA fragments that have been nebulised (for size selection) and adapter ligated are fixed to these beads and PCR-amplified (Voelkerding et al., 2009). The DNA-bead complexes are placed in wells with enzymes and sequencing occurs by adding nucleotides in a previously established order. The addition of nucleotides creates a signal that is captured on camera and identified (Voelkerding et al., 2009). 454 sequencing ensures high quality reads by generating millions of identical copies with PCR. This can hinder the sequencing of repetitive DNA and genomes with high GC content (Hommelsheim et al., 2014). Large homopolymers stretches are also not well resolved by this technique (Margulies et al., 2005). Single molecule real time sequencing (SMRT) also known as PacBio sequencing is a DNA sequencing technique that uses a single DNA fragment as template per sequencing read produced (Eid et al., 2009). Sequencing occurs inside many zero-mode waveguide (ZMW) wells containing a DNA polymerase enzyme and a DNA fragment. Incorporation of phospholinked nucleotides results in the cleavage of the dye molecule and the phosphorescent signal is identified by a detector that assigns nucleotides accordingly (Levene et al., 2003). PacBio has the advantage of producing long reads (~ 10 kb) that can span and resolve small repeats. As a single template is used, PacBio has an average error rate of 13% (Quail et al., 2012). Currently, Illumina and more recently PacBio sequencing are the preferred methods for genome sequencing (Bennett, 2004; Bentley et al., 2008; Levene et al., 2003).

Initially, before the sequencing step, DNA was cloned in plasmid vectors, bacteriophages or phagemids, however this only yielded molecules smaller than 10 kb. Polymerase chain reaction (PCR) was used to obtain single stranded DNA for sequencing using specific primers to generate the target DNA temple (Scharf et al., 1986). The genome was assembled using mainly a combination of the clone contig approach and the whole genome shotgun approach (Fig. 1.1).

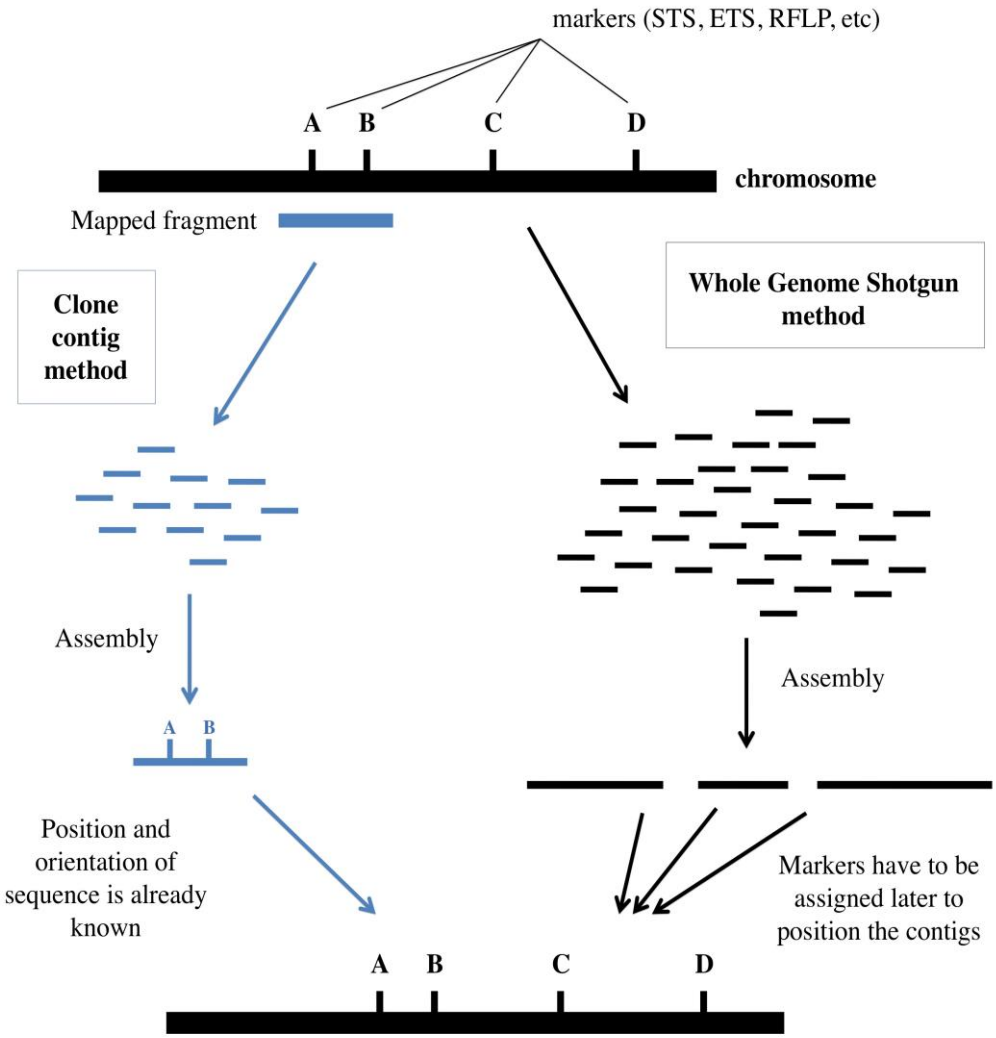

Figure 1.1 – Workflow of Clone contig sequencing method and whole genome shotgun sequencing method for *de novo* assembly of large genomes (Weber and Myers, 1997).

In the shotgun sequencing strategy a final sequence is constructed by overlapping sequencing reads from cloned fragments of a larger DNA segment

(Anderson, 1981). Extracted DNA was sonicated to randomly fragment the genome. Gel electrophoresis was employed to select fragments less than 20 kb that could then be amplified in plasmids, phages or cosmids. End sequences obtained from the clones were sequenced and assembled into unique contigs representing regions of the genome by overlapping the reads (Weber and Myers, 1997). The clone contig approach used restriction and physical maps to guide the placement and orientation of the assembled contigs. The genome was broken into fragments, preferably up to 1.5 Mb using restriction enzymes. The fragments were cloned in vectors such as yeast artificial chromosomes (YACs) or bacterial artificial chormosomes (BACS) (Monaco and Larin, 1994; Shizuya et al., 1992). YACs yield up to 1000 bp segments of DNA, although, exact replication of the inserted DNA is hindered by deletions and rearrangements that produce chimeric artefacts of the original sequence (Burke et al., 1987; Green and Olson, 1990; O'Connor et al., 1989). BACs can hold up to 300 kb and are more stable than YACs with fewer occurrences of rearrangements (Morrow et al., 1974; O'Connor et al., 1989; Stone et al., 1996). After cloning, the fragments are sequenced and contigs formed by identifying overlapping reads. The cloning contig approach uses prior knowledge of physical, restriction and/or genetic maps to guide the overlapping, orientation and positioning of the clones along the chromosomes (Cohen et al., 1993; Donis-Keller et al., 1987; Gyapay et al., 1994; Hudson et al., 1995; Osoegawa et al., 2001). Restriction maps were generated by digesting clones with restriction enzymes (Schwartz et al., 1993). The resulting products are separated by electrophoresis. Two YAC or BAC clones containing overlapping regions are identified by the common bands. Physical maps were created with sequence tagged sites (STSs), expressed sequence tags (ETSs) and

microsatellites. These are unique sequences easily amplified by PCR whose locations in the genome are known (Adams et al., 1991; Hudson et al., 1995). BAC clones were also used to complete chromosomal regions by employing chromosome walking. To solve gaps from regions where markers could not be placed, paired-end sequencing libraries were generated. Paired-end reads help define orientation and positioning of assembled contigs (Roach et al., 1995).

1.1.2 Characteristics and types of variation

The extent of sequence similarity between any two individuals is believed to be around 99.9% (Adekoya et al., 2001). The dissimilarities that occur in the genomes include changes in the structure (structural variations, > 3 kb) and quantity of chromosomes, such as rearrangements and heteromorphisms (Bobrow et al., 1971; Kim et al., 1999; Maegenis et al., 1978), which can be observed at the microscope. Smaller scale differences, mainly observable through DNA sequencing, constitute the majority of genome variation. These include single nucleotide polymorphisms (SNPs), insertions and deletions (indels) of base pairs, inversions and duplications and various repetitive short DNA sequences (micro and minisatellites) (Korbel et al., 2007; Verma et al., 1978). Genomic variations impacts on gene expression (Stranger et al., 2007) and may also cause health disorders such as velocardiofacial syndrome (Freeman et al., 2006; Lupski and Stankiewicz, 2005).

1.2 Functional Organisation of the Human Genome

After completion of the human genome, the next step was to identify and annotate all genes and functional elements (Birney et al., 2007). Importantly, comparative genomics studies revealed that the majority of the genome (~99%) consisted of non-coding sequences that also included trait-associated loci involved in disease and susceptibility (Kleinjan and van Heyningen, 2005; Lander et al., 2001; Lindblad-Toh et al., 2011; Ponting and Hardison, 2011). The ENCODE project was created to determine the nature and role of the non-coding regions of the human genome (Consortium, 2004). An array of elements, such as promoters, non-coding RNAs and histone modifications, were known to influence gene regulation (Birney et al., 2007). New technologies and strategies, such as ChIP-seq and RNA-seq were developed to aid in the identification of novel regulatory elements (Johnson et al., 2007; Morin et al., 2008).

1.2.1 Chromatin structure

The chromatin of eukaryotes has multiple levels of organisation. The first one is the nucleosome (Olins and Olins, 1974). Nucleosomes are a chromatin structure comprised of an octamer of four core histones (H2A, H2B, H3, and H4) and 146 bp of DNA wrapped around in a 1.75 turns (Finch et al., 1977; Kornberg, 1974; Luger et al., 1997). The N-terminal tails of histones are subjected to a vast number of modifications, such as phosphorylation, methylation and acetylation (Chen et al., 1999; Pokholok et al., 2005). These are

crucial to the control of transcription activation and repression (Han and Grunstein, 1988; Lorch et al., 1987). Arrays of nucleosomes known as 10-nm 'beads-on-a-string' make up the next level of DNA packaging (Kornberg, 1974; Olins and Olins, 1974). In mitotic chromosomes, the 10-nm fibres are coiled in a fractal manner (Nishino et al., 2012). During interphase, the local structure of chromatin is observed as euchromatin and heterochromatin (Kustatscher et al., 2014). Euchromatin is lightly packaged and closely associated with RNA polymerase, whereas tightly packed heterochromatin generally comprises inactive regions or structural regions such as telomeres and centromeres (Raisner et al., 2005; Sullivan and Karpen, 2004).

1.2.2 Histone modifications and chromatin modulation

Histone modifications are involved either in activation or repression of transcription (Table 1.1)(Bannister et al., 2001; Barski et al., 2007; Benevolenskaya, 2007; Bernstein et al., 2006; Birney et al., 2007; Guenther et al., 2007; Heintzman et al., 2007; Joshi and Struhl, 2005; Ong and Corces, 2011; Schotta et al., 2004; Talasz et al., 2005; Wang et al., 2008). Different types of histone modifications are responsible for influencing gene expression (transcription activation/repression), DNA replication and repair, and chromatin condensation (Barski et al., 2007). The amino acids lysine (Lys or K) and arginine (Arg or R) are usually the main target for modifications in histones (Allfrey et al., 1964; Li et al., 2007; Wang et al., 2008). Modifications in serine

(Ser or S), threonine (Thr, T) and Tyrosine (Tyr, Y) also occur (Daujat et al., 2005; Dawson et al., 2009; Kim et al., 2013).

Table 1.1 - List of common histone modifications and associations

Histone Modification	Transcription regulation	Association
H3K4me1	Activation	Enhancers
H3K4me2	Activation	Promoters and enhancers
H3K4me3	Activation	Promoters
H3K9me2	Repression	Heterochromatin
H3K9me3	Repression	Heterochromatin
H3K9ac	Activation	Active regulatory regions
H3K27me1	Activation	Euchromatin
H3K27me2	Repression	Polycomb-repressed regions
H3K27me3	Repression	Polycomb-repressed regions
H3K36me3	Activation	Transcribed regions
H4K20me1	Activation	Transcribed regions
H3K27ac	Activation	Active regulatory regions

Methylation of Histone H3 in the 4th lysine is strongly associated with transcription activation. Tri-methylation (H3K4me3) is associated with strong promoters of active genes and early-transcribed regions (Heintzman et al., 2007; Wang et al., 2008). Di-methylation (H3K4me2) is also bound to promoters and as mono-methylation (H3K4me1) is preferentially associated to enhancers (Ong and Corces, 2011; Wang et al., 2014). Acetylation of the 9th lysine in H3 (H3K9ac) is indicative of functional regulatory elements and may contribute to

gene activation and chromatin remodelling (Roh et al., 2005). Acetylation of H3K27 is also associated with transcriptional activation and it helps to differentiate between active and poised enhancers (Creyghton et al., 2010; Suka et al., 2001; Wang et al., 2008). Tri-methylation of lysine 36 on H3 (H3K36me3) and mono-methylation of lysine 20 in H4 is associated with transcribed regions binding extensively across the gene body (Joshi and Struhl, 2005; Schotta et al., 2004; Talasz et al., 2005). Importantly, ChIP-seq peaks of H3K36me3 preceded by peaks of H4K4me3 are indicative of the existence of transcripts transcribed by RNA Polymerase II (Guttman et al., 2009). For repression of transcription, tri-methylation of H3K9 is a marker for heterochromatin, and also plays an important role in the formation of heterochromatin and gene silencing in repetitive sequences (Bannister et al., 2001). H3K27me3, also a marker of inactive genes is associated with Polycomb-repressed regions prompting gene silencing (Boyer et al., 2006; Roh et al., 2006).

A study by Ernst *et al*, inferred through ChIP-seq analysis the existence of six states of chromatin depending on its profile: promoter, enhancer, insulator, transcribed, repressed, and inactive (Ernst et al., 2011). The promoter state can further be divided into strong, weak and poised depending on their expression levels and enhancers differing on expression of proximal genes are called strong and weak candidate. Transcribed regions can have strong or weak transcript enrichment and Polycomb repressed regions can be heterochromatic or repetitive, with this last state being enriched for H3K9me3 (Ernst et al., 2011). The typical length of the states and their coverage of the genome also varies. Promoters and enhancers average 500 bp and representing less than 1% of the genome while inactive regions cover more than 70% with an average length of

10 kb. Interestingly, the location of these states varies between cell lines (Ernst et al., 2011). ChIP-seq can also be used to identify promoters and insulators through the binding of RNA Pol II and the transcription factor CTCF, respectively (Kim et al., 2007; Kim et al., 2005).

1.3 Genome packaging

Regulation of gene expression in the human genome is not restricted to histone modifications, the location of specific proteins and regulatory sequences. The spatial conformation of chromatin within the nucleus also modulates the expression of genes (Bickmore and van Steensel, 2013; Finlan et al., 2008). The view that the positioning of chromatin in the interphase nucleus is functionally relevant for transcription has been discussed for a few decades (Blobel, 1985; Hilliker and Appels, 1989; Vogel and Schroeder, 1974). Numerous techniques such as ChIP, DamID and FISH revealed genetic loci have a non-random positioning within the human nucleus (Bickmore and van Steensel, 2013; van Steensel et al., 2001). The DamID technique, for mapping chromatin-associated proteins, has shown that there are around 1,400 lamina-associated domains (LADs), up to 10 Mb in length, in mammalian genomes (Kind et al., 2013; van Steensel et al., 2001). These interphase contacts consist of heterochromatin, marked with H3K9me2, and have constrained mobility, rarely mixing with nearby euchromatin (Kind et al., 2013). Chromosomes in the interphase nucleus are organised into chromosomal territories (CTs)(Fig. 1.2).

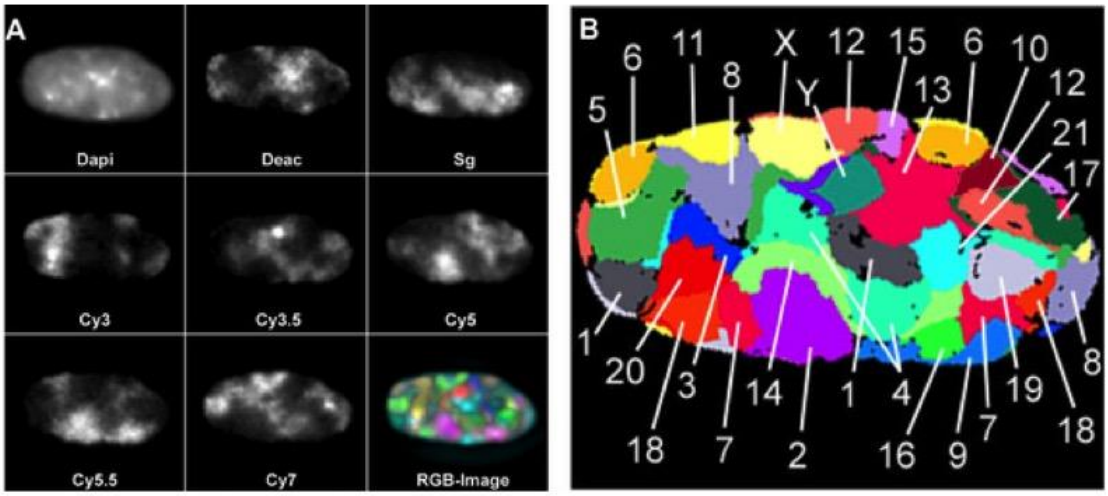

Figure 1.2 – Colour 3D FISH representation and classification of chromosomes in a human G0 fibroblast nucleus. A- A deconvoluted mid-plane nuclear section recorded by wide-field microscopy in eight channels: one channel for DAPI (DNA counter stain) and seven channels for the following fluorochromes: diethylaminocoumarin (Deac), Spectrum Green (SG), and the cyanine dyes Cy3, Cy3.5, Cy5, Cy5.5, and Cy7. Each channel represents the painting of a CT subset with the respective fluorochrome. RGB images of the 24 differently labeled chromosome types (1–22, X, and Y) were produced by superposition of the seven channels (bottom right). B - False color representation of all CTs visible in this mid-section after classification with the program goldFISH. Figure from (Bolzer et al., 2005)

FISH combined with high-resolution microscopy revealed CTs have a diameter between 1 to 3 μm, are irregularly shaped and are comprised of much smaller subdomains (Cremer and Cremer, 2001; Gilbert et al., 2005; Parada et al., 2002). CTs from neighbouring chromosomes intermingle at their periphery (Branco and Pombo, 2006) and are passed from parent to daughter cell in a semi conserved manner (Parada et al., 2002). The local positions of large areas of identity in CTs maintain their nuclear positions in different species (Tanabe et al., 2002). Proximity patterns for CT clusters vary between cell types from different tissues (Parada et al., 2004). This strengthens the concept that nuclear architecture provides an additional level of epigenetic regulation and genome maintenance depending on the transcriptional needs of the cell type (Cremer et al., 2006; Parada et al., 2004). The folding of individual loci can be observed at

low resolution by light microscopy and electron microscopy offers high-resolution but lacks connectivity to the DNA sequences (Schermelleh et al., 2010). Although FISH enables visualisation of multiple loci the treatment might influence chromosomal conformation (Lichter et al., 1988; Pinkel et al., 1986; Schermelleh et al., 2010). The chromosome conformation capture (3C) technique was developed to study the spatial organisation of chromosomes in their natural state at high resolution (Dekker et al., 2002). 3C can be used to create a 3D cast of nuclear structure. Initially applied to the yeast genome where it showed chromosome 3 forms a contorted ring (Dekker et al., 2002). 3C revealed chromatin loops connect regulatory sequences and their target genes in mammalian genomes (Tolhuis et al., 2002). Remarkably, 3C enabled the identification of enhancers previously unknown to regulate the CFTR gene (Gheldof et al., 2010). Control of gene expression by enhancers can be blocked by insulator sequences bound by proteins such as CTCF (Phillips and Corces, 2009; Wallace and Felsenfeld, 2007). Chromatin loops are formed by CTCF sites in contact with each other (Splinter et al., 2006; Zhao et al., 2006). CTCF also recruits other factors to its binding sites including cohesin and TAF3, which are thought to help in the formation of loops (Hadjur et al., 2009; Parelho et al., 2008; Wang et al., 2011; Wendt et al., 2008).

However, the 3C method is hindered by the limited number of loci that it can target. Numerous methods have been developed based on the 3C technology, such as 4C and 5C (Dostie et al., 2006; Simonis et al., 2006). Hi-C is a comprehensive technique to infer the 3D architecture of chromatin in a genome-wide fashion with high resolution (Lieberman-Aiden et al., 2009). Through the use of high-throughput sequencing, Hi-C allows the inquisition of all possible

contacts in a genome at 1 kb resolution (Jin et al., 2013; Rao et al., 2014). The usage of FISH and multiple methods for chromosome conformation capture revealed that the human genome is organised in topologically associated domains (TADs) of sizes ranging from 100 kb to 1 Mb (Lieberman-Aiden et al., 2009). Interestingly, the position of TADs in the genome is generally conserved between cell types (Dixon et al., 2012). The size of TADs resemble the size of replication domains and their boundaries are enriched for SINEs, transfer RNAs, housekeeping genes, and CTCF, and also correlate with regions that halt the spread of heterochromatin (Dixon et al., 2012; Pope et al., 2014).

Currently, the functional nuclear organisation model is described as the 4D nucleome (Chen et al., 2015; Tashiro and Lanctot, 2015). To reconcile the nuclear space-time organisation with function, an integrative model has been proposed. Chromatin is organised into two co-aligned network compartments, the active nuclear compartment (ANC) and the inactive nuclear compartment (INC). The INC is formed by the transcriptionally inactive and tightly packed core of chromatin domain clusters (CDCs)(Fig. 1.3).

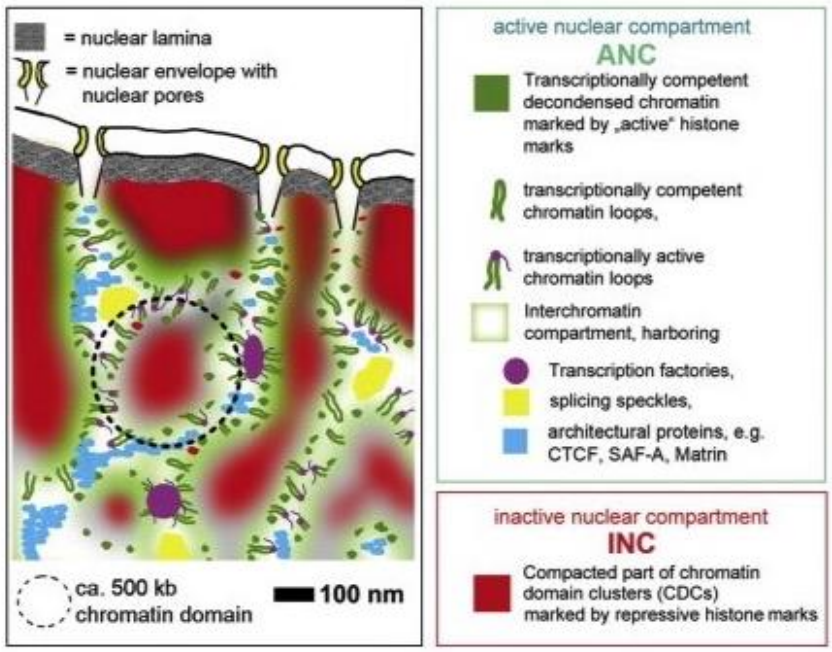

Figure 1.3 - ANC-INC network model of nuclear organization based on spatially co-aligned active and inactive nuclear compartments. Nuclear organization according to co-aligned 3D networks of an active (ANC) and an inactive nuclear compartment (INC). The ANC is a composite structural and functional entity of a 3D-channel network, the ''Interchromatin-Compartment'' (IC) together with the decondensed periphery of a higher order chromatin network, which pervades the nuclear space and is built up from chromatin domain clusters (CDCs). The decondensed periphery of CDCs is known as the perichromatin region (PR). Reprinted from FBES Letters, Cremer et al, "The 4D nucleome: Evidence for a dynamic nuclear landscape based on co-aligned active and inactive nuclear compartments", Copyright (2014), with permission from Elsevier (Cremer at al 2015).

The ANC comprises the perichromatin regions, which is the transcriptionally active periphery of CDCs and the interchromatin compartment, IC. ICs are a network of channels, mostly devoid of DNA, that start and end at nuclear pores that pervade the heterochromatin layer underneath the nuclear lamina and extend between and within CTs (Albiez et al., 2006; Hubner et al., 2013). ICs contain the transcription and splicing machineries and DNA replication and repair complexes (Albiez et al., 2006; Hubner et al., 2013; Markaki et al., 2010). The IC is outlined by the perichromatin region (PR), which corresponds to decondensed chromatin containing the coding and regulatory sequences of active genes (Hubner et al., 2013; Rouquette et al., 2009; Smeets et

al., 2014). The PR is highly enriched for histone modifications associated with transcription activation and for RNA Pol II (Markaki et al., 2010; Niedojadlo et al., 2011). In contrast, the CDCs are characterised by histone modifications associated with transcriptionally silent chromatin (Markaki et al., 2012; Popken et al., 2014; Smeets et al., 2014).

1.3.1 Genome packaging in disease

The spatial organisation of the nucleus has been linked to the control of gene expression, DNA replication and DNA repair (Burgess et al., 2014; Nagano et al., 2013). Gene expression is influenced by numerous regulatory elements, which can be located in close proximity or megabases upstream or downstream of their target genes (Miele and Dekker, 2008; Montgomery et al., 2010). Defects in chromatin organisation are a potential cause for disease (Misteli, 2010). Multiple single nucleotide polymorphisms (SNPs) have been associated to numerous diseases and regulatory pathways in genome-wide association studies (GWAS) (Pomerantz et al., 2009; Zhang et al., 2012a). Recently, promoter capture Hi-C has revealed that regions that interact with promoters are highly enriched for SNPs that have been associated with disease (Jager et al., 2015). This means that defective regulatory elements of the genome could be linked with the genes and molecular pathways they influence (Jager et al., 2015). CTCF, Cohesin and SATB1 are, among other roles, chromatin organisers and have been implicated in numerous human diseases (Han et al., 2008; Libby et al., 2008; Tonkin et al., 2004; Vega et al., 2005; Witcher and Emerson, 2009). It is

possible that disrupting the manner in which chromatin folds affects the action of enhancers and insulators, disrupting gene expression. A Hi-C study on Huntington-Guilford Progeria Syndrome showed that the accumulation of progerin in the nuclear lamina causes altered patterns of H3K27me3 and a loss of global spatial compartmentalisation organisation that may result in transcriptional misregulation (McCord et al., 2013). Changes in chromatin organisation have been observed through Hi-C data after overexpression of ERG, an oncogenic transcription factor (Rickman et al., 2012). Overexpression of ERG in prostate cancers is the result of gene fusion (Pflueger et al., 2009). Failure to define gene boundaries, either by incorrect behaviour of chromatin insulators or by incorrect patterns of histone modifications is predicted to cause widespread erroneous gene expression (Wendt et al., 2008). Genomes of cancer cells often have an abnormal number of chromosomes (Nicholson and Cimini, 2013). This can potentially disrupt the cell type-established chromosomal territories and provoke changes in gene regulation.

1.4 Missing regions in the Human Reference Genome

Around 5-10% of the human genome is missing from the assembled reference genome (Altemose et al., 2014). Key missing regions include the centromeres, telomeres and short arms of acrocentric chromosomes. These areas correspond in their majority to heterochromatic regions comprised of repetitive sequences. Between 66% and 69% of the human genome is comprised of repetitive and repeat-derived elements (de Koning et al., 2011). The various

classes of repetitive DNA sequences include the telomere repeat, subtelomeric repeats, microsatellite and minisatellite repeats, Alu repeats, L1 repeats, alpha satellite DNA, satellite I, II and III repeats and cot1 DNA (Catasti et al., 1999). Repeats vary in length and periodicity (Fig. 1.4).

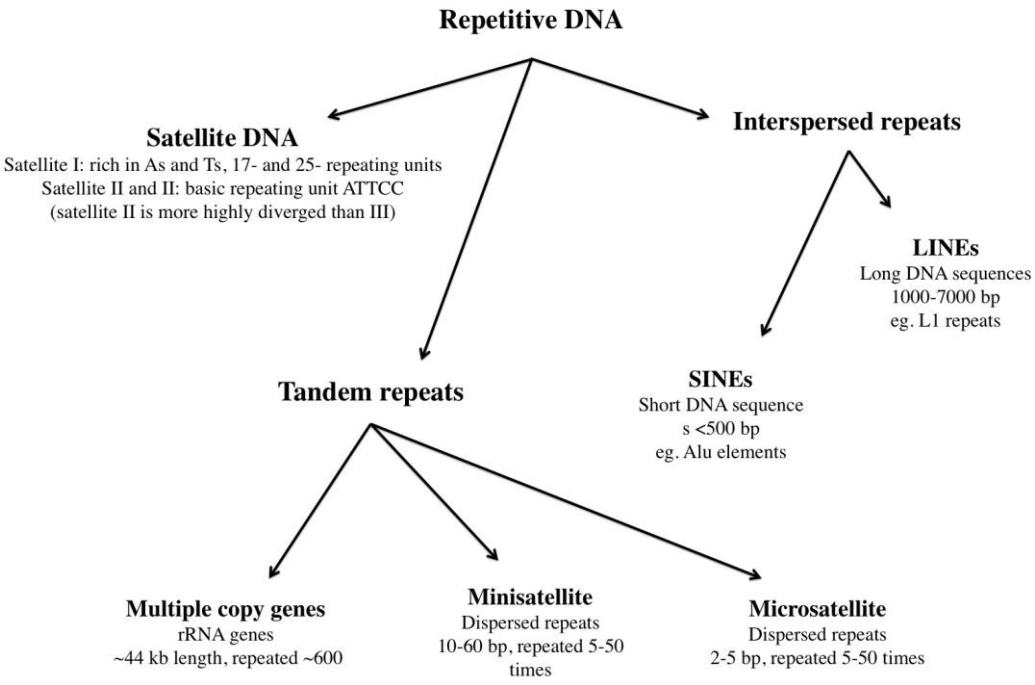

Figure 1.4 - Examples of repeats found in the human genome (Duitama et al., 2014; Jones and Corneo, 1971; Smit, 1996; Sylvester et al., 1986).

Repetitive sequences are difficult to sequence and assemble. Some sequencing technologies employ PCR amplification steps to increase the library input. Repeats with high GC content are difficult to sequence with these technologies due to the tendency of these regions to form secondary structures, such as hairpins, and the inefficient incorporation of dye terminators (Aird et al., 2011; Chen et al., 2013; Kozarewa et al., 2009).

Although the accuracy of sequencing reads is improving and their lengths increasing, repetitive regions remain a challenge for most assembly algorithms

(Miller et al., 2010). If the repeats are longer than the reads, the repeats will not be resolved and will usually collapse into a single sequence. This behaviour by the assembler can help identify repeating sequences, as they will have a higher mapping coverage than non-repetitive regions. If the repeats are inexact, careful alignment with high coverage might separate the repeat copies into different sequences but it is dependant on the high accurate reads. Paired-end reads can help resolve repeats longer than the reads but if the periodicity of the repetitive sequences is high, different insert libraries are needed. Repeats present in different chromosomes also hinder assemblies as they might be assembled together due to their similarity and insert incorrect junctions in the contigs.

The short arms of the five human acrocentric chromosomes remain unsequenced and missing from the current human reference genome due to their repetitive nature. Different kinds of repetitive elements can be found in them, such as large tandem repeats (ribosomal genes ~44kb), segmental duplications (<130 kb), and telomeric and centromeric repeats. Although unassembled, the short arms of acrocentric chromosomes contain crucial features to the proper functioning and survival of the cell, the nucleolar organiser regions (NORs). NORs are the site of formation and maintenance of nucleoli (McClintock, 1934; Pederson, 2011). The nucleolus is a key regulator of cellular growth and responsible for ribosome biogenesis (Henderson et al., 1972; McConkey and Hopkins, 1964).

As the main theme of my research thesis, is to explore these missing regions of the genome, in the next sections I will present a brief review of what is known about the biology of the nucleolus and the genomic architecture of nucleolar organiser regions.

1.5 Nucleolus

In human cells, nucleoli form around the arrays of ribosomal genes positioned on the short arms of the five acrocentric chromosomes (Bloom and Goodpasture, 1976; Henderson et al., 1972). The largest and densest of the nuclear compartments, the nucleolus is a cytogenetic entity responsible for the synthesis of ribosomal RNA and the assembly of ribosomes (Brown and Gurdon, 1964). Other nucleolar functions include cell cycle progression, DNA replication and DNA repair, the sensing of cellular stress, replication of viral DNA, cell survival and initiation of apoptosis, and the assembly of signal recognition particles (Boisvert et al., 2007; Boulon et al., 2010; Pederson, 2011). Nucleoli are composed of three parts defined by their visual appearance in electron microscopy (Figure 1.5).

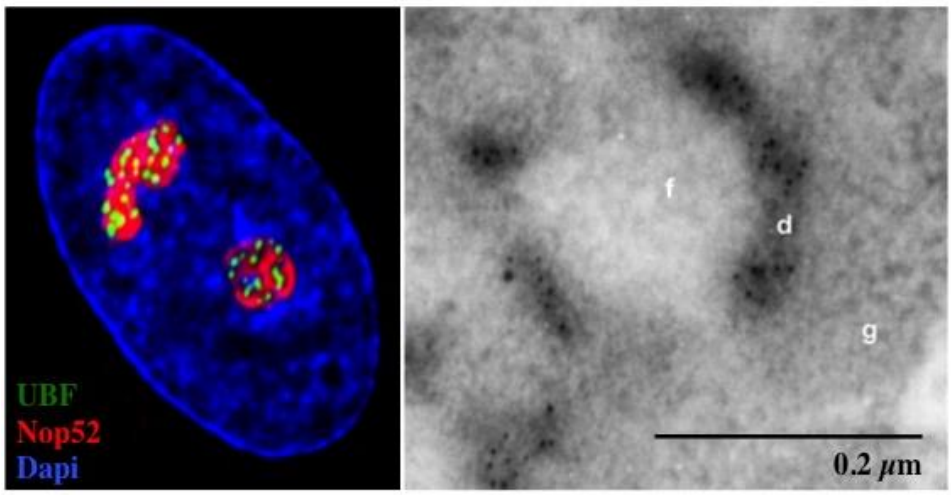

Figure 1.5 – Internal structure of the nucleolus. Left image shows cells stained with UBF (in green) depicting the fibrillar centre and Nop52 (in red) staining for the granular centre (picture from Chelly van Vuuren). Right EM image from ©Koberna et al., 2002. Originally published in J Cell Biol. Vol 157, 743-748. (f) fibrillar centre, (d) dense fibrillar component and (g) granular component.

These distinct compartments, fibrillar centre (FC), dense fibrillar centre (DFC) and granular component (GC), also reflect the events taking place within them (Bernhard and Granboulan, 1963). Ribosomal genes are located in the fibrillar centre and transcription of pre-rRNA occurs at the boundary of the fibrillar centre and the dense fibrillar component (Koberna et al., 2002; Raska et al., 1989; Thiry and Lafontaine, 2005). Early processing of rRNA such as post-transcription modifications is observed at the dense fibrillar component and late rRNA processing and the beginning stages of ribosome assembly occur in the granular component (Koberna et al., 2002; Puvion-Dutilleul et al., 1997; Thiry and Lafontaine, 2005). Each of these compartments also contains the proteins and processing factors, such as UBF, necessary to perform these tasks (Hernandez-Verdun, 2011; Jantzen et al., 1990; Russell and Zomerdijk, 2006). In each cell cycle, nucleoli assemble and disassemble (Sirri et al., 2000). Upon mitosis and during prophase, transcription is inhibited and nucleoli disappear

(Gebrane-Younes et al., 1997). Nucleoli reappear at the end of mitosis in telophase (Muro et al., 2010). Transcription of rDNA resumes simultaneously in active NORs that subsequently fuse to form larger nucleoli (Roussel et al., 1996; Savino et al., 2001). Not all available repeats are active and contribute to nucleolar formation (Dammann et al., 1995). The activity status of rDNA is influenced by epigenetic switches (Lawrence and Pikaard, 2004) and various histone modifications occurring in the intergenic spacer (Zentner et al., 2011). In interphase cells, inactive NORs remain dissociated from nucleoli and can be observed as condensed foci devoid of UBF and Pol I (McStay and Grummt, 2008).

Nucleoli in cancer cells have been observed to have irregular shapes and be larger than nucleoli from healthy cells (Derenzini et al., 1998; Pianese, 1896). The high levels of rDNA transcription demanded by the elevated proliferation rates lead to the amorphously shaped nucleoli (Hanahan and Weinberg, 2011; Stults et al., 2009).

1.6 Nucleolar Organiser Regions

Chromosomes possess a primary constriction; the centromere. Human cells have an additional secondary constriction located on the short arms of the five acrocentric chromosomes (Fig. 1.6-A). These constrictions indicate the location of the nucleolar organiser regions around which nucleoli form (McClintock, 1934). NORs are conserved and shared across all acrocentric chromosomes and are responsible for the assembly and regulation of the

nucleolus (Floutsakou et al., 2013; McStay and Grummt, 2008; Pederson, 2011). A single NOR is comprised of tandem arrays of ribosomal gene repeats, rDNA, arranged in a head to tail orientation, and their adjacent sequences (Fig. 1.6-B) (Floutsakou et al., 2013; Henderson et al., 1972).

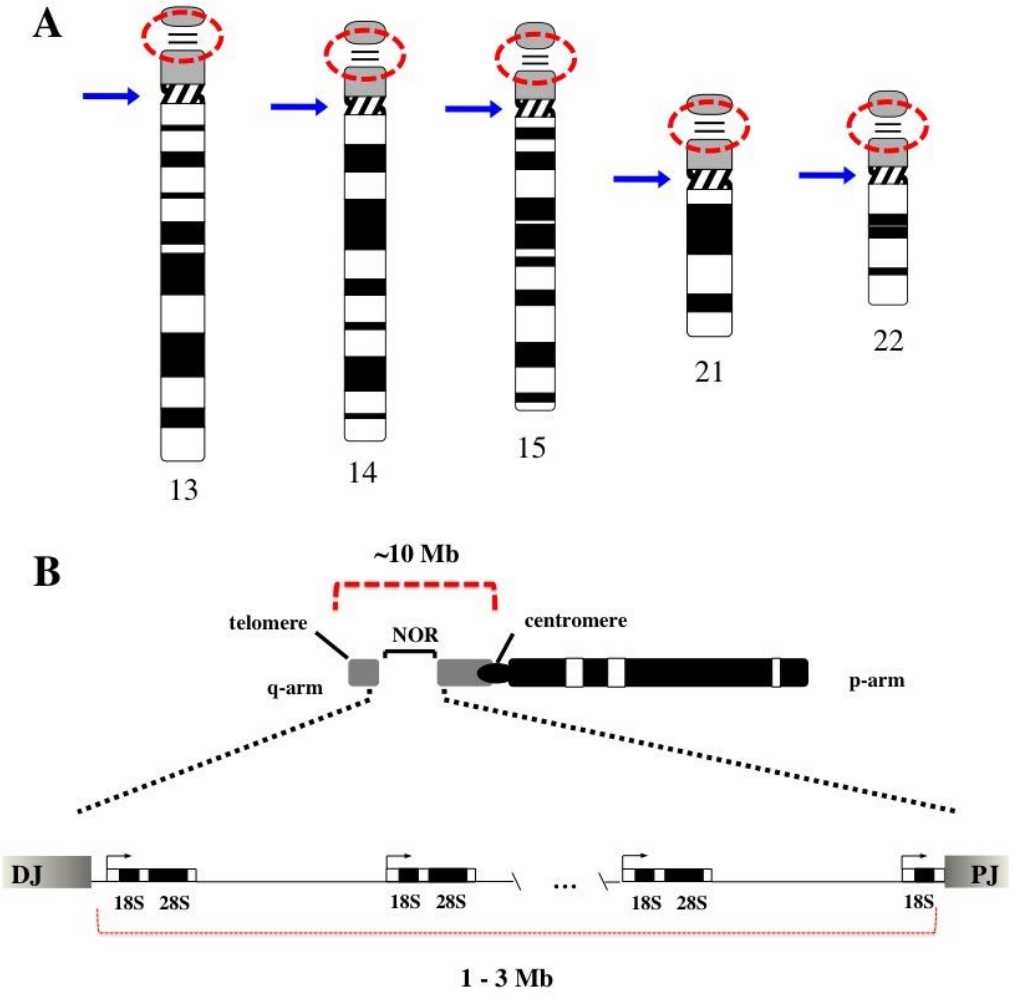

Figure 1.6 - The five human acrocentric chromosomes, 13, 14, 15, 21, and 22, have an asymmetric conformation due to the location of their centromeres near one end of the chromosome. A - Blue arrows indicate the location of centromeres and red dotted circles indicate the location of NORs in human acrocentric chromosomes. Chromosome pictures from Idiogram Album: Human copyright © 1994 David Adler. B – Nucleolar Organiser Regions are located in the short arms of the acrocentric chromosomes. NORs contain around 1 – 3 Mb of ribosomal gene clusters.

The rDNA gene contains the sequences for the 18S, 5.8S and the 18S, the ITS1 and ITS2 internal transcribed spacers and a non-transcribed intergenic spacer, IGS (Fig. 1.7)(Long and Dawid, 1980; Stults et al., 2008).

Figure 1.7 - Human rDNA repeat extracted from BAC AL592188 [105424 - 149395] bp. The entire repeat is almost 44 kb and contains the sequences for 18S, 5.8S and 28S ribosomal subunits in the first 13 kb followed by a large intergenic spacer that is not transcribed. Base pair coordinates of the rDNA components are, 5'ETS [1, 3654], 18S [3655, 5523], ITS1 [5524, 6600], 5.8S [6601, 6765], ITS2 [6766, 7924], 28S [7925, 12994], 3'ETS [12995, 13392].

The rDNA repeat is transcribed by RNA Polymerase I (Masson et al., 1996), unlike the majority of the transcriptome, which is transcribed by Pol II (Kedinger et al., 1970; Roeder and Rutter, 1969). The direction of transcription occurs from the telomere towards the centromere. Sequences on either side of the rDNA clusters (Fig. 1.8) are also part of nucleolar organiser regions and contribute to the formation and regulation of nucleoli (Floutsakou et al., 2013). These contigs were constructed and identified by screening cosmid libraries, sequencing clones, searching GenBank, and performing BAC walking.

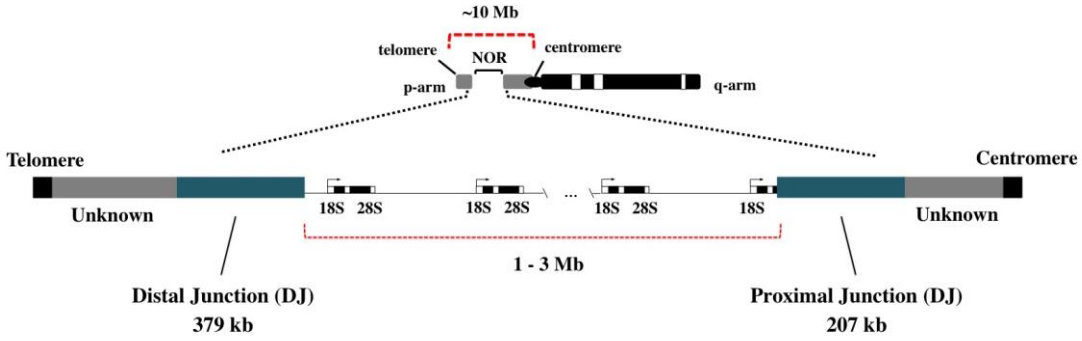

Figure 1.8 - Location of sequences identified adjacent to the rDNA repeats that also comprise NORs. The Distal Junction is located on the telomere side and is 379 kb in length. The Proximal Junction is positioned on the telomere side of rDNA and is 207 kb in length.

Work by Floutsakou *et al*, identified sequences adjacent to the rDNA clusters that are shared between all acrocentric chromosomes (Floutsakou et al., 2013). On the telomere side of the rDNA repeats, the Distal Junction, DJ, is almost 380 kb in length and possesses a large inverted repeat centred at a 6 kb spacer and arms length of ~109 kb and ~111 kb with 79.5% sequence identity between the two sequences (Fig. 1.9-A). Segmental duplications are common features in the human genome and highly enriched near centromeres (Bailey et al., 2001; She et al., 2004). The DJ shows a low degree of segmental duplication (7.3%) to the rest of the genome (Fig. 1.9-B). The Proximal Junction, PJ, is almost entirely segmentally duplicated (92.4%) with long segments mapping to peri-/centromeric regions (Fig. 1.9-B). Duplications in the DJ are short (no more than 5 kb with at least 85% identity) and are restricted to euchromatic and telomeric regions. The distal junction also displays transcriptional activity. Histone modifications for activation and repression of transcription were identified in the DJ through ChIP-seq, together with CTCF binding sites and Pol

II peaks. RNA-seq and RT-PCR revealed that there are spliced and polyadenylated transcripts originating in the DJ (Floutsakou et al., 2013).

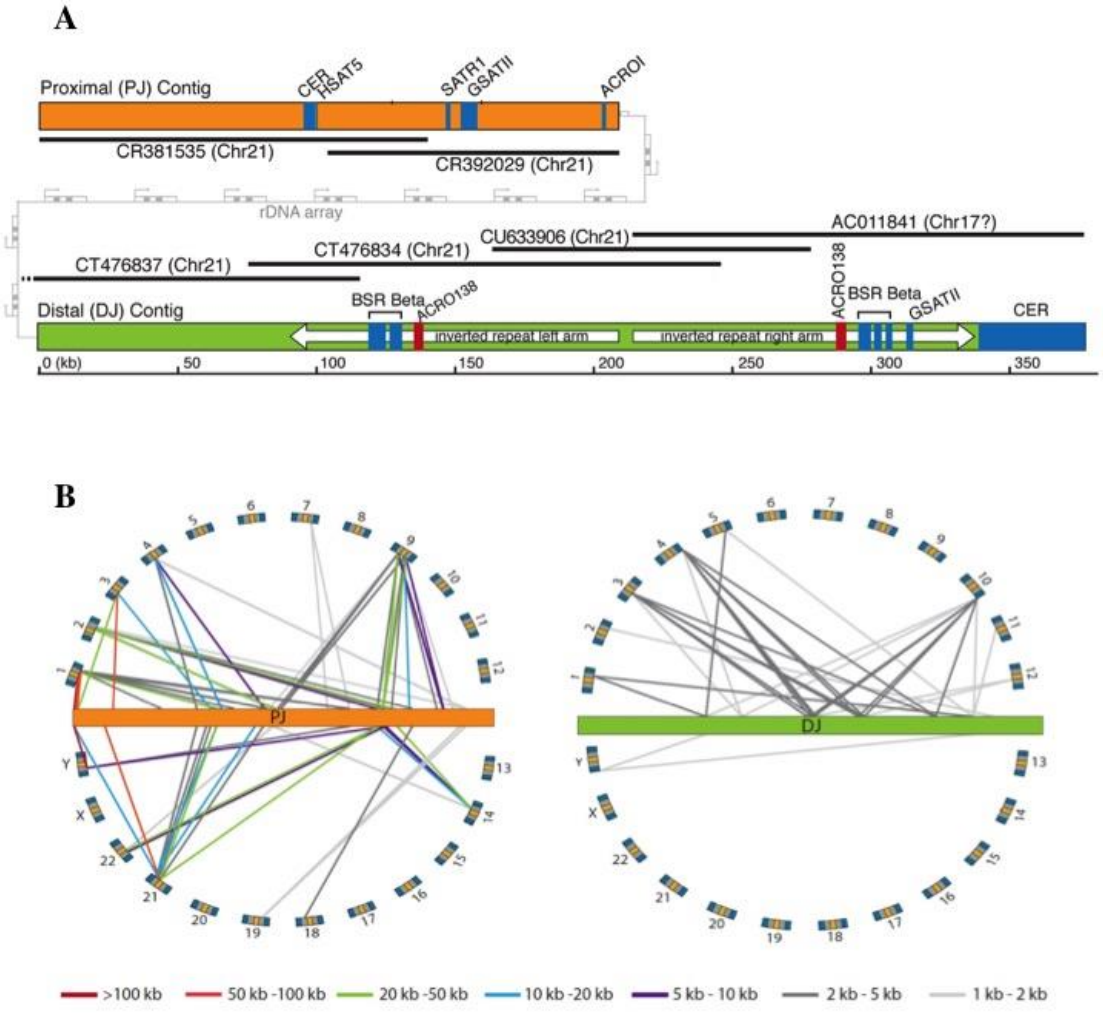

Figure 1.9 – Sequence characterisation of the DJ and PJ. A - Main genomic features of the PJ (orange) and DJ (green) contigs. Black lines indicate BACs used to construct the distal and proximal contigs with BAC names and chromosomal origins indicated (chr17 annotation of AC011841 is incorrect). White arrows indicate the position of the large inverted repeat in the DJ and in gray the rDNA array between the PJ and the DJ. Blue shows the location of satellite repeats. B – Analysis of segmental duplications. Lines coloured according to length of segments (below) connect the PJ (orange) and DJ (green) to the location of duplications in other human chromosomes. Figure from (Floutsakou et al., 2013).

During interphase, the DJs from acrocentric chromosomes that contribute to nucleoli relocate to the periphery of the nucleolus (Fig. 1.10). Significantly, inhibition of RNA Pol I, through double strand breaks or actinomycin D (Perry, 1962; Reich et al., 1961), leads to withdrawal of rDNA from the nucleolus

interior to form caps adjacent to its corresponding DJ (Floutsakou et al., 2013; Schofer et al., 1996).

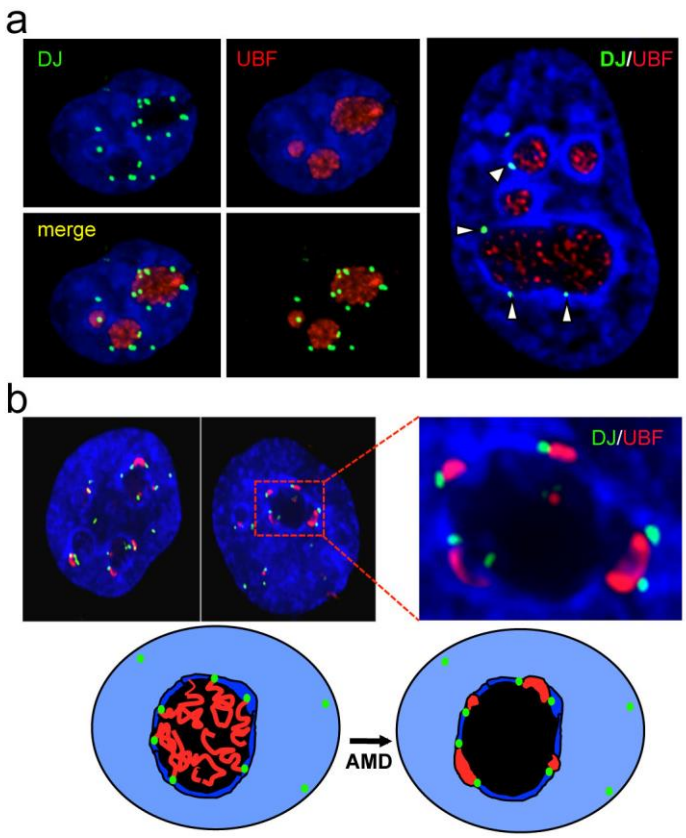

Figure 1.10 - Localisation of the DJ to the nucleolar periphery during interphase. A - The DJ (BAC CT476834 in green) acts as an anchor to the rDNA repeats (UBF antibody in red). B – Inhibition of transcription causes the rDNA to retreat to the DJ. Figure from (Floutsakou et al., 2013).

1.7 Technologies for functional and spatial organisation analysis

The following sections introduce relevant technologies, publicly available data sets, and cell lines utilised during the course of my thesis work.

1.7.1 Chromatin Immunoprecipitation sequencing, ChIP-seq

Chromatin immunoprecipitation followed by high-throughput sequencing, ChIP-seq, is a technique used to determine histone modifications or DNA binding sites for a protein of interest (Barski et al., 2007; Johnson et al., 2007). The ChIP-seq method starts with experimental steps to enrich for DNA that is bound to the protein of interest (Fig. 1.11).

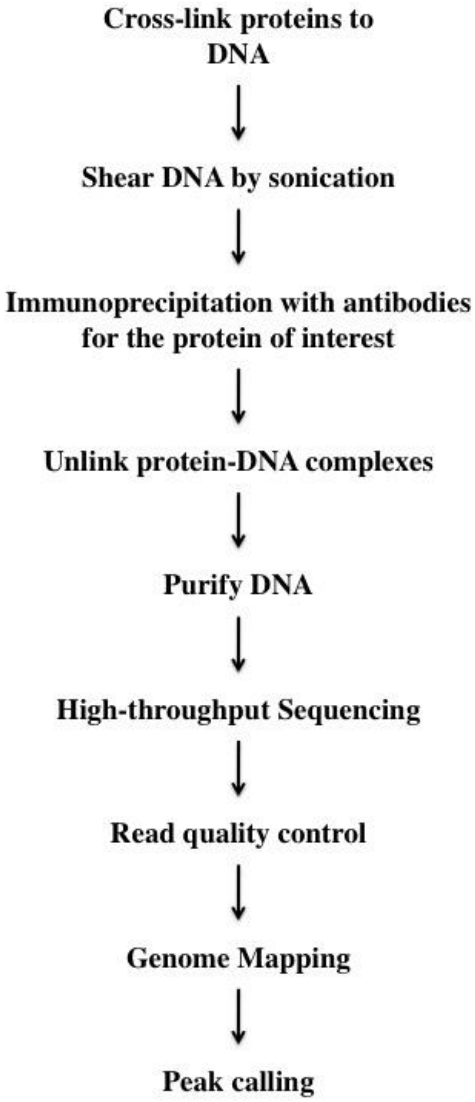

Figure 1.11 - Workflow of ChIP-seq method and analysis.

First, proteins are cross-linked to DNA with formaldehyde. Chromatin is then sheared with sonication or micrococcal nuclease. The protein-DNA complexes are immunoprecipitated with antibodies against the target protein. The cross-link in the selected complexes is reversed by heat followed by purification of DNA (Orlando, 2000). After size selection, DNA fragments are subjected to high-throughput sequencing. To identify sites in the genome, that are enriched for the protein or protein modification of interest, the sequencing reads are first quality controlled and filtered to ensure only high quality reads are used in the subsequent mapping step. Reads are then mapped to the genome of interest usually allowing only one or two mismatches per read. Reads that map to more than one location are discarded. This eliminates reads that map to repeats or that are not unique by chance (Johnson et al., 2007). The next step (peak calling) is carried out by finding local concentrations of reads. A target protein-binding site, peak, is called if the number of reads that constitute that peak surpasses the number of reads at the same location in the control sample by a defined threshold (Johnson et al., 2007).

1.7.2 Whole transcriptome shotgun sequencing, RNA-seq

Numerous methods, such as cDNA microarrays, were developed to measure gene expression and characterise transcription at the exon level and transcript levels (DeRisi et al., 1996; Saha et al., 2002; Velculescu et al., 1999).

Detection of transcripts by microarray analysis, however, is limited to the known genomic annotations (Ota et al., 2004). Quantifying and deducing RNA presence in a cell allows the identification of genes that are expressed in different cell types and states (Nagalakshmi et al., 2008). Transcriptome sequencing, or RNA sequencing (RNA-seq) is a technology that uses short sequencing reads to assess transcriptional start and end sites, transcript abundance, identification of novel exons and exon-exon attachment in matured transcripts, identification of new transcripts, SNPs and mutations, post-translational modifications and alternatively spliced transcripts (Morin et al., 2008; Nagalakshmi et al., 2008; Ozsolak and Milos, 2011; Wilhelm et al., 2008). RNA-seq can also be used to determine if a particular region in the genome is transcribed (Morin et al., 2008). The RNA-seq technology starts with total RNA isolation from the cell sample (Fig. 1.12).

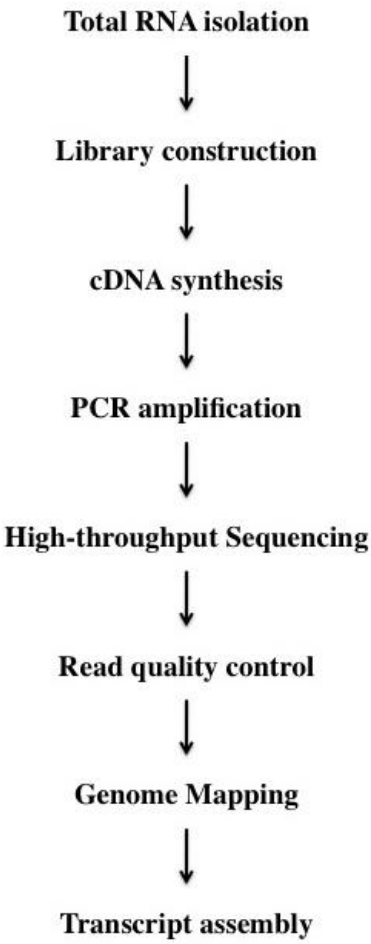

Figure 1.12 - Workflow of RNA-seq technology and analysis.

If the purpose is to analyse coding RNA, poly(T) oligos in magnetic beads are used to target the 3' polyadenylated tail of mRNA (Morin et al., 2008; Mortazavi et al., 2008). This is also the method to isolate PolyA- RNA (non coding RNA) by retaining the flow-through after capturing the beads (Morin et al., 2008). Specific RNA types can be selected from the PolyA- RNA sample through size-selection in a size exclusion gel (Morin et al., 2008). The next step is the synthesis of a cDNA library from the captured RNA. The cDNA fragments from the mRNA sample are cut to smaller fragments before PCR amplification. RNA is sequenced usually in a paired-end fashion. The sequencing reads are

subjected to quality control and filtered before being mapped to the reference genome. Mapped reads outline putative exons and help estimate transcript abundance. The next step is transcript assembly from the aligned reads.

1.7.3 Chromosome conformation capture, 3C, 4C and 5C

Chromosome conformation capture (3C) is a methodology to define the presence and frequency of contacts between genomic loci (Dekker et al., 2002). The experimental technique starts with the isolation of intact nuclei. Formaldehyde is used to fix protein-DNA interactions (Fig. 1.13).

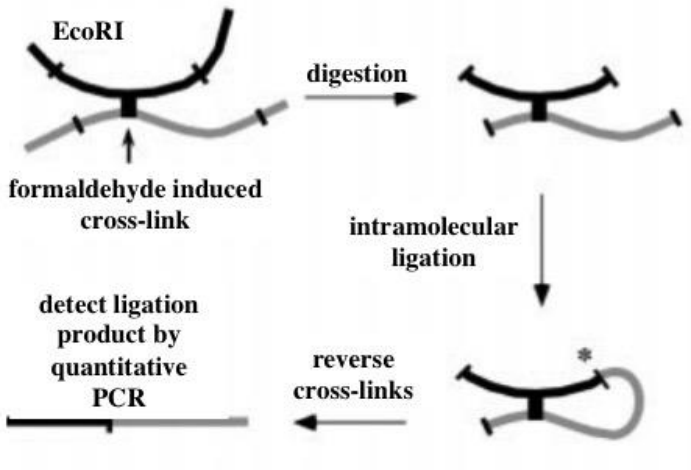

Figure 1.13 - Methodology for chromosome conformation capture (3C). After formaldehyde cross-linking, chromatin is digested with EcoRI followed by intramolecular re-ligation and quantitative PCR to detect interacting fragments after reversal of cross-links. From (Dekker et al., 2002). Reprinted with permission from AAAS.

After formaldehyde cross-linking, chromatin is digested with EcoRI which cuts the non-cross-linked DNA. DNA fragments are ligated into rings and

cros-linking is reversed. This creates hybrid fragments containing both interacting fragments. Detection of contacts is carried out by PCR with primers specific for the interactions of interest. However, 3C is hindered by its low resolution as the results depend on the abundance of PCR products from the targeted interaction compared to a control template. 3C also requires a choice of target loci and it is experimentally laborious to look for many interacting loci.

Variants of 3C include conformation capture on chip (4C), a technology to ascertain all contacts across the genome to a genomic site of interest (one-to-all) (Simonis et al., 2006). In 4C, a second restriction digest with a different cutter is employed and the resulting fragments are self-circularised. Inverse PCR with primers that map to the known region amplify the unknown interacter in the middle of the fragment. Quantification is performed with microarrays (Simonis et al., 2006). Another variant, 3C-carbon copy (5C), has lower resolution than 4C but generates interaction maps more accurately than 3C (Dostie and Dekker, 2007). 5C uses universal primers to amplify the ligation products (many-to-many). Interactions are detected through microarrays or DNA sequencing (Dostie and Dekker, 2007).

1.7.4 High-throughput conformation capture, Hi-C

Hi-C is a chromosome conformation capture technique followed by high-throughput sequencing that enables the unbiased detection of long-range interactions in a genome-wide fashion. This technique follows 3C quite closely (Fig. 1.14).

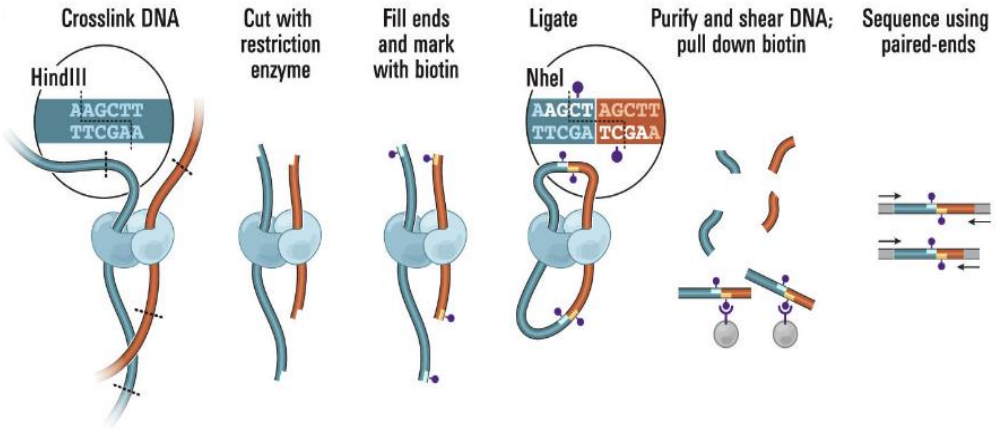

Figure 1.14 - The Hi-C method. Cells or nuclei are cross-linked with formaldehyde to preserve chromatin segments that are spatially adjacent. A ubiquitous restriction enzyme is used to cut the chromatin and the resulting sticky ends are filled with biotinylated nucleotides. Subsequent ligation in extremely dilute conditions creates new chimeric molecules. DNA is sheared and biotinylated segments are pull down with streptavidin beads. Fragments go through paired-end sequencing with the resulting pairs representing the two interacting/spatially adjacent chromatin fragments. From (Lieberman-Aiden et al., 2009). Reprinted with permission from AAAS.

Cells are cross-linked with formaldehyde and lysed. Chromatin is digested with an ubiquitous enzyme (4-6 cutter) that leaves a 5' overhang and the DNA ends are filled with biotinylated nucleotides. The blunt-end fragments are ligated in dilute conditions to favour intramolecular ligations and DNA is sheared. Cross-linking is reversed and streptavidin beads are used to pull down the biotinylated segments. The DNA hybrid fragments are subjected to paired-end sequencing. After quality control, reads are aligned independently to the genome and their mapping positions are used to create heatmap matrices of the intrachromosomal interactions (Fig. 1.15).

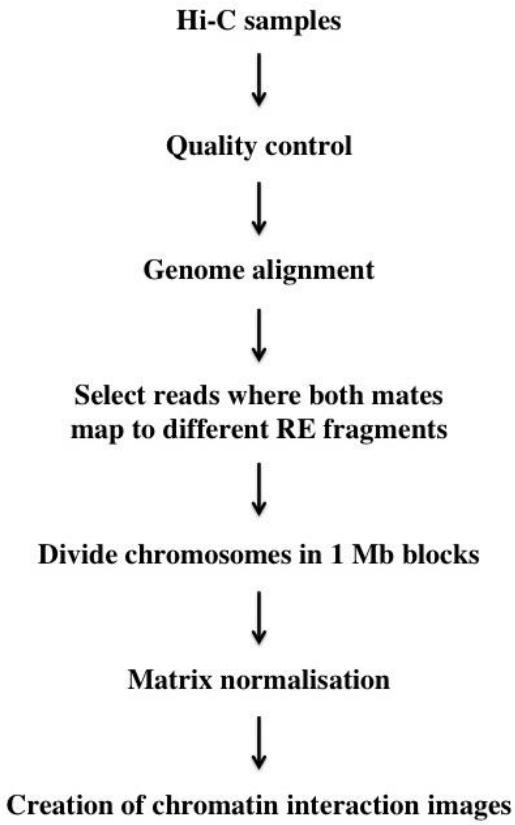

Figure 1.15 - Workflow of Hi-C reads analysis.

1.7.5 DNA combing technique

The observation of repeated regions with repeats longer than sequencing reads is difficult to attain with DNA sequencing (Hattori et al., 2000). DNA molecular combing was developed to study the structure of such regions. Microscopy glass cover slips with a monolayer of silane molecules are used to anchor the ends of single DNA molecules in solution. The cover slips are slowly lifted from the solution at a constant rate and this controlled movement enables full extension of the bound molecules across the surface. Different fluorescent

probes spanning the areas of interest are hybridised onto the stretched DNA molecules. Samples can then be observed under a microscope as any other FISH sample and placement of probes in relation to each other can be assessed in the single DNA strands (Bensimon et al., 1994). Although the technique is designed to capture and stretch single DNA molecules it is possible that a few DNA strands might end up stretched together in the cover slip biasing the hybridisation step.

1.8 Cell lines used in this project

The hTERT-immortalised retinal pigment epithelial cell line, hTERT RPE-1 is a karyotypically normal cell line, with a single derivative X chromosome that has additional material at the end of the q-arm. However, as the rDNA repeats are not localised in the X chromosome the derived chromosome is not a hindrance to the study of the nucleolus. The RPE-1 cell line has been immortalised with human telomerase reverse transcriptase, hTERT (Bodnar et al., 1998).

Contrary to RPE-1 cells, HeLa cells are not karyotypically normal. HeLa are an immortalised cervical cancer cell line that in addition to the human genome has also incorporated DNA from human papillomavirus 18 (HPV18). HeLa typically have more than the expected 46 chromosomes, with almost the double of that number and massive rearrangements derived from chromothripsis (Landry et al., 2013).

Recently, Pacific Biosciences released their shotgun sequencing dataset with ~54x coverage of the human genome. The data were created by sequencing a human cell line, CHM1htert, and is being used to generate an alternative human reference genome tailored to study structural variation. The CHM1 cell line is ideal for this purpose. CHM1 generated from a hydatidiform mole, this is an abnormal pregnancy where an egg with no nuclear DNA gets fertilised and duplication of the sperm DNA occurs. The full dataset was made publicly available on the 12th February 2014 and it contains over 22.5 million reads with an average read length of almost 8kb.

1.9 Aims of this thesis

The specific aims of my thesis were:

1) Establish the organisation of ribosomal gene repeats
2) Determine the spatial organisation of the sequences distal to rDNA genes
3) Extend and characterise sequences on the distal side of rDNA genes

2 Molecular Biology and Bioinformatics Methods

For this thesis, Perl, R and Python were the main scripting languages used. Several publicly available tools such as SAMtools (Li et al., 2009), BamTools (Barnett et al., 2011), fastQC (Patel and Jain, 2012) and Trimmomatic (Bolger et al., 2014) were employed. The hTERT-immortalised retinal pigment epithelial cell line, hTERT RPE-1, was the main cell line used experimentally in the project.

2.1 Tissue Culture

RPE-1 cells (hTERT-immortalised retinal pigment epithelial - hTERT-RPE-1) were grown in DMEM/Nutrient Mixture F-12 Ham media, supplemented with 17.3 mL Sodium bicarbonate (v/v)(Sigma), 50 mL 10% fetal bovine serum (v/v) (BioSera), 5 mL 1% L-Glutamine 200mM (v/v)(Sigma), and 5 U/mL (100µg/ml) of penicillin/streptomycin (Sigma) per 500 mL bottle. Cells were maintained in T175 culture flasks at 37ºC with 5% CO_2. For seeding new flasks when cells reached confluency, media was removed and cells were washed with PBS (phosphate buffered saline). Cells were detached from the flasks with 1x Trypsin (Sigma) with 1 mM EDTA pH 8.0, and incubated at 37ºC for 5 minutes. If necessary, the flask was gently tapped to help cells detach from the surfaces of the flask. Trypsin was neutralised with an equal volume of media. Single cell suspension was obtained by pipetting up and down. Cells were re-seeded at the

required dilution into new T175 or 150 mm dishes containing fresh media.

2.2 Isolation of nucleoli

RPE-1 cells were seeded on to 15 x 150 mm dishes containing media (described in section 2.1.) and grown until >80% confluence. Subsequent steps were performed on ice. Media was discarded and cells from all dishes were harvested by scraping into 40 mL of PBS (40 ml used per ~ 4-5 dishes). The resulting cell solution was divided by 50 mL falcon tubes and centrifuged at 1200 rpm at 4ºC for 5 minutes. Supernatant was discarded and 40 mL of PBS added to the first 50 mL falcon tube. Samples pooled to a single 50 mL tube and centrifuged again at 1200 rpm at 4ºC for 5 minutes. Supernatant was discarded. Cells were resuspended in 15 mL PBS and transferred into a 15 mL tube. The cell suspension was centrifuged for a third time at 1200 rpm at 4ºC for 5 minutes. Liquid was discarded and 5 mL of Buffer A was added (10 mM Hepes, pH 7.9, 10 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM DTT - 5 µL DTT was added just before the start of experiment). After 5 min, a small drop of the cell suspension (~ 5 µL) was put on a glass slide and checked under a phase contrast microscope. Cells should be swollen but not burst. Solution was transferred to a Dounce tissue homogeniser and 1 mL Buffer A was added. Suspension was homogenised 10 times (repeated as necessary) and checked under a microscope after each repetition. Homogenisation was performed until >90% of cells were burst leaving intact nuclei. Homogenised cells were transferred to a new 15 mL tube and Buffer A was used to clean and transfer cells left in the homogeniser. This

new tube was centrifuged at 1200 rpm at 4ºC for 5 minutes. Pellet was resuspended in 3 mL S1 solution (0.25 M Sucrose, 10 mM $MgCl_2$ + half tablet Protease inhibitor, complete mini EDTA-free) and poured into a new 15 mL tube already containing a 3 mL S2 solution (0.35 M Sucrose, 0.5 mM $MgCl_2$ + one tablet Protease inhibitor, complete mini EDTA-free). The two-layered solutions should be cleanly separated. This was followed by centrifugation at 1430xg for 5 min at 4ºC. Supernatant was discarded and pellet (nuclei) resuspended in 4.5 mL S2 and transferred to a new 15 mL. The nuclear pellet was sonicated on ice for 5 x 10 second bursts (with 10 second interval) and repeated if necessary. Sonicated nuclei were checked under a microscope after each round of sonication, looking for no intact cells and nucleoli observed as dense and refractile bodies. Over-sonication of nucleoli leads to their disruption. Sonicated sample was layered over 3 mL of S3 solution (0.88 M Sucrose, 0.5 mM $MgCl_2$ + one tablet Protease inhibitor, complete mini EDTA-free) and centrifuged at 3000xg for 10 min at 4ºC. Supernatant was discarded and nucleolar pellet resuspended in 0.5 mL of S2 solution and further sonicated at 1430xg for 5 min at 4ºC. Resulting pellet contained highly purified nucleoli (checked under microscope) and stored at -80ºC in 0.5 mL S2 solution as necessary.

2.3 DNA extraction from purified nucleoli

DNA extraction from purified nucleoli was carried out by washing cells with PBS and incubating overnight in 400 µL TE (22 mM tris pH 8.0, 2 mM EDTA), 0.5% SDS and 0.3 mg/mL proteinase K (Roche). Tubes were

centrifuged at 1000 rpm for 1 min and the liquid was discarded. Afterwards, 10 µL proteinase K were added and samples were incubated for 2 hours at 50ºC, turning the tubes every half hour. Added 40 µL 3 M sodium acetate. Added an equal volume (tube volume) of phenol-chloroform to each tube. Samples were centrifuged for 10 min at maximum speed. The lower half of the tube was discarded and the upper half was transferred to new centrifuge tubes. 1 mL of 100% ethanol was added (usually 2-2.5 times the volume) and tubes were placed at -80ºC for 20 min. Tubes were centrifuged for 10 min at maximum speed. Liquid was discarded and pellet was centrifuged again to remove any ethanol left. 50-100 µL TE were added and samples were put at 37ºC for ~4 hours tapping the tubes to mix every half hour. Samples were kept in the fridge overnight. RNA was removed by discarding the liquid and pellets were resuspended in 100-200 µL TE. 2-4 µL RNase (25mg/mL) were added and tubes were incubated at 27ºC for an hour. DNA precipitation was carried out with 1/10 volume of sodium acetate (3M, pH 5.2) was added followed by 2.5-3 volumes of >95% ethanol. Sample was incubated for 30 min at -20°C. This was followed by centrifugation at 14,000xg for 5 min at 4°C. Supernatant was discarded and the pellet was rinsed in 70% ethanol and centrifuged again at 14,000xg for 15 min at 4°C. The pellet was air-dried for 3-5min to ensure removal of residual ethanol. Pellet was resuspended in 25 µL TE 1:0.1.

2.4 Measurement of nucleic acids concentration and purity

Concentration of extracted purified DNA was measured using the Picodrop spectrophotometer (Picodrop Limited). High concentration DNA was diluted in TE. Pure DNA has an A260/A280 ratio around 2.0.

2.5 Gel electrophoresis

DNA fragments were typically run on a 1% - 1.5% (w/v) agarose (as required by product length) dissolved in 1xTAE (40mM Tris, 20mM Acetic acid, glacial, 1mM EDTA) with 0.5µg/ml ethidium bromide (EtBr) to help visualisation of DNA. A 10x DNA loading dye (40% (w/v) sucrose, 0.25% (w/v) Xylene cyanol (XC)) was added to the DNA samples prior to loading. DNA Hyperladder (Bioline) with the appropriate sizes was also loaded to the gel. Gels were run usually at 100v.

Smaller DNA fragments (<200 bp) were loaded on a 1.5% (w/v) Agarose in 1xTBE buffer (pH 8.5) (80mM Tris, 80mM Boric acid, 2mM EDTA pH8.0) with 0.5µg/ml EtBr. DNA samples were loaded with 10x TBE loading buffer (10mM Tris pH8, 10mM EDTA pH8, 50% (w/v) sucrose, 0.15% (w/v) Bromophenol Blue).

DNA products were visualised on a UV transilluminator (Gbox imager Syngene) and images were captured using GeneSnap (Syngene).

2.6 Fluorescence in Situ Hybridisation, FISH

1μg of BAC DNA or nucleolar DNA was labelled with Green dUTP or Red dUTP (rDNA and DJ probes were also labelled accordingly as necessary) using the Nick Translation kit (Abbott Molecular) according to the manufacturer's protocol.

A probe comprised of 100 ng BAC/genomic DNA (5 μL of labeling reaction) and 50 ng rDNA/DJ (2.5 μL of labeling reaction), 2.5 μL Human Cot-1 DNA (1 mg/mL), 5 μL Herring Sperm DNA (10 mg/mL) and 1/10 volume of sodium acetate (3 M, pH 5.5). DNA was precipitated with 2.5 volumes of 100% ethanol, washed with 70% ethanol and air-dried. The pellet was resuspended in 25 μL Hybrisol® VII (Qbiogene). Metaphase were denatured for 5 min in a 75°C water bath with ~40 mL M-FISH Denaturation buffer. The metaphase slide was then dehydrated for 2 min in

70% ethanol followed by 2 min in 90% ethanol and 2 min in 100% ethanol and air-dried. The DNA probe was denatured at 75°C for 5 min and applied to a previously warmed (37ºC) coverslip. The slide was lowered to the cover slip and glued together with rubber cement (Marabu-Fixogum). Slides were incubated in at 37°C in a humidity chamber overnight. The slide was then washed in 0.4% SSC/0.3% NP-40 at 74°C for 2 min and in 2xSSC/0.1% NP-40 for 5 min at room temperature. The slide was air-dried and mounted in VectorShield® with DAPI (Vector Laboratories).

2.7 Mosaik alignment of Roche 454 reads

Mosaik Aligner (version 2.2.3) was used to map all 454 reads (single-end and paired-end) to the rDNA repeat extracted from AL592188. Default parameters for alignment were used, and the options ”-mmp 0.10 -bw 51” were set to align reads with a mismatch threshold of 10% of the total length of the read and align Roche 454 reads as suggested by the Mosaik manual for increased performance. Unlike other aligners, Mosaik requires the user to create numerous commands to perform every step of the alignment. The first and second commands convert the reference genome and input sequences to the input format accepted by Mosaik. The subsequent commands perform the alignment, sort the alignments and convert to SAM format. The last command calculates the average coverage for all alignments.

2.8 Quality control of sequencing reads

To trim low quality bases from the PacBio reads the fastq_quality_trimmer tool (Giardine et al., 2005) was used. For Illumina reads from ChIP-seq and Hi-C studies, Trimmomatic (version 0.32) was used for quality control and filtering of reads using the appropriate sequence adaptors (Bolger et al., 2014).

2.9 Alignment of PacBio reads

BLASR aligner (version 1.1) was used to align PacBio reads to the rDNA repeat extracted from AL592188. The options "-minPctIdentity 80 -minMatch 10 -minReadLength 100 -header -maxExpand 4" were used to establish the minimum percentage of identity between the read and the reference, set the minimum seed length, discard reads smaller than 100 bp and define the number of search iterations, respectively.

2.10 Generation of a consensus sequence

The sorted and indexed BAM files created from the SAM alignment files were used in conjunction with SAMtools, bcftools and vcfutils.pl to produce reference-guided consensus sequences. The SAMtools options "-C 50" was used when there was high coverage of low precision reads. The bcftools options "-c –e -g" were used to call variants, enforce maximum-likelihood inference and call the genotypes at variant sites.

2.11 Generation of sub-sequences from PacBio reads

Rearrangements in the rDNA repeats have been reported to have a palindromic nature. Rearrangements will be observed in PacBio reads that

contain non canonical sequences of the 18S and 28S regions. Information on the orientation and order of paired-end reads can be used in alignments to infer structural variation in the reference sequence. PacBio sequences are long single-end sequencing reads. The average read length for the RPE1 and HeLa samples is around 3kb. Reads longer than at least 2kb were selected and artificial paired-end reads created by extracting 500 bp from each end of the sequences in a 100 bp sliding window. The same method was used in the CHM1 PacBio reads. Sequences that were longer than 2kb were selected and the paired-end generated by extracting 500 bp from each side.

2.12 Culture and storage of BAC/plasmid clones

Agar plates were prepared for the BAC clones in DH10B *e. coli* stab-cultures. Single colonies were picked and grown in 10 mL Lysogeny broth (LB) supplemented with the appropriate antibiotic (12.5 µg/mL Chloramphenicol or 50 µg/mL Ampicillin as suggested by vendor) and placed in a 37ºC incubator.

A glycerol stock was prepared for long-term storage at -80ºC. An 800 µL aliquot from the cell culture was added to 200 µL of pre-warmed glycerol (100%) in a screw cap tube to achieve a final concentration of 20% glycerol. The tube was vortexed and stored at -80ºC.

2.13 Plasmid purification from small cultures

Cultures in 10 ml LB broth with the appropriate antibiotic were grown overnight at 37°C. Cultures were centrifuged at 4000xg for 15 minutes at 4°C. Plasmid DNA was isolated with the NucleoSpin® Plasmid kit (Macherey-Nagel). The DNA was eluted in 50-100µl TE (10 mM Tris pH8.0, 0.1 mM EDTA). Alternatively colonies were streaked out on a large area of the LB agar plate and grown overnight at 37°C. The bacteria were scraped of the LB agar plate with an inoculation loop and directly dissolved in the first buffer from the kit.

2.14 Plasmid purification from large cultures

The BAC/plasmid culture from the previous sections was later transferred to a large flask with 800 mL LB medium and the appropriate antibiotic and kept in a shaking incubator at 37ºC overnight. Afterwards, cells were centrifuged at 4000xg for 15 min at 4ºC. Low copy BAC DNA and high copy Plamid DNA was extracted using the NucleoBound Xtra Maxi Plus kit (Machery-Nagel Cat No 740414.50) as per manufacturer's instructions. DNA was precipitated with equal volume of Isopropanol followed by centrifugation at 4000x g for 15 min at 4°C. Pellets were washed in 70% ethanol and resuspended in 100 µL TE (10mM Tris pH8.0, 0.1mM EDTA). Plasmids/BAC yield was determined by UV spectrophotometry. Plasmid digestion was performed with the appropriate restriction enzyme and the integrity of the plasmid confirmed by gel

electrophoresis.

2.15 Analysis of transcriptome profile

Alignment of RNA-seq data to the reference genome was performed with TopHat (version 1.4.1) using the options "-i 30 --segment-mismatch 2 --segment-length 38" to define the minimum intron length, the minimum length for cutting a read into segments, and the number of mismatches per independently matched segment.

After alignment of ChIP-seq and RNA-seq sequencing reads to the target reference, duplicated reads (originating from PCR amplification artefacts) were removed with Picard tools.

Assembly of aligned RNA-seq data into transcripts was carried out with Cufflinks (version 1.3.0) using the option "--multi-read-correct" to employ a 'rescue method' for multi-reads. Multi-reads are sequencing reads that align to multiple locations in the genome. Cuffmerge was used to merge cufflink assemblies (originating from fastq files from the same sample).

2.16 BLAST search

The search for human sequences in the nucleotide database to extend the distal side of NORs was carried out with BLAST with options "-dust yes -penalty -1 -gapopen 0 -gapextend 2 -reward 1 -perc_identity 80". These options were chosen to filter repeats and low information content sequences, set the penalty for nucleotide mismatches, opening or extend gap, set the score for matching nucleotides and set the percentage identity threshold between reads and reference.

2.17 PCR/RT-PCR

Polymerase chain reaction, PCR, was performed using DNA polymerases Taq or Q5, depending on the size of expected product. Primer pair mixes were created by adding 10 µL of forward and reverse primer in 80 µL TE. Master reaction contained 1x Taq buffer (10 mM Tris pH 9.0 (25°C), 50 mM KCl, 1.5 mM $MgCl_2$, 0.1% (v/v) Triton X-100)) or 5x Q5 buffer (NEB) with 200 µM dNTPs (Bioline) and 0.2 µM of each primer. To the reaction, 50-100 ng genomic DNA (gDNA), nucleolar DNA or 10-30 ng plasmid was added. GC-rich PCR products were amplified in the presence of 1 M Betaine (Sigma Aldrich). Reactions were carried out in an Eppendorf Mastercycler gradient (Eppendorf). A typical program for Taq reaction was 5 min at 95°C, 25-35 cycles of 30 sec at 95°C, 30 sec at 50-65°C, 72°C for 1kb/1min. For Q5®, 30 sec at 98ºC followed

by 35 cycles of 10 sec at 98ºC, 20 sec at 50-70ºC, 4 min 72ºC for 20-30 sec per kb and final extension for 2 min at 72ºC.

Reverse transcriptase PCR, RT-PCR, was performed in two steps. In the first step, the ProtoScript M-MulV First Strand cDNA Synthesis Kit (New England Biolabs), was used to synthesize cDNA from an RNA sample as per manufacturer's instructions. The second step, PCR was carried out as described above.

2.18 Purification of PCR products

DNA from PCR runs was isolated with the NucleoSpin Extract II (Macherey-Nagel) as per manufacturer's instructions. Alternatively, 1/10 volume of sodium acetate (3M, pH 5.2) was added followed by 2.5-3 volumes of 100% ethanol to precipitate the DNA. Sample was incubated for 30 min at -20°C. This was followed by centrifugation at 14,000xg for 5 min at 4°C. Supernatant was discarded and the pellet was washed in 70% ethanol and centrifuged again at 14,000xg for 15 min at 4°C. The pellet was air-dried for 3-5min to ensure removal of residual ethanol. Pellet was resuspended in TE 1:0.1 and 25 µL of the appropriate buffer.

2.19 cDNA cloning and sequencing

cDNA from RT-PCR was amplified using CloneJET PCR Cloning Kit (Life Technologies) as per manufacturer's instructions. DNA sequencing of cloned cDNA was performed by Source Biosciences (LifeScience).

2.20 Bowtie alignment of Illumina reads

Bowtie aligner (version 1.0.0) was used to map Hi-C or ChIP-seq Illumina reads to custom genomes, comprised of CRCh37+rDNA+PJ+DJ, CRCh35+rDNA+PJ+DJ or CRCh37+rDNA+PJ+DJ+AL591856. The options "--best -m 1" were used to enforce unique mapping and report alignment by stratum (number of mismatches).

If necessary, SRA formatted files were converted to FASTQ using the SRA toolkit. Picard tools were used to remove duplicates from the ChIP-seq data sets.

3 Rearranged rDNA repeats

3.1 Background

Nucleoli, the sites of ribosome biogenesis and key regulators of cellular growth and proliferation (Grob et al., 2014; Pederson, 2011), form around nucleolar organiser regions, NORs, positioned in the short arms of the five human acrocentric chromosomes (Fig 3.1)(Henderson et al., 1972).

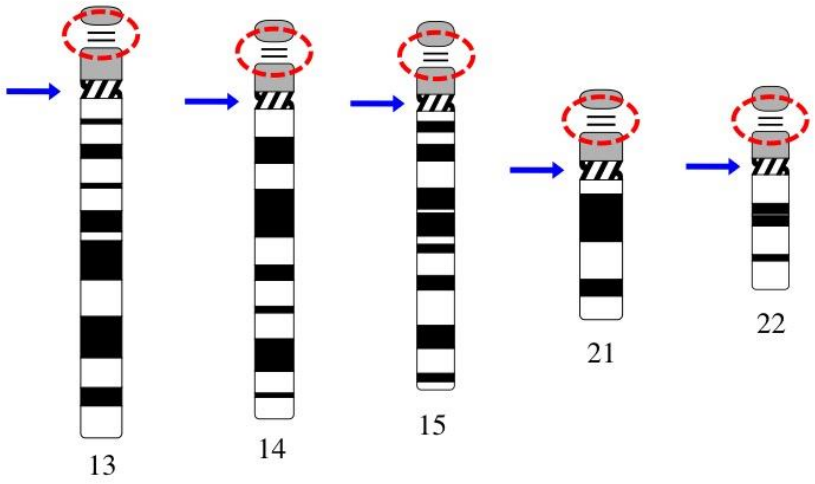

Figure 3.1 - The five human acrocentric chromosomes, 13, 14, 15, 21 and 22. The acrocentric chromosomes have their centromeres close to one end of the chromosome. This results in an asymmetric conformation where the short arm is much smaller than the long arm. The red circles indicate the location of nucleolar organiser regions and the blue arrows the location of centromeres. Chromosome figures from Idiogram Album: Human copyright © 1994 David Adler.

In humans, NORS contain 1Mb to 3 Mb ribosomal gene (rDNA) arrays, organised in a head-to-tail orientation (Fig 3.2) and the adjacent sequences on

either side of the tandem repeats, proximal junction toward the centromere and distal junction toward the telomere (Floutsakou et al., 2013; McClintock, 1934).

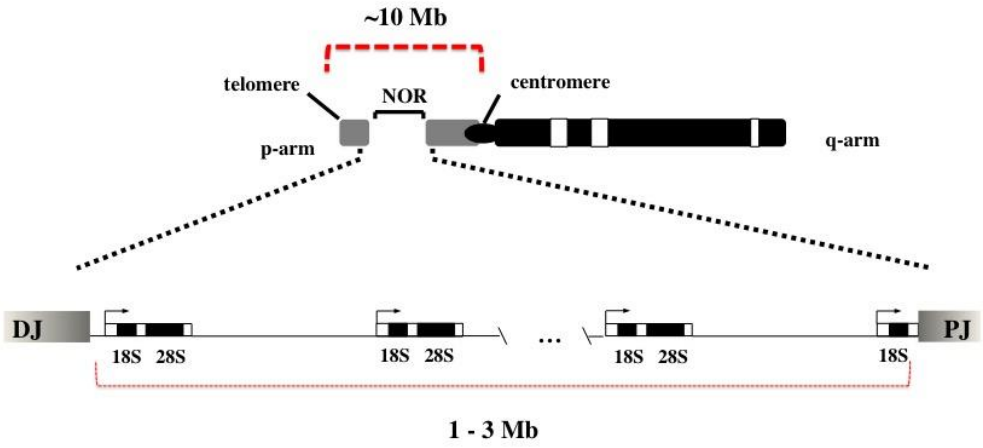

Figure 3.2 - Nucleolar organiser regions, NORs, are located in the short arms of the acrocentric chromosomes. NORs are comprised of tandem arrays of ribosomal genes organised in a head-to-tail orientation towards the centromere and the distal junction (DJ) and proximal junction (PJ) on either side of the block of repeats. Transcription, by RNA polymerase I, proceeds from the DJ towards the PJ.

Each individual repeat of rDNA is almost 44 kb in length and contains the coding sequences for the 18S, 5.8S and the 28S ribosomal subunits (Fig 3.3) located in the first 13kb of the repeat. As the human genome is estimated to be 3.2 billion bases long, the ribosomal genes represent less than 0.05% of the genome.

5'ETS	1 - 3654
18S	3655 - 5523
ITS1	5524 - 6600
5.8S	6601 - 6765
ITS2	6766 - 7924
28S	7925 - 12994
3'ETS	12995 - 13392
IGS	12393 - 43972

Figure 3.3 - Human rDNA repeat extracted from BAC AL592188. The repeat is 43972 bp and contains the sequences for the 18S, 5.8S and 28S ribosomal subunits separated by internal transcribed spacers followed by a large intergenic spacer.

The most commonly used consensus for the rDNA is a 43 kb sequence created by sequencing four fragments generated by EcoRI digestion; genbank accession number: U13369.1 (Gonzalez and Sylvester, 1995; Gonzalez et al., 1992). However, this consensus is inaccurate and contains many sequencing errors, including an inexact gene promoter and differences to the repeat extracted from AL592188 (clone RP11-337M7). U13369.1 is a ~43 kb sequence comprised of the ribosomal gene sequence only. AL592188 is a much longer BAC (~162 kb). It contains a slightly longer rDNA gene sequence (~44 kb) and part of the distal junction, meaning it is a representation of the last rDNA repeat in a NOR on the telomere side. Alignment of high-throughput and single molecule real time sequencing reads to both consensuses revealed AL592188 has better coverage and fewer mismatches than U13369.1 and therefore was considered a better representation of the ribosomal gene.

The entire short arms of the five human acrocentric chromosomes are missing from the current human reference genome assembly (GRCh38). The

heterochromatic and repetitive nature of these regions makes it very difficult to sequence them with sufficient coverage and authenticity. One way to visualise the distribution of rDNA in the genome is to perform FISH on individual DNA fibres, harvested by the molecular combing technique. The transcribed regions are examined with two specific probes that cover the entire length of the coding regions (Caburet et al., 2005). This data revealed the rDNA genes are organised in repeating arrays in the same orientation. Importantly, molecular combing not only confirms the standard rDNA tandem head-to-tail arrays but also reveals unorthodox patterns (Fig 3.4).

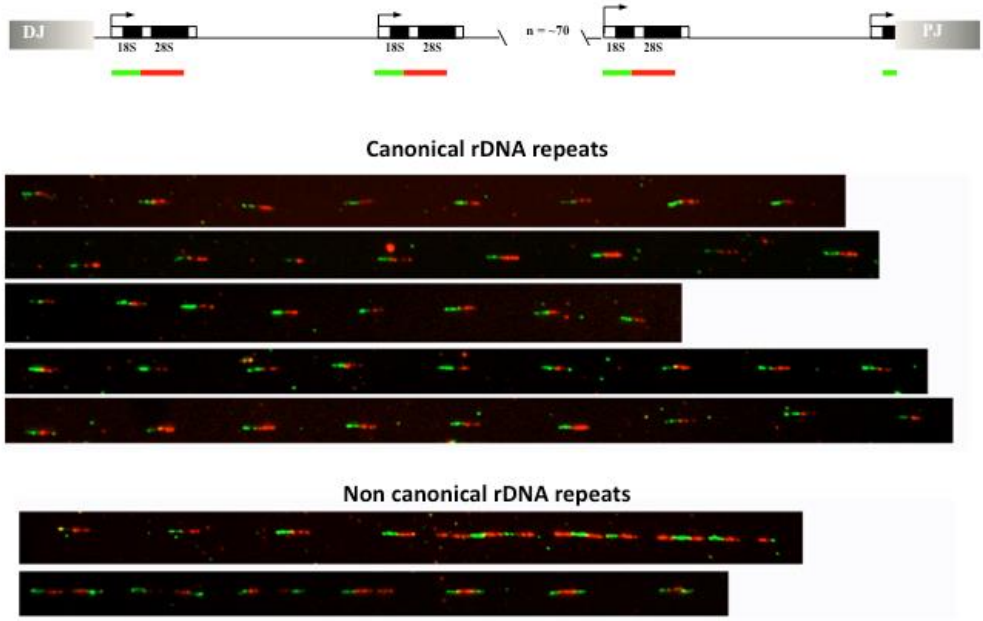

Figure 3.4 - Combing of rDNA reveals canonical organisation for 18S (green) and 28S (red) regions in 1Mb DNA fibres. However, around 30% of probed strands depict rearrangements in the rRNA coding regions. Images generated by Prof Brian McStay (HeLa DNA fibres).

These noncanonical units depict rDNA rearranged into palindromic segments that do not follow the expected configuration. These rearrangements could carry consequences on the nucleolus, e.g. causing genomic instability of

the NOR, produce defective rRNA and/or introduce convergent transcription. All these events can disrupt nucleolar biogenesis and induce nucleolar and nuclear stress. Molecular combing data from Caburet and colleagues (Caburet et al., 2005) and also from the McStay lab are, presently, the only available evidence of the occurrence of rearrangements in the rDNA repeats. Another possibility is that these structures are experimental artefacts, such as several DNA fibres adhered together. Therefore, it is important to gather independent evidence to confirm the existence of rearrangements and subsequently explore their role in cellular activity.

A different approach is to directly sequence these regions of the genome and characterise them at the nucleotide level. Next generation sequencing technology such as 454 offers a higher yield than Sanger sequencing at a more effective cost and timeframe with longer reads than Illumina. Pacific Biosciences (PacBio) reads on the other hand offer longer read length with lower accuracy than 454 reads. The technology employed in 454 chemistry produces high precision sequencing reads. Roche 454 has high accuracy due to the fact that thousands of identical molecules (from PCR amplification) are sequenced and averaged into consensus reads.

Pacific Biosciences developed DNA single molecule real time sequencing by synthesis, SMRT (Levene et al., 2003). In this technology, a single DNA polymerase enzyme is attached to the bottom of a zero-mode waveguide cell, ZMW, with a single molecule of DNA as template. Using a single DNA molecule as template has the drawback of lower accuracy as each fragment is only sequenced once. On the other hand the longer length of PacBio reads can help in the resolution of complex repeats.

SMRT technology is interesting to this project as it uses single-purified non-amplified DNA and have longer sequences than 454, at the cost however of accuracy. However, contrary to next generation sequencing, the precision errors are random and more probable to be mismatches than insertions and deletions. As such, with sufficient coverage the low accuracy of PacBio reads might be overcome.

Mosaik aligner is indicated for next generation technologies such as Roche 454 and it allows gapped alignments. In addition, Mosaik can also produce reference-guided assemblies of the gapped pairwise alignments.

Aligners currently used for mapping next generation sequencing are designed for short high accuracy reads. BLASR, a mapping tool for single molecule sequencing reads using local alignment with successive refinement (Chaisson and Tesler, 2012) was develop by Pacific Biosciences to tackle the challenge of aligning longer reads with higher error rate.

In this chapter I describe purification of nucleoli from RPE-1 and HeLa cells and extract DNA from the isolated nucleoli. I also describe how to prepare DNA for 454 and PacBio sequencing and describe the analysis of RPE-1 454 nucleolar paired-end sequencing data to look for evidence of rearranged rDNA repeats. The RPE-1 and HeLa PacBio sequencing and the publicly available genomic PacBio (from the Platinum Genome human alternate assembly) data with high coverage of the human genome will also mined to search for rearrangements. A new consensus for rDNA repeat was generated from mapping of shotgun and paired-end 454 nucleolar reads, nucleolar PacBio and genomic PacBio.

3.2 Results

3.2.1 DNA preparation for 454 sequencing

To address the existence of rearranged rDNA repeats, it was necessary to prepare DNA for sequencing. In a single cell, although there are around 400 copies of rDNA these only represent a very small portion of the human genome. Therefore, there was a need to enrich for acrocentric short arms in the sequencing data as much as possible. The strategy employed to ensure enrichment of rDNA sequences was to extract DNA from nucleoli, which form around active rDNA repeats.

The physical characteristics of the nucleolus can be explored to purify nucleoli from nuclei, as they are the densest of nuclear sub-domains and resistant to sonication. The cell line hTERT-RPE1, a female immortalised retinal pigmented epithelial cell, was chosen for sequencing as it is mainly karyotypically normal, with a single derivative X chromosome. The RPE-1 cell line has been immortalised with human telomerase reverse transcriptase, hTERT. However, as the rDNA repeats are not localised in the X chromosome the derived chromosome is not a hindrance to this project.

RPE-1 cells were grown until 80-90% confluence. This yielded around $1.5x10^7$ cells per dish. For details regarding tissue culture method employed for cell maintenance please see section 2.1.

Cells were washed, nuclei were prepared and nucleoli were isolated through sonication. Nucleoli were separated from the "nucleoplasmic fraction" by sedimentation through sucrose cushion. Every step of the nucleoli isolation

procedure was monitored using a phase contrast microscope. For detailed information on this procedure, please refer to section 2.2. – Isolation of Nucleoli. Harvested nucleoli were stored as specified in the methods section.

Previous studies used different cell lines when isolating nucleoli. An incipient study on genomics of the nucleolus used HeLa cells (Nemeth et al., 2010) to identify nucleolar-associated domains (NADs). In this study, cells were also grown to a high confluence but were fixed with formaldehyde prior to harvesting. The addition of formaldehyde forces NADs to remain associated with nucleoli during the isolation procedure. As many of these domains belong to other chromosomes, adding formaldehyde exacerbates the presence of other chromosomal regions in the nucleolar data. This was already a concern when sequencing nucleolar DNA due to the presence of segmental duplications in the short arms of acrocentric chromosomes (Floutsakou et al., 2013). Excluding this step from the nucleoli purification protocol facilitates the enrichment for acrocentric short arm sequences.

A similar study confirmed that around 4% of the genome associates with nucleoli without fixing cell before nucleoli isolation (van Koningsbruggen et al., 2010). In the same study, high coverage Illumina sequencing, with read lengths up to 50 bp, is used to reveal that the majority of chromosomes have nucleolar-associated regions. However, the cell line used, HT1080, is a fibrosarcoma cell line that contains abnormally enlarged nucleoli in addition to a deviant karyotype. Potentially this might influence chromatin organisation in the nucleus resulting in different association of chromosomal regions with nucleoli or more importantly, the organisation of rDNA repeats within the arms. Many cancer cells show an increase in rRNA synthesis in order to produce enough proteins to

shorten cell cycle time (Derenzini et al., 1998). These highly transformed cancer cells usually display abnormally large nucleoli and due to their atypical karyotype the DNA content from their acrocentric short arms could have significant changes relative to normal cells. To reiterate, we are interested in the organisation of rDNA repeats in normal cells.

Thus, in order to avoid unnecessary chromosomal regions being sequenced in our nucleolar sample, a decision was made to sequence karyotypically normal cells, using initially 454 which gives longer reads than those offered by Illumina and to enrich for nucleolar regions as much as possible. Also, longer sequencing reads offer a more accurate alignment of the DNA fragments.

DNA extraction was then carried out on the purified nucleoli. Nucleoli were resuspended in TE, SDS proteinase K to denature and digest proteins. To remove the proteins, Sodium acetate was added followed by a phenol-chloroform extraction. Nucleolar DNA was washed and precipitated with ethanol and resuspended in TE (for detailed method see section 2.3).

Concentration of extracted DNA was measured and gel electrophoresis was performed to ensure good fragment length in the extracted DNA (see sections 2.4-2.5 for details). The sample contained DNA fragment length above 10kb (Fig. 3.5) and was therefore suitable for 454 sequencing.

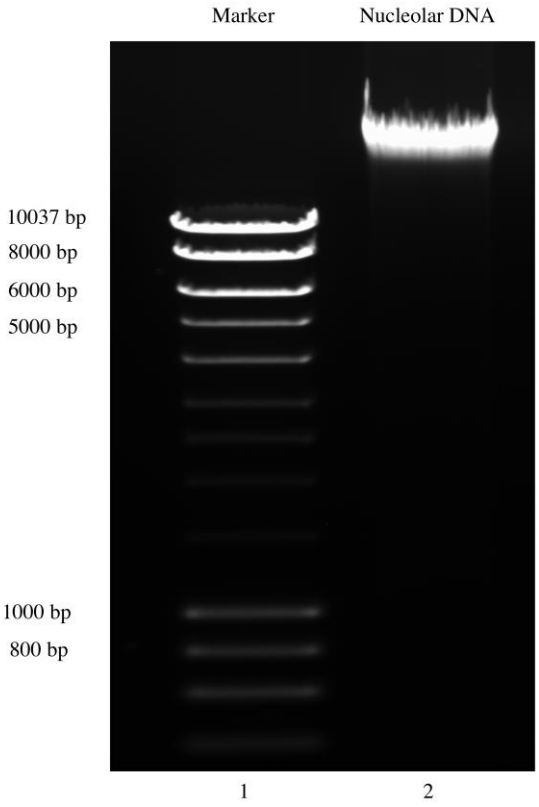

Figure 3.5 - Gel electrophoresis of purified nucleolar DNA. Hyperladder 1kb was used as marker in lane 1. Nucleolar sample in lane 2 shows a clear, thick band with very high molecular weight.

Enrichment of acrocentric short arms was confirmed by fluorescent in situ hybridisation, FISH, prior to sequencing (see section 2.6). All acrocentric chromosomes were represented in the sample (Fig 3.6). Limited hybridisations to other chromosomes regions were observed. These were, probably, due to segmental duplications from the sequences adjacent to NORs (Floutsakou et al., 2013).

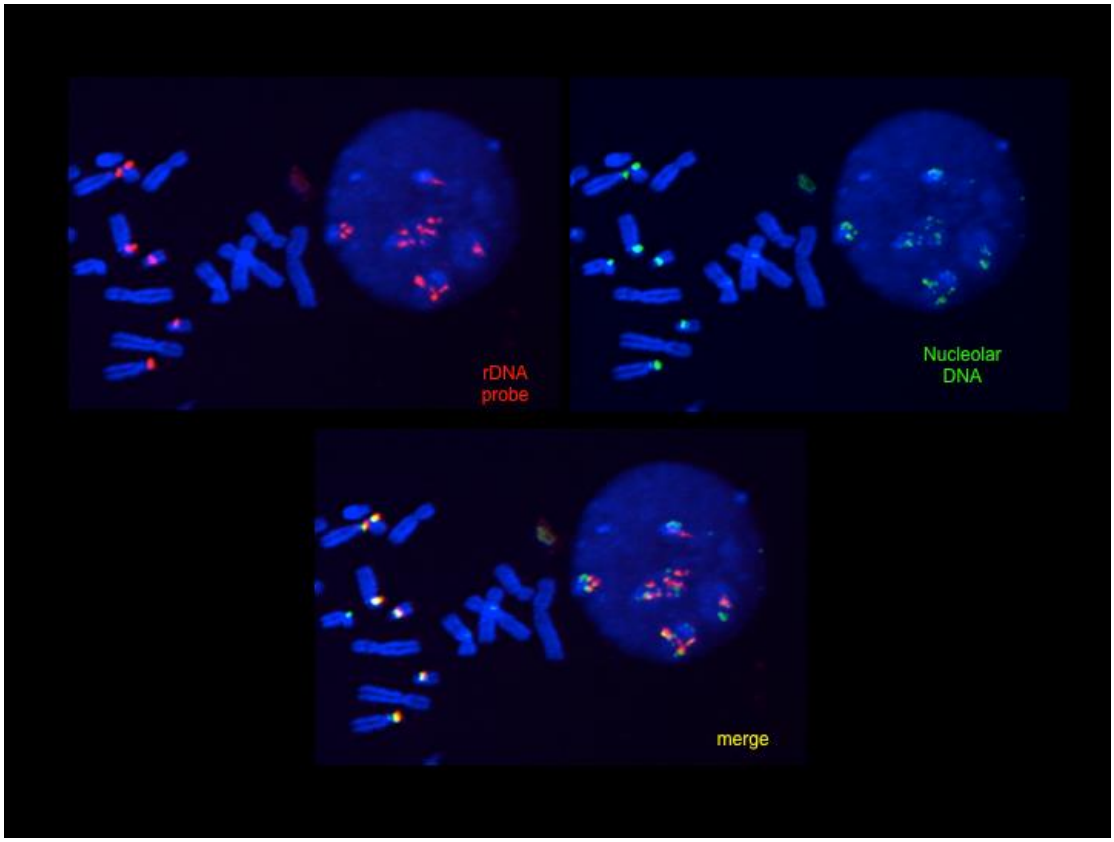

Figure 3.6 - FISH of nucleolar DNA. FISH confirmed enrichment for all short arms of acrocentric chromosomes with rDNA probe in red and nucleolar DNA in green.

Nucleolar DNA was sent for 454 sequencing in the 454 laboratories/Roche. Two sequencing libraries were generated, a single-end (shotgun) library and a paired-end sequencing library.

3.2.2 Roche 454 sequences and quality control

Two sequencing libraries, with single-end and paired-end sequences were received from Roche along with summary report (Table 3.1 and 3.2). Only paired-end reads that tested positive for the linker (Table 3.2), and were therefore comprised of left and right mate, were included in the analysis. Overall quality of the raw sequence data from the two libraries was assessed with FastQC.

Table 3.1 - Summary of statistics for shotgun (single-end) and paired-end 454 libraries

SID	Sample	HQ reads	HQ bases	Avg Read Length
17426	RPE-1	665,998	233,740,085	351
17463	RPE-1 paired-end	563,816	199,000,909	353

Table 3.2 - Summary of statistics of the paired-end reads

SID	# Reads	%Reads	Avg left length	Avg right length
17363	423,876	75.18%	165	166

The FastQC report (Fig 3.7) shows that the quality of the bases towards the end of the shotgun reads decreased considerably. The majority of sequences have around 500bp in length, with very few reaching over 900bp. Also a considerable number of reads has length of 70-80 bp. The report also showed GC content of 41% and no overrepresentation of sequences.

Figure 3.7 - Quality scores of 454 shotgun sequencing data across all bases and sequence length distribution.

For the paired-end file, the overall quality of the reads tended to decrease towards the end of the sequences with optimal average quality dropping after 350

bp (Fig 3.8). On the other hand, the majority of reads have length of 300 bp, indicating overall good quality of the paired-end reads. Over 152,000 reads are shorter than 50 bp though.

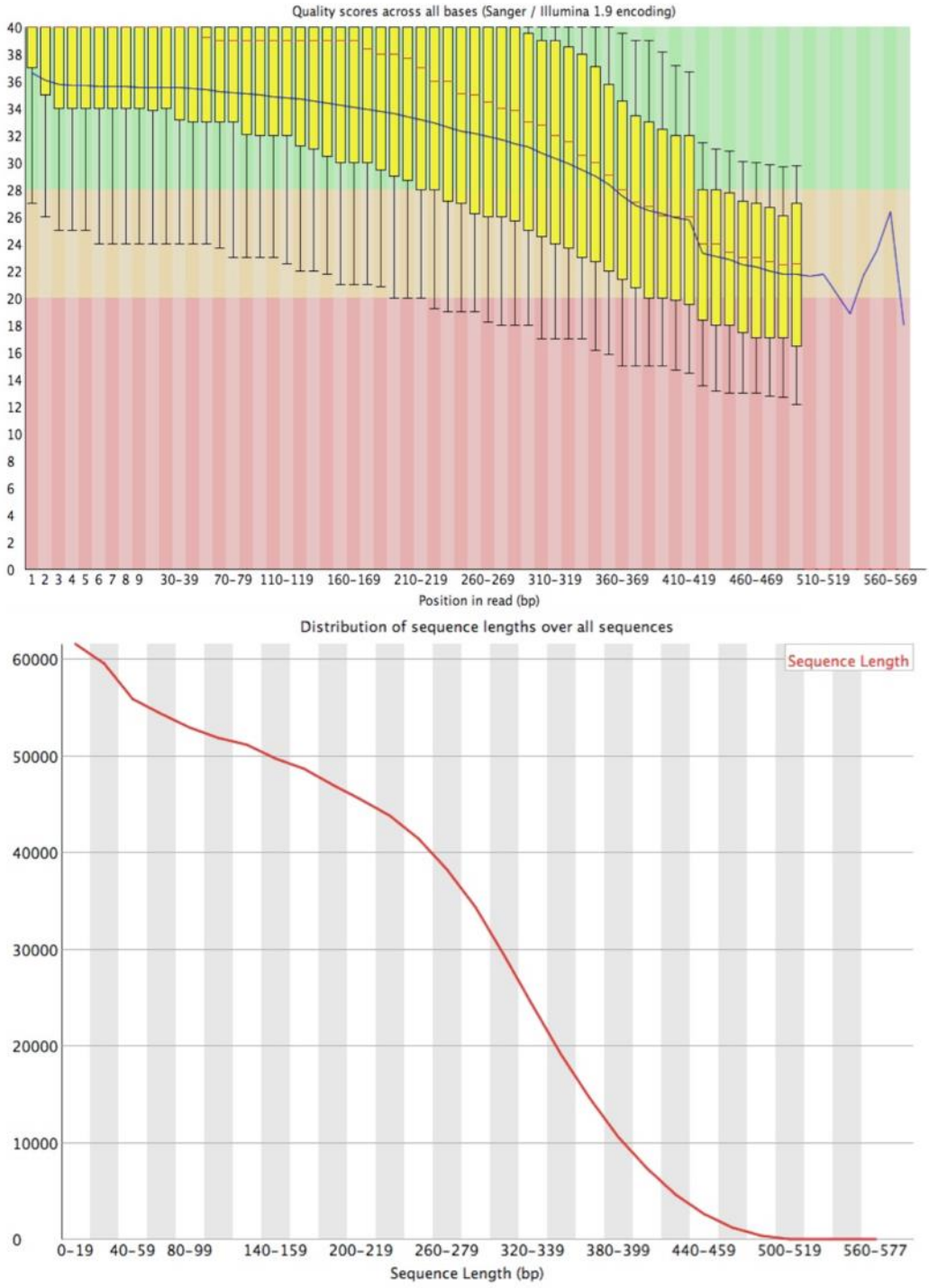

Figure 3.8 - Quality scores across all bases for the 454 paired-end file and distribution of sequence lengths.

Curiously, the read lengths showed an uneven distribution between the lengths of left and right mates within each pair (Fig. 3.9). In most cases one read was much longer than its mate, with some reads being over 500 bp and their mates less than 25 bp. Some of the smallest reads after removing linker were only 1 bp or higher whilst their mates were over 400 bp; these very small reads were discarded during subsequent analysis.

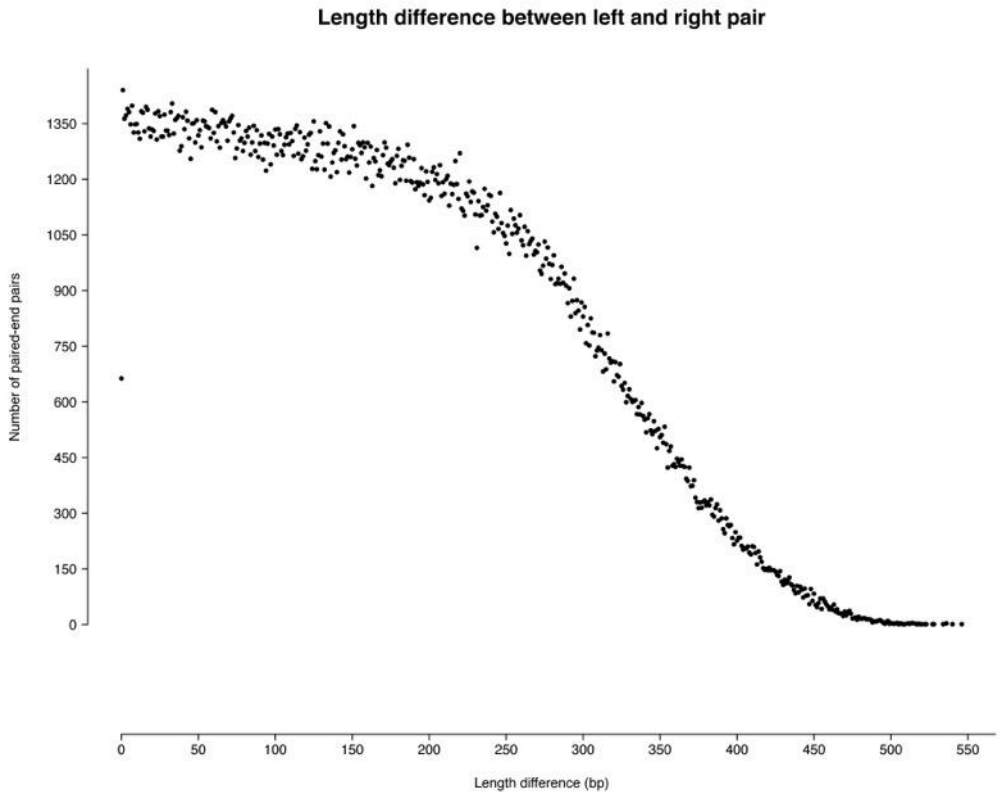

Figure 3.9 - Difference in the lengths of left and right paired-end reads. Over 600 pairs have same length reads but many more have one read much longer than its mate.

3.2.3 Alignment of 454 sequences to rDNA repeat

All 454 reads (shotgun and paired-end reads) were mapped to the rDNA repeat extracted from AL592188 (the most representative rDNA sequence available) using Mosaik aligner (Lee et al., 2014) (please see section 2.7. for alignments details). Although the average coverage was of 176.7x, we found very unequal coverage across the rDNA repeat (Fig. 3.10-A). The first 13 kb of the repeat (Fig. 3.10-B), where the coding sequences for the 18S, 5.8S and 28S ribosomal subunits are located also show uneven coverage. The 28S region displayed relatively high coverage (Fig. 3.10-C) but some regions showed no reads (Fig. 3.10-D).

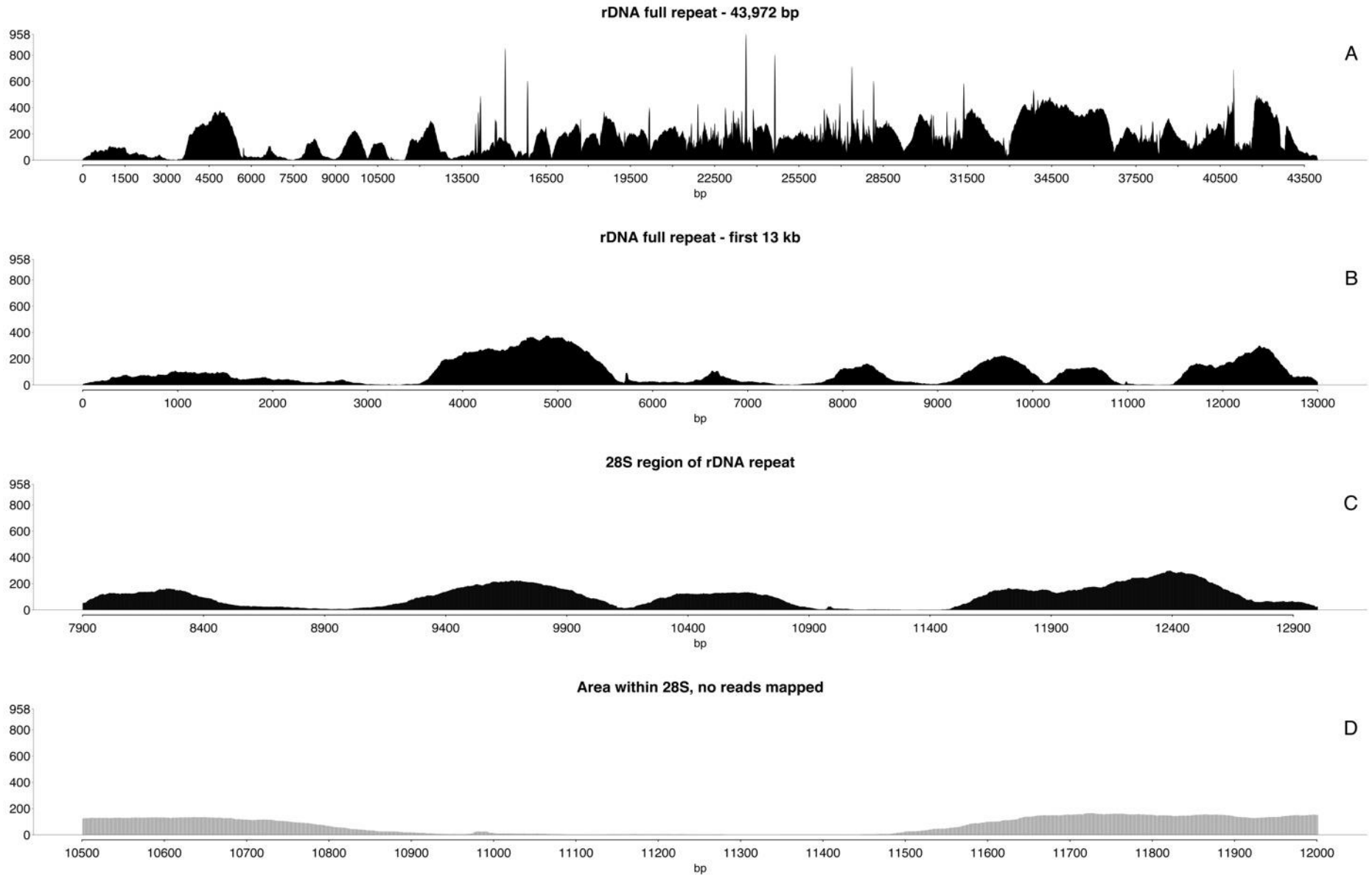

Figure 3.10 - All 454 reads mapped against the rDNA repeat extracted from AL592188. A - Coverage of the entire repeat by shotgun and paired-end reads. B - Coverage of the first 13kb of the rDNA repeat. C - Coverage of the 28S region. D - Example of 28S region where no reads mapped.

3.2.4 Aligning 454 reads to rDNA with 10% mismatches

All 454 reads were remapped to the rDNA repeat, allowing a larger number of mismatches per read (10%, compared to the default of 4 mismatches per read, regardless of read length). Given the variability in read length an error threshold based on the proportion of mismatches it is preferable in this case to a threshold consisting of a fixed number of mismatches. Unsurprisingly, this resulted in a greater proportion of mapped reads (3.1% of reads aligned compared to 2.2% using the default threshold) and higher coverage of the rDNA repeat (Fig. 2311). The majority of alignments occurred in the IGS region (Fig 11A) with the transcribed region still displaying a poorer coverage.

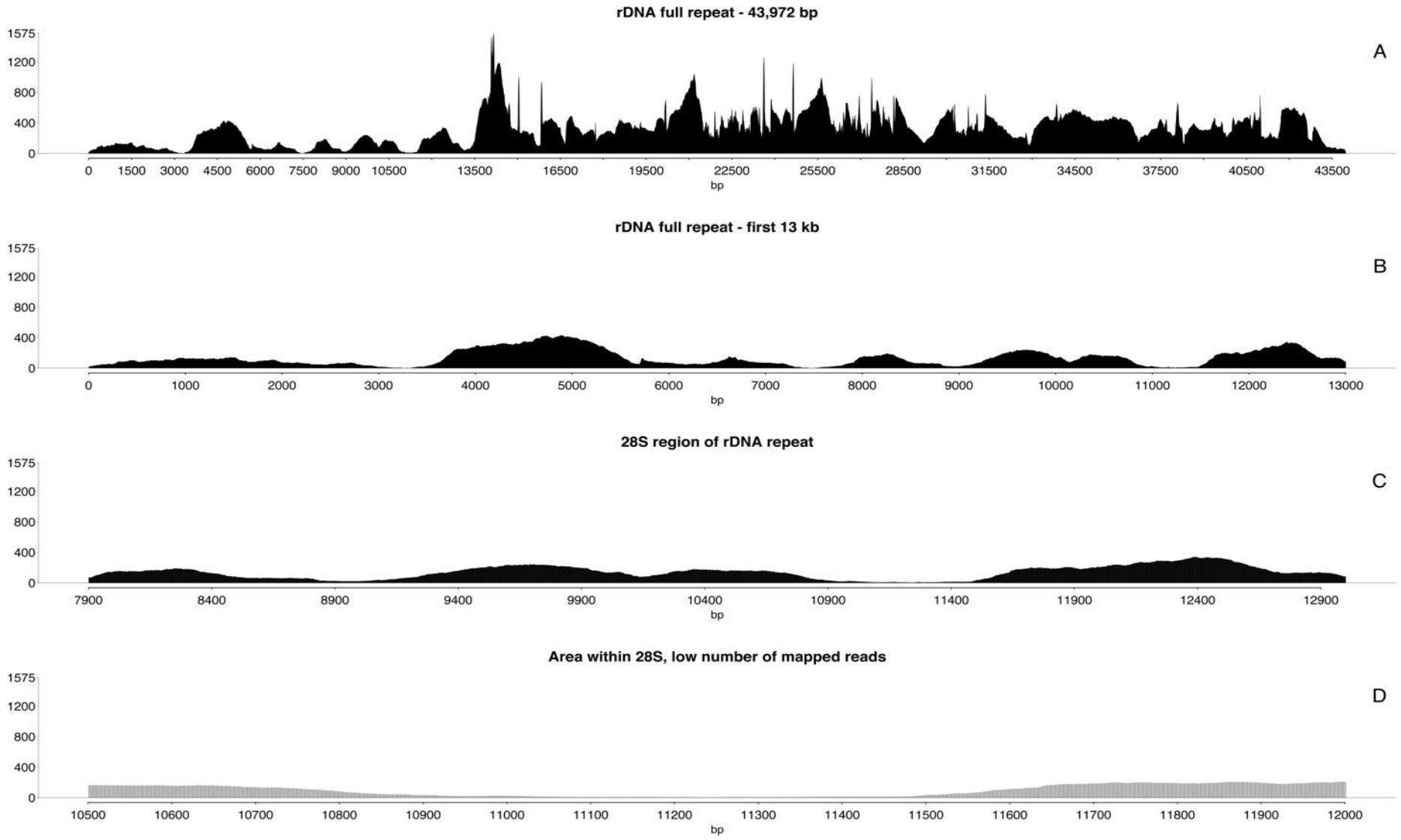

Figure 3.11 - Remapping of all 454 reads with 10% mismatches. A - Coverage of the full rDNA repeat, 3.1% of reads mapped, of which 2.4% are uniquely mapped reads. B – Mean coverage of 454 reads in the transcribed region improved slightly but it is still uneven. C – 28S region shows a moderate improvement of number of reads mapped. D – Region within 28S where very few reads mapped after increasing the mismatch threshold.

3.2.5 Paired-end alignments against rDNA repeat with 10% mismatches

Currently available aligners cannot map rearranged reads. A decision was made to use only the paired-end reads to look for proof of rearrangements. Technology employed by 454 determines that the paired-end sequences were delivered in a forward-forward orientation (Fig 3.12).

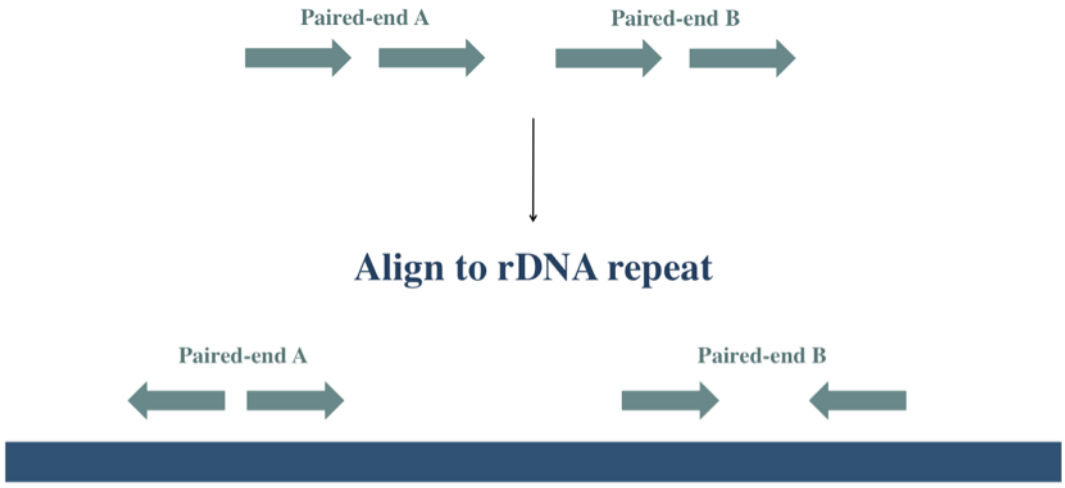

Figure 3.12 – Strategy to look for rearrangements using the 454 paired-end reads from nucleolar DNA. The reads within each pair have a forward – forward orientation. After alignment to the rDNA repeat, pairs that mapped in a backward – forward and forward - backward orientation should be indicative of rearrangements.

Consequently, proof of rearrangements can be achieved by finding uniquely mapped pairs (pairs where both reads only mapped once to the repeat) that have opposing mapping orientations. The paired-end library has a fragment insert of 3 kb, allowing a larger search area for rearrangements than the search area given by using individual reads. Also, as the length of reads varies greatly, only pairs with both reads larger or equal to 50 bp were included in the analysis.

The reads that mapped to the first 13kb in the coding sequences for the 18S, 5.8S and 28S ribosomal of the repeat were investigated (Fig. 3.13).

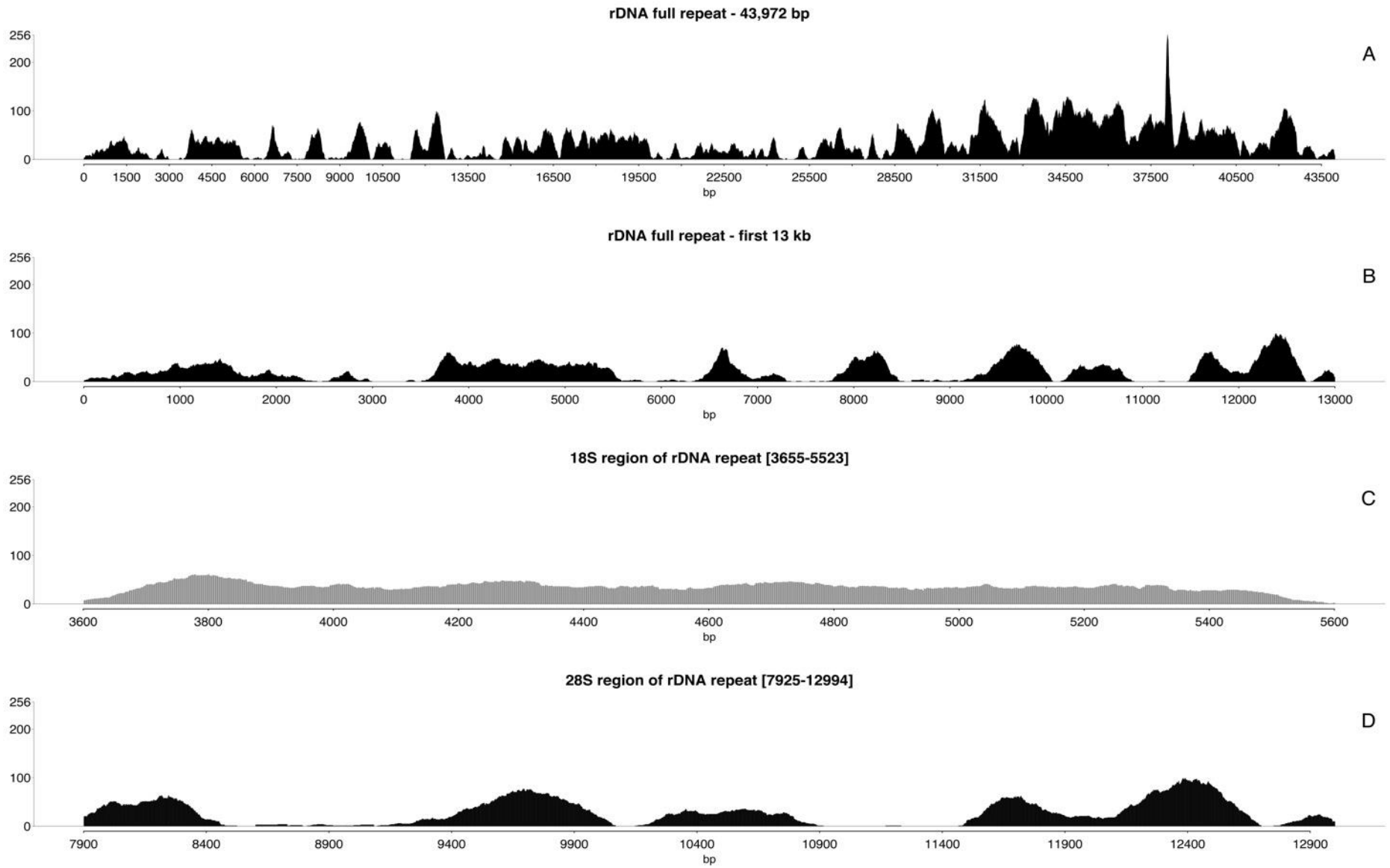

Figure 3.13 - Alignment of 454 paired-end reads to the rDNA repeat, allowing 10% mismatches per read. A - Coverage across the rDNA repeat, 1.1% of reads mapped (unique and non-unique alignments). B - Coverage of the transcribed region of the rDNA repeat shows as before unequal distribution of alignments. C – 18S region shows even coverage across the sequences. D – 28S region of the rDNA repeat sows fragmented distribution of alignments.

Pairs that mapped uniquely in opposing directions in the transcribed region of the rDNA repeat were isolated. There were no pairs indicative of rearrangements in the coding regions of rDNA.

However, this was not sufficient to preclude the existence of rearrangements. The secondary structure of the ribosomal coding sequences and the high GC content of the rDNA repeat could have led to a bias against these regions. The PCR amplification step employed by the 454 technology selects against sequences with these characteristics. A different sequencing method, with longer reads and more importantly with no PCR amplification could, in principle, circumvent the poor representation of these regions.

3.2.6 Generation of a new consensus rDNA sequence

A new consensus sequence for the rDNA repeat was generated (section 2.10 for method) from the alignment of all 454 reads with 10% mismatches. The new sequence is 43972 bp long and contained 161 mismatches relative to the rDNA repeat extracted from AL592188. Of the 73 mismatches present in the transcribed region, 53 were from unassigned bases due to lack of coverage in the alignment of 454 reads. The remaining 20 mismatches (Fig. 3.14) are listed in table 3.3.

Figure 3.14 - Comparison between the rDNA repeat extracted from AL592188 and the consensus generated from RPE-1 454 reads. Red lines mark the position of mismatches on AL592188 rDNA. The transcribed region (first 13 kb) is indicated by a blue line.

Table 3.3 - Nucleotide mismatches between AL592188 rDNA repeat and the new consensus generated from alignment of 454 reads

Nucleotide position	AL592188 rDNA	454 consensus
762	T	K (G or T)
1756	C	Y (C or T)
2338	A	R (A or G)
2512	C	G
2602	C	Y (C or T)
2891	G	C
2955	T	C
3714	G	A
5734	A	R (A or G)
5819	A	R (A or G)
6273	C	M (A or C)
6523	G	T
6524	G	T
7434	C	G
7789	T	Y (C or T)
7794	C	T
10965	G	A
12830	C	Y
12834	G	A
13013	C	T

3.2.7 DNA sample preparation for SMRT sequencing

For the next stage of the project, nucleolar DNA of two cell lines, RPE-1 and HeLa cells, was sequenced, using single molecule sequencing technology from Pacific Biosciences. HeLa cells unlike RPE-1 cells are not karyotypically normal. HeLa is an immortalised cervical cancer cell line that in addition to the human genome has also incorporated DNA from human papillomavirus 18 (HPV18). These cells typically have more than the expected 46 chromosomes, with almost the double of that number and massive rearrangements derived from chromothripsis (Landry et al., 2013).

Nucleolar DNA from RPE-1 was extracted as before for the 454 sequencing sample and gel electrophoresis (Fig. 3.15) was performed to check for ideal fragment length of extracted DNA (For methods please refer to section 2.2 – 2.5). FISH on normal metaphase chromosomes showed good enrichment for acrocentric short arms (Fig. 3.16) with small hybridisations to other chromosomes.

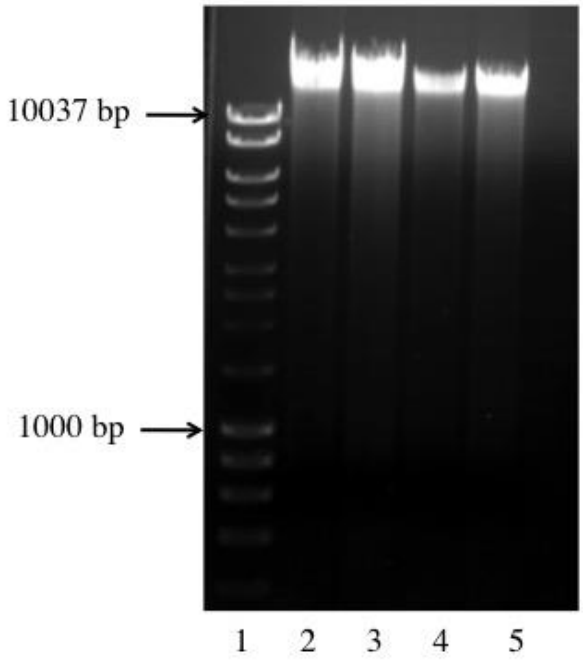

Figure 3.15 - Gel electrophoresis of purified nucleolar DNA from HeLa (second and third lane) and RPE-1 (fourth and fifth lane) cells. Hyperladder 1 kb was used. All samples show thick, clear bands with very high molecular weight DNA.

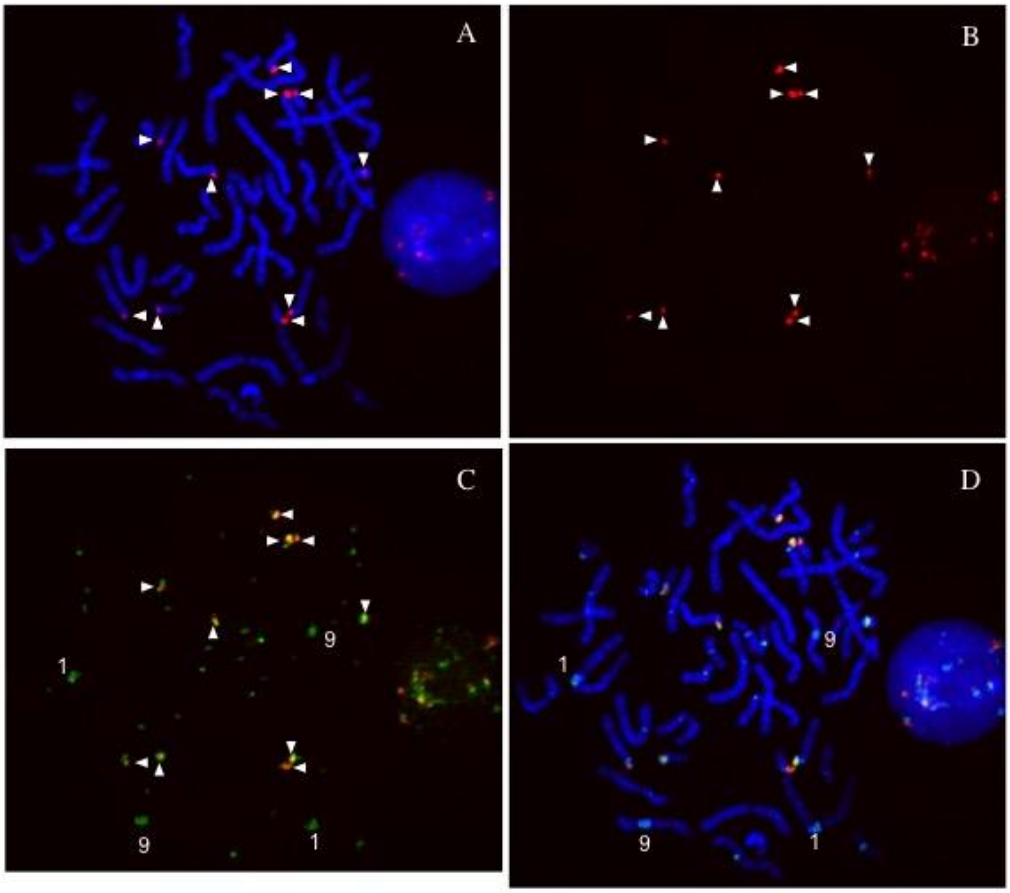

Figure 3.16 - FISH of RPE-1 nucleolar DNA (green) and rDNA (red). As expected, there is hybridisation to all acrocentric short arms (indicated by white arrows) and cross hybridisation to other chromosomes due to segmental duplication. FISH for rDNA marker in red (A and B) on DAPI stained chromosomes (A). Overlap of rDNA and nucleolar sample can be observed in panel C and on DAPI stained chromosomes in panel D.

For the HeLa sample, nucleoli were isolated from purified nuclei (section 2.2) followed by DNA extraction, (section 2.3.). FISH confirmed enrichment of acrocentric short arms (Fig. 3.17) and short cross-hybridisations with other chromosomes most likely due to segmental duplications from the sequences adjacent to NORs (Floutsakou et al., 2013).

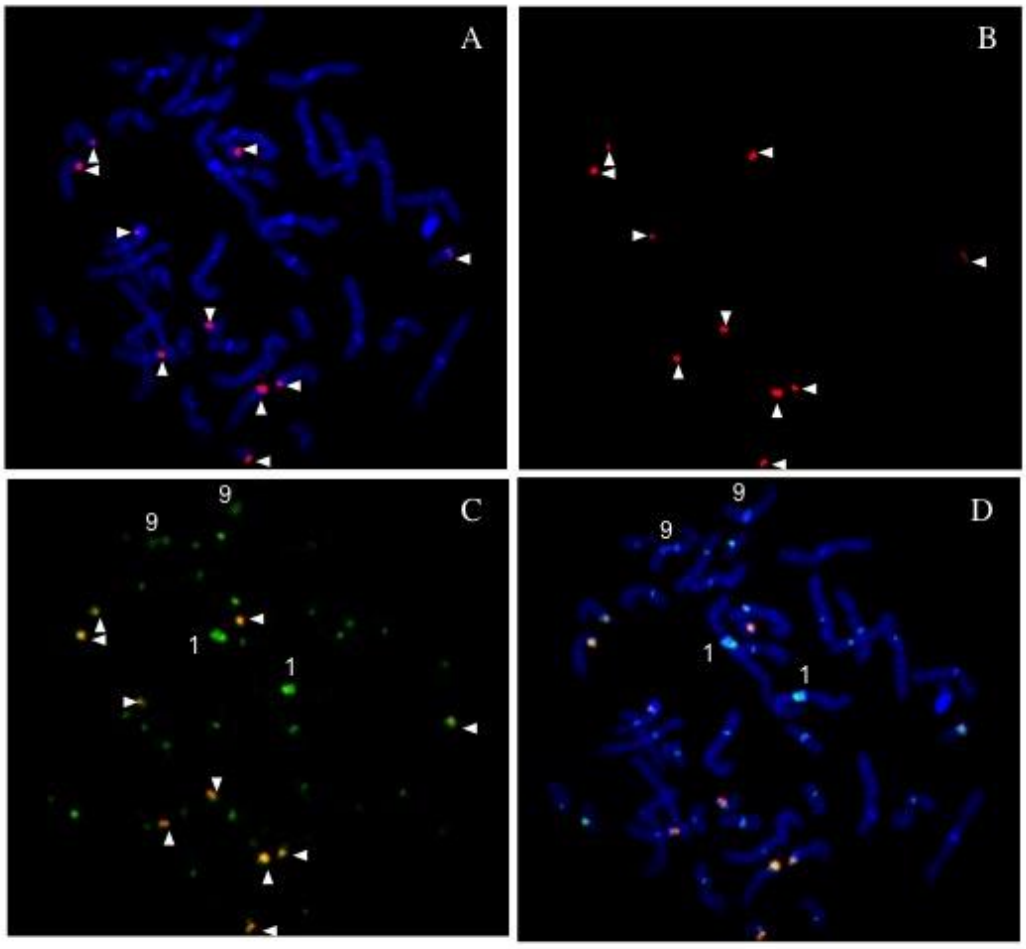

Figure 3.17 - FISH of HeLa nucleolar DNA (green) and rDNA (red). All acrocentric short arms are represented with hybridisation of sequences in and around rDNA. Evidence of cross-hybridisation with other chromosomes can also be observed. FISH for rDNA marker in red (A and B) on DAPI stained chromosomes (A). Overlap of rDNA and nucleolar HeLA sample can be observed in panel C and on DAPI stained chromosomes in panel D.

3.2.8 PacBio sequencing and sequence quality control

Over 75,000 unfiltered PacBio reads were delivered by GATC Biotech for each cell sample along with quality reports for the RPE-1 (Fig. 3.18) and the HeLa (Fig. 3.19). Low quality reads were removed prior to alignment with the fastq_quality_trimmer tool from FASTX toolkit (section 2.8 for details).

Figure 3.18 - PacBio quality report for RPE-1 nucleolar sample. According to the report Pacific Biosciences sent, after quality filtering there were 50,170 reads with average read length of 3117 bp and mean quality of 0.843.

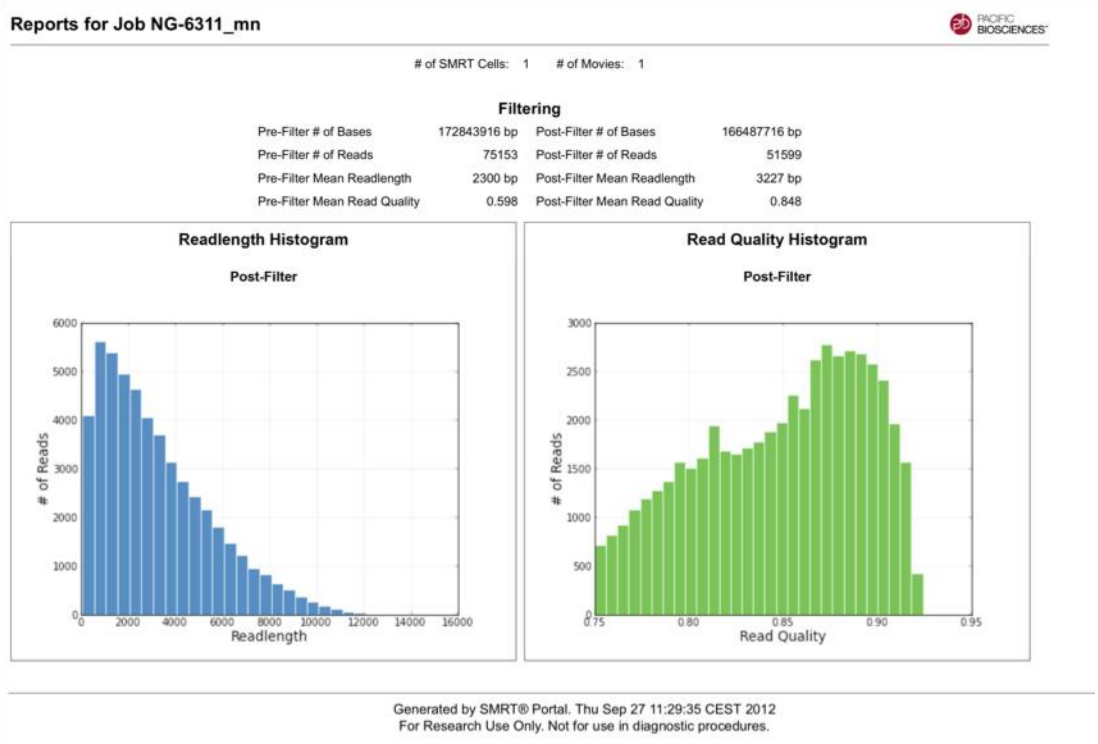

Figure 3.19 - PacBio quality report for the HeLa nucleolar sample. After filtering there were 51,999 reads with average read length of 3227 bp and mean quality of 0.848.

3.2.9 PacBio alignments against rDNA using the BLASR aligner

All filtered PacBio reads, from both samples, were mapped to the rDNA repeat using BLASR (section 2.9). There was a higher proportion of reads mapped for both the RPE-1 (Fig. 3.20) and the HeLa sample (Fig. 3.21) compared to the RPE-1 454 data. Average coverage of the transcribed region improved, when compared to the alignment of 454 RPE-1 nucleolar reads (Fig. 3.20-B and 3.21-B). The 28S showed good coverage, as all regions showed mapping of sequences (Fig 3.20-D and 3.21-D), unlike the alignment of 454 reads where some segments were not represented (e.g. regions [3000, 3500] and [10900, 11500]). Interestingly, high peaks can be observed in the IGS (Fig. 3.20-A and 3.21-A). These might be repetitive elements that can be found in other areas of the genome, increasing the overall mappability capacity of these regions of rDNA.

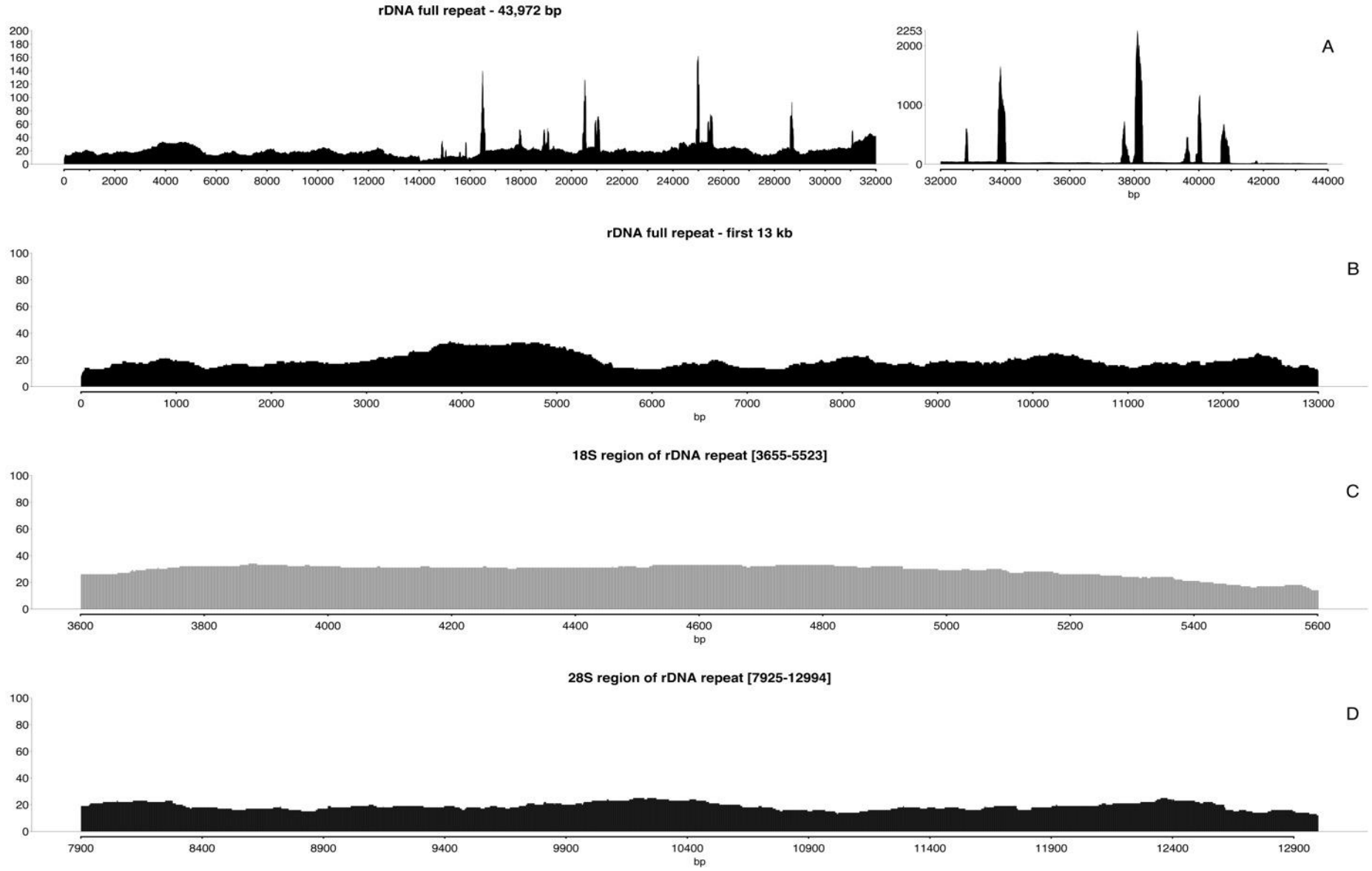

Figure 3.20 - Alignment of RPE-1 PacBio reads to the rDNA repeat extracted from AL592188. Around 14% of filtered reads mapped. A - Coverage of the entire rDNA repeat shows uneven distribution with the majority of reads mapping to the IGS. Scale of the y-axis was adjusted to better illustrate the coverage of different regions. B – Coverage of the transcribed region of rDNA. C – Distribution of PacBio reads across the 18S region. D – Distribution of PacBio reads across 28S region of rDNA.

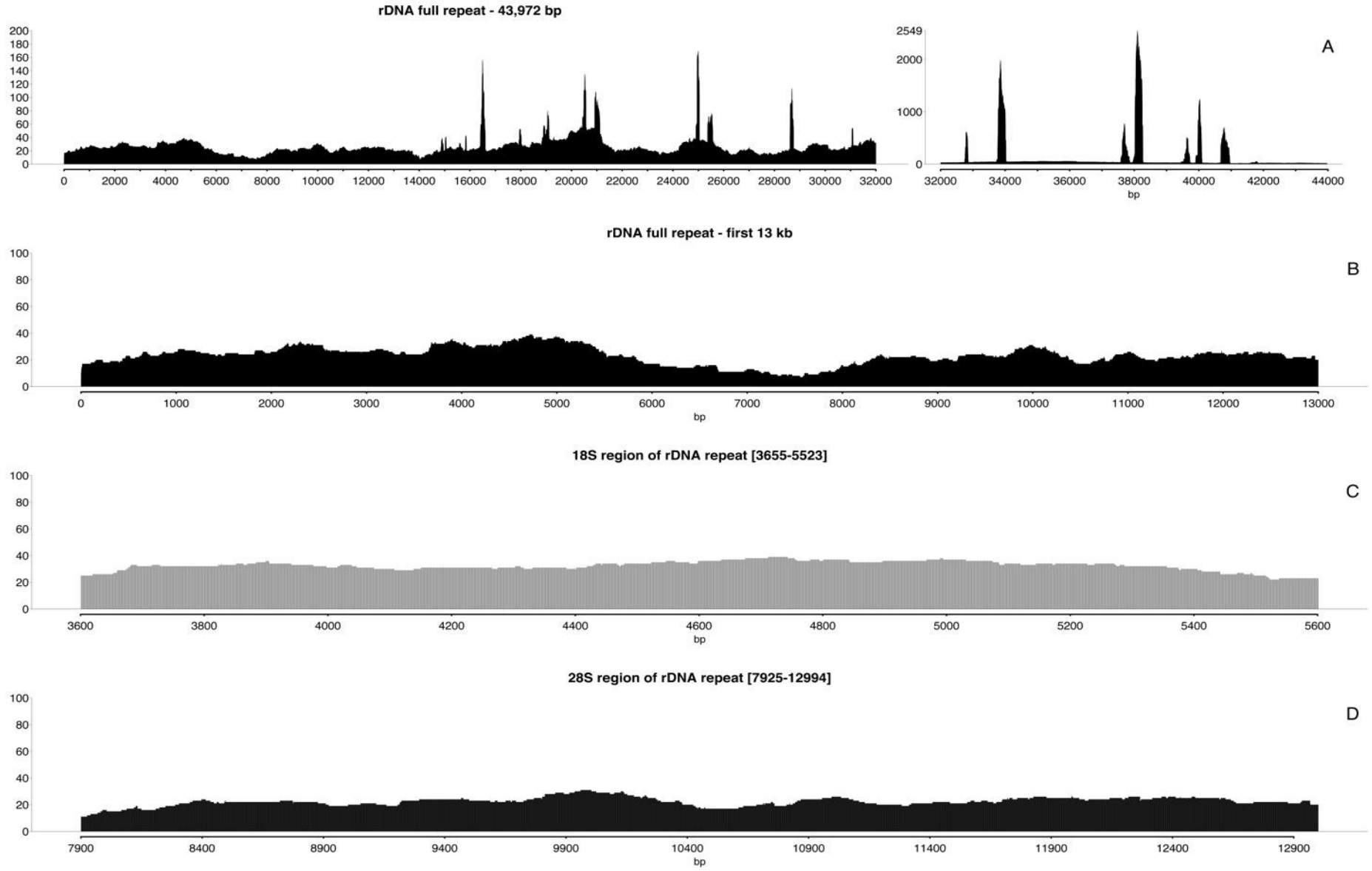

Figure 3.21 - Alignment of nucleolar HeLa PacBio reads to the rDNA repeat. 15% of reads mapped. A - The majority of reads mapped to the IGS. B - The transcribed region of rDNA shows no unmapped regions. The y-axis scale was adjusted to better illustrate each region. C – Distribution of PacBio reads across the 18S region of rDNA. D – Distribution of PacBio reads across the 28S region of rDNA.

The AL592188 sequence comprises the first rDNA repeat after the Distal Junction sequence (Floutsakou et al., 2013). Although there is high sequence identity to all the following repeats, there is the possibility of small differences between the sequences.

The improved coverage of the transcribed region of rDNA and the number of reads aligned provided an opportunity to generate new consensus sequences for the rDNA repeats from the RPE1 and HeLa PacBio reads.

3.2.10 Generation of new rDNA consensus from nucleolar PacBio

The new consensus sequences were generated from the alignments (section 2.10) for the two sequenced cell lines. The generated consensus for RPE-1 was 43972 bp long. Assessment of identity between this new consensus and the AL592188 rDNA repeat using BLAST revealed query cover of 100% and 99.8% similarity with 91 mismatches (Fig. 3.22). The majority of mismatches were present in the IGS, with only 3 mismatches in the transcribed region (located in the 5' ETS, the 18S region and in the ITS2, respectively).

The consensus for the HeLa sample showed the same characteristics (when aligned to AL592188), 100% query cover and 99.8% identity to rDNA with 88 mismatches (Fig. 3.23). Of the 3 mismatches in the transcribed region, 2 were in the same nucleotide positions as in the RPE-1 consensus, in the 18S and the ITS2 (Table 3.4). The other mismatch was also located in the 5'ETS but in a different position.

Table 3.4 - Reported mismatches for the RPE-1 and HeLa consensuses relative to the rDNA repeat from AL592188

Nucleotide Position	rDNA AL592188	RPE-1	HeLa
1756	C	T	C
2602	C	C	T
3714	G	A	A
7794	C	T	T

Comparison between the two new consensuses revealed a total of 61 mismatches, with 2 of those in the 5'ETS region, as expected, and the remainder in the IGS. The differences found in the IGS were not concerning, as this region is not expected to be as conserved as the transcribed sequences.

Alignment of the two consensuses against the U13369 rDNA repeat showed a larger number of mismatches, 532 for RPE-1 (Fig. 3.24), with 82 mismatches and 110 gaps in the transcribed region and 523 for HeLa (Fig. 3.25), with 78 mismatches and 110 gaps in the transcribed region.

Figure 3.22 - Comparison between the rDNA repeat from AL592188 and the consensus generated from RPE-1 PacBio reads. Red lines mark the position of mismatches on AL592188. The transcribed region (first 13 kb) is indicated by a blue line.

Figure 3.23 - Comparison between the rDNA repeat from AL592188 and the consensus generated from HeLa PacBio reads. Red lines mark the position of mismatches on AL592188. The transcribed region (first 13 kb) is indicated by a blue line.

Figure 3.24 - BLAST alignment of generated consensus sequence from RPE-1 PacBio reads against the U13369 rDNA repeat. Many mismatches can be observed between the two sequences, with the majority occurring in the IGS.

Figure 3.25 - BLAST alignment of generated consensus sequence from HeLa PacBio reads against the U13369 rDNA repeat. Many mismatches can be observed between the two sequences, with the majority occurring in the IGS.

3.2.11 Search for rDNA rearrangements with RPE-1 and HeLa PacBio reads

The search for rearrangements in the RPE-1 and HeLa PacBio sequences was performed using the same strategy employed previously in the 454 reads. PacBio sequencing only included single-end reads, however, each read can be treated as a DNA fragment prior to sequencing. Sub-sequences from either side of any read longer than 2kb were mapped to the rDNA repeat (Fig. 3.26).

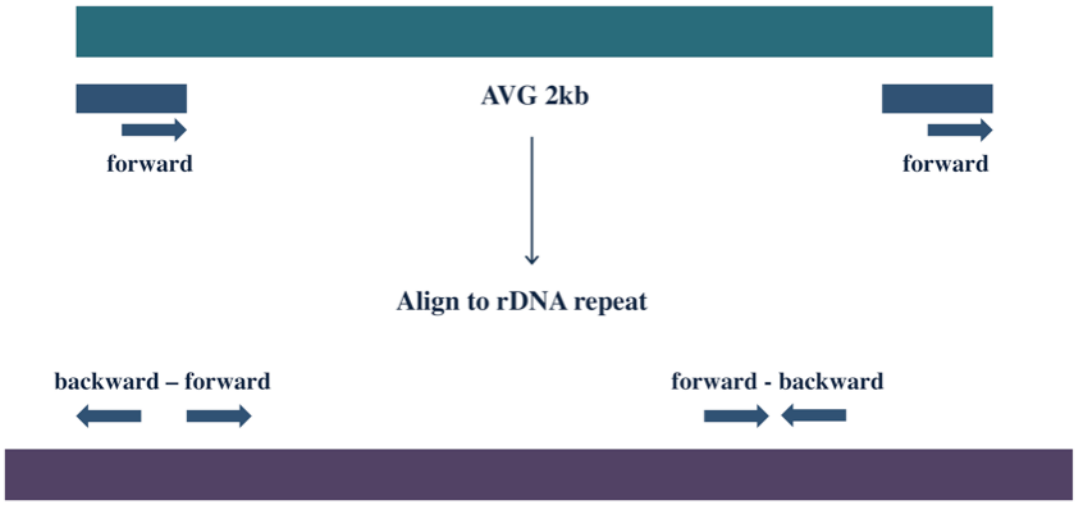

Figure 3.26 – Sequences from either side of the PacBio reads were used to look for rearrangements. As with the 454 paired-end data, unique alignments in opposing orientations would indicate possible rearrangements.

The orientation of the sub-sequences is the same as the original read, forward-forward, therefore, rearrangements can be searched by looking for two reads from a pair that mapped in opposing orientations. Alignments were carried out in BLASR against the first 13kb of the rDNA repeat. Of the unique

alignments reported, 7 pairs showed rearrangement for RPE1 and 4 for HeLa. However, these reported rearrangements were due to sequencing artefacts.

Following this analysis, it was encouraging to know that our strategy worked and existing rearrangements could be found with it. To test our strategy, we aligned to the rDNA repeat a small set of sequences artificially generated to depict rearrangements as observed in molecular combing. This approach confirmed the efficacy of our method, as all reads were reported as containing rearrangements,

Repeating this analysis in a different data set with, preferably, higher coverage of the rDNA repeat would give us another opportunity to look for the existence of rearrangements. Recently, Pacific Biosciences produced and released to the public their own 54x human genome data set that was used to create a new *de novo* reference assembly.

3.2.12 Analysis of CHM1 PacBio reads

PacBio reads were retrieved in FASTA format in several files. Therefore, quality control was not employed before undergoing alignments. Setting a minimum percentage identity between the reads and the references should throw out reads with a high number of errors.

Alignments were carried out using BLASR (section 2.9) and showed excellent coverage of the rDNA repeat (Fig 3.27).

3.2.13 Search for rearrangements with CHM1 Pacbio reads

The same strategy of using sub-sequences from either side of PacBio reads and aligning these to the rDNA repeat was used (please refer to sections 2.9 and 2.11).

As before, only pairs of sub-sequences that mapped uniquely were considered, and pairs with sub-sequences that mapped in opposing orientations in the 18S, 5.8S and 28S regions of the rDNA repeat were searched after. Of all the pairs that mapped, only 27 were marked as rearrangements. However, none of the original sequences revealed any of the patterns observed in DNA combing, with the majority of reads only showing either an inversion in the 18S or 28S. Molecular combing data showed that one third of fibres depict rearrangements. We found a low number, 27, of rearrangements reported relative to the number of reads that mapped to the transcribed region (over 25,000). The majority of these reads also contained inverted copies of either only the 18S or 28S sequences.

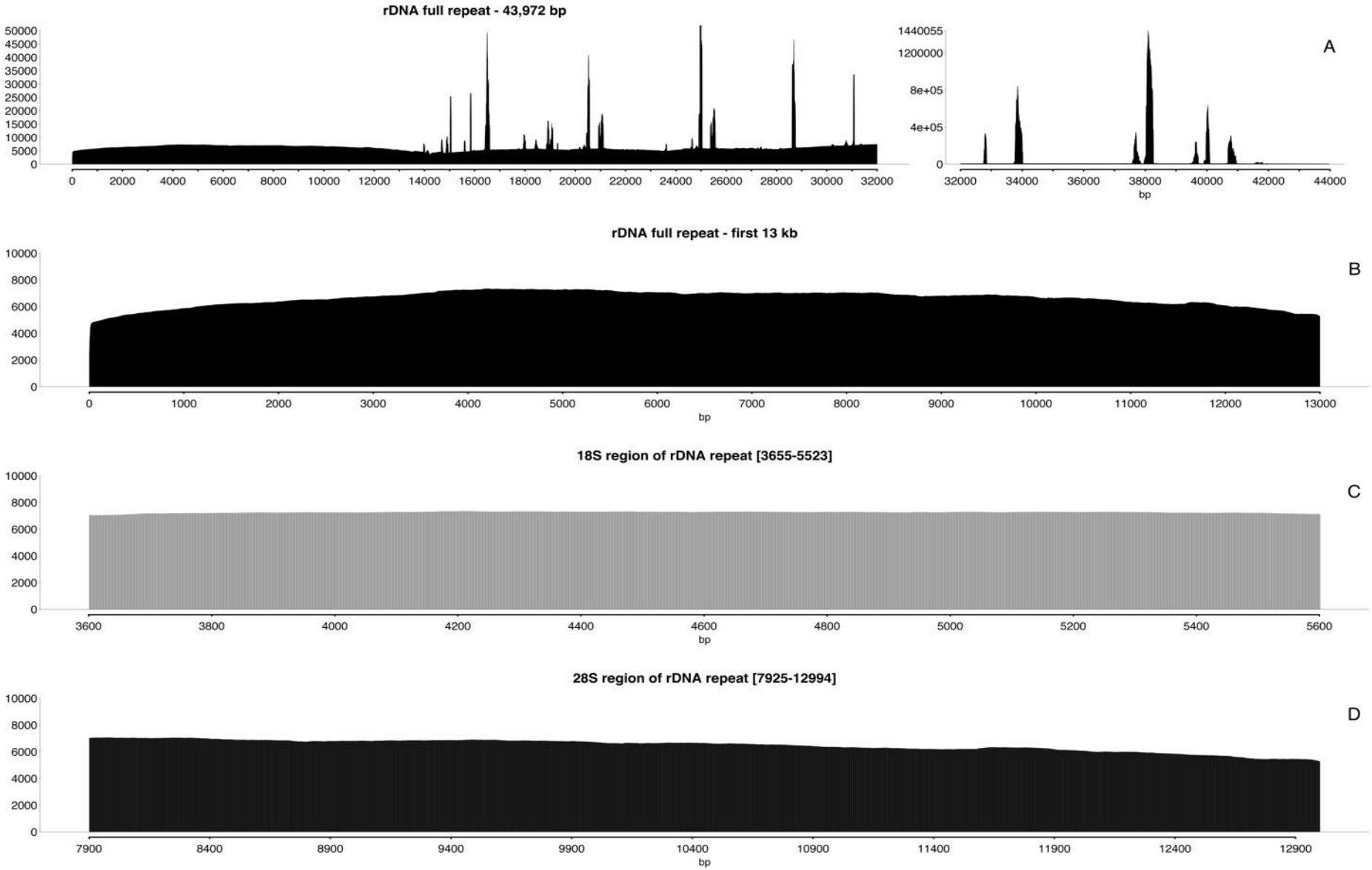

Figure 3.27 - Alignment of CHM1 PacBio reads against the rDNA repeat extracted from AL592188. A - Alignment coverage of the full repeat shows the majority of reads mapped to the IGS. The y-axis was adjusted to better illustrate coverage in the different regions of the repeat. B – Distribution of PacBio reads on the rDNA transcribed region. C – Distribution of PacBio reads in the 18S region. D – Distribution of aligned reads in the 28S region of rDNA.

3.2.14 Generation of a new consensus sequence from CHM1 reads

A new consensus sequence was generated from the alignment (section 2.10) for the CHM1 cell line. The generated consensus was 43972 bp long. Assessment of identity between this new consensus and the AL592188 rDNA repeat using BLAST revealed query cover of 100% and 99.94% similarity with 22 mismatches (Fig. 3.28 and Table 3.5). The majority of mismatches occurred in the transcribed region, with only 4 mismatches in the IGS (Table 3.5). Comparison between this new consensus and the rDNA repeat U13369 revealed many mismatches (Fig. 3.29).

Figure 3.28 - Comparison between the rDNA repeat from AL592188 and the consensus generated from CHM1 PacBio reads. Red lines mark the position of mismatches on AL592188 rDNA. The transcribed region (first 13 kb) is indicated by a blue line.

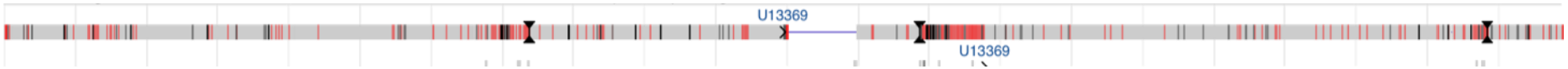

Figure 3.29 – Blast alignment of CHM1 consensus against the rDNA repeat U13369.

Table 3.5 - Mismatches between rDNA repeat from AL592188 and CHM1 consensus

Nucleotide position	AL592188 nucleotide	CHM1 consensus nucleotide
139	T	Y (C or T)
1756	C	Y (C or T)
2338	A	R (A or G)
2512	C	S (C or G)
2602	C	Y (C or T)
2891	G	S (C or G)
2955	T	Y (C or T)
3207	C	M (A or C)
3714	G	R (A or G)
5734	A	R (A or G)
5819	A	R (A or G)
6273	C	M (A or C)
6523	G	K (G or T)
6524	G	K (G or T)
6526	G	S (C or G)
7794	C	Y (C or T)
12834	G	R (A or G)
13013	C	Y (C or T)
13521	G	R (A or G)
13556	A	M (A or C)
14399	T	Y (C or T)

3.2.15 Improvement of CHM1 alignments against the rDNA repeat

All alignments reported by BLASR were local alignments with at least 85% identity to the reference. However, BLASR does not offer options to enforce global mapping or to only report reads that mapped in their entirety. The number of reads that mapped to the rDNA repeat (5,210,137) was quite high. The total number of reads in the data set is 22,565,609, which gives a 54x coverage of the human genome. This means that around 23% of the PacBio reads from CHM1 mapped to the rDNA sequence, which is not realistic as the rDNA repeats represent less than 0.05% of the human genome. The SAM file was then parsed to only consider alignments that were at least 80% of the length of the read. The threshold of 90% was chosen to account for gaps in the alignment and the accuracy of PacBio reads. After this selection, the number of alignments dropped to 39,706 (32,564 reads). Despite the low number of alignments, coverage of the rDNA repeat revealed no gaps or unaccounted regions (Fig. 3.30). Interestingly, there was a visible dip in the coverage distribution around 14,000 bp. A new consensus for rDNA was created from the parsed alignments. The consensus differed from the AL591856 repeat in 20 mismatches - 100% query cover and %99.95 identity (Fig. 3.31 and Table 3.6). The majority of mismatches occurred in the transcribed region, with only 3 mismatches in the IGS. Comparison between the previous CHM1 generated consensus and this new one showed the same mismatches to the AL592188 rDNA repeat. The mismatches at 6273, 6526 and 13013 were not found in the new consensus and two new mismatches at 2620 and 6525 were reported (Table 3.6).

Table 3.6 - Mismatches between rDNA repeat from AL592188 and CHM1 consensus (85% identity and alignment length at least 90% read length)

Nucleotide position	AL592188 nucleotide	CHM1 consensus nucleotide
139	T	Y (C or T)
1756	C	Y (C or T)
2338	A	R (A or G)
2512	C	S (C or G)
2602	C	Y (C or T)
2620	A	M (A or C)
2891	G	S (C or G)
2955	T	Y (C or T)
3207	C	M (A or C)
3714	G	R (A or G)
5734	A	R (A or G)
5819	A	R (A or G)
6523	G	K (G or T)
6524	G	K (G or T)
6525	G	S (C or G)
7794	C	Y (C or T)
12834	G	R (A or G)
13521	G	R (A or G)
13556	A	M (A or C)
14399	T	Y (C or T)

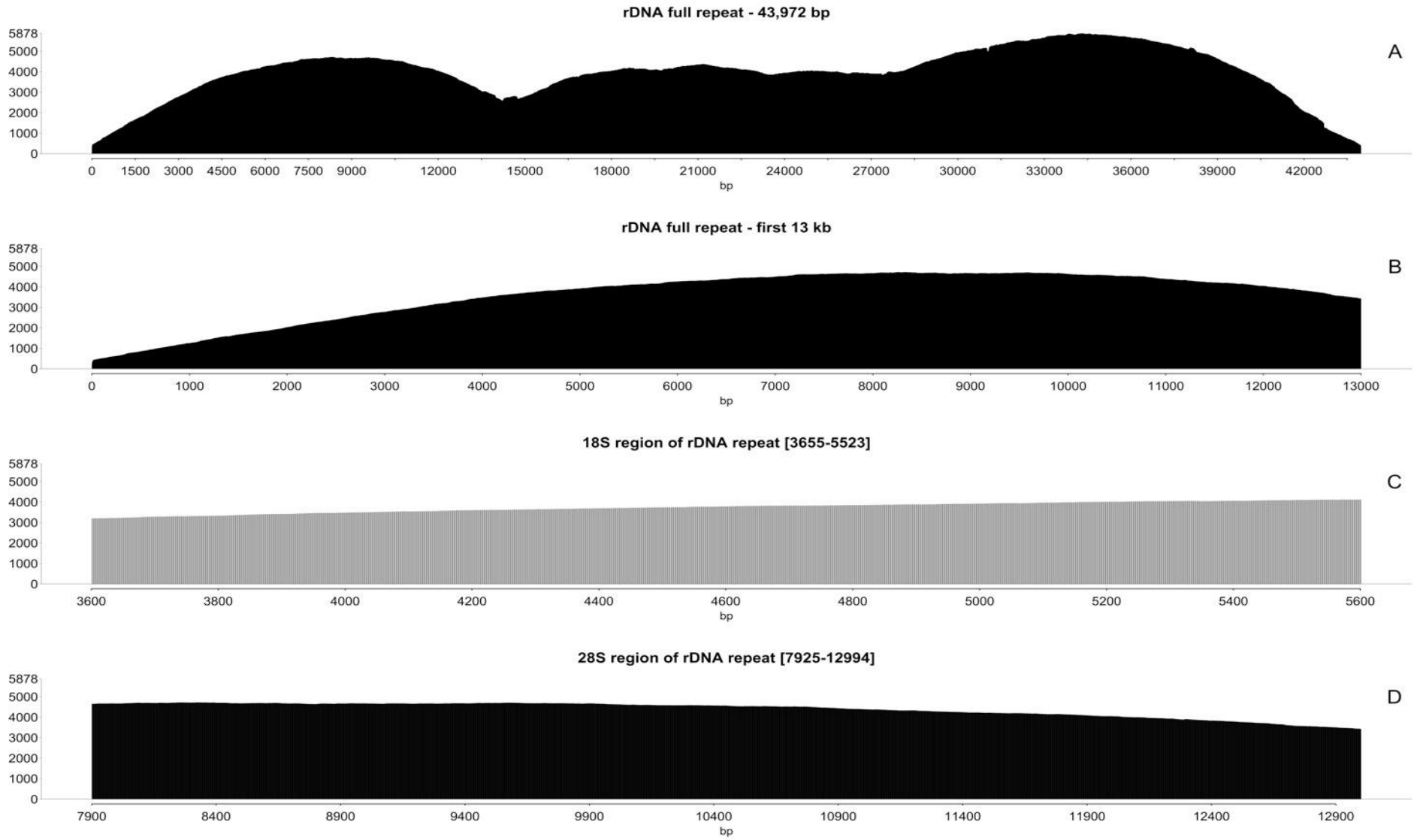

Figure 3.30 - Alignment of CHM1 PacBio reads to the rDNA repeat extracted from AL592188. Only alignments with at least 80% of the read length aligned with 85% identity were considered. A- Distribution of alignments to the full repeat (~44 kb). B - Alignments in the transcribed region (first 13 kb). C - Distribution of alignments in the 18S region of rDNA repeat. D - Alignments in the 28S region of rDNA repeat.

Figure 3.31 - Comparison between the rDNA repeat from AL592188 and the consensus generated from CHM1 PacBio reads (85% alignment identity with at least 80% length of read aligned). Red lines mark the position of mismatches on AL592188 rDNA. The transcribed region (first 13 kb) is indicated by a blue line.

3.3 Discussion

The majority of rearrangements in the rDNA genes observed through molecular combing are in the form of 3'-3' or 5'-5' palindromic units or closely spaced, inverted 18S and 28S units (Caburet et al., 2005). Sequencing paired-end reads can be used to search for rearrangements if reads in a pair map in opposing orientations. Unfortunately, the nucleolar 454 paired-end reads failed to report any rearrangements. These sequences had however, very low coverage in the transcribed region. Potentially, these could be a reason for not finding evidence of these events. Molecular combing data depicting rearrangements was performed in HeLa cells by the McStay lab and in fibroblasts and lymphoblastoid cells from patients with Werner syndrome (WS) and from control individuals (Caburet et al., 2005). Although the molecular combing showed more rearrangements in the WS samples, the healthy cells also reported rearrangements. Following this line of thought, we decided to sequence nucleolar DNA from RPE-1 and HeLa cells using PacBio technology. These two cell lines were chosen for their availability in the lab and their nature. RPE-1 is a healthy karyotypically normal cell line and HeLa is a cancer cell line with atypical number and organisation of chromosomes. Sequence coverage was quite high when the full PacBio reads were aligned to the rDNA repeat extracted from AL592188. However, rearrangements found from sub-sequences from either side of the reads were not real rearrangements. When further explored it was revealed the reported rearrangements were due to the PacBio adapter used in the technology still being present in the original reads (Fig 3.32). The DNA Polymerase generated a new sequence until reaching the adapter, at which stage,

turned back and continued adding nucleotides to the new sequence having as template the DNA molecule it was previously sequencing.

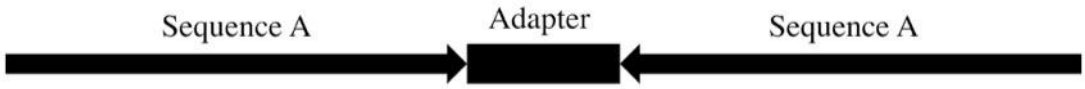

Figure 3.32 – Scheme of PacBio reads, containing the adapter, that reported rearrangements.

Genomic sequencing PacBio reads from the CHM1 cell line were also searched to look for rearrangements. However, a very low number of reads reported rearrangements compared to the number of reads that mapped to the rDNA repeat. This suggests these reads are not examples of rearrangements but possibly sequencing artefacts.

Nucleoli form around active rDNA repeats. If rearranged rDNA repeats are not being transcribed due to the defective rRNA that it would produce, it is possible that if rearrangements are real, they are not being captured through sequencing data. However, the PacBio reads from the CHM1 cell line were generated from genomic DNA and should have reads representing both silent and active repeats. Very few rearrangements were reported in that data set. This suggests that more sequencing data is needed to continue the search for these events. Furthermore, the rearrangements observed in the combing data could be artefacts (several fibres adhered together or even DNA strands that were being replicated and that got pulled together). In order to pursue this matter, molecular combing should be repeated on non-cycling cells. That is for example, non-dividing cells in the G0 phase. This can be achieved through serum starvation.

High-throughput and single molecule sequencing data of nucleolar and genomic DNA from 3 different cell lines enabled us to improve the consensus sequence for the rDNA gene. The rDNA repeat used as reference to align the reads, a sequence extracted from BAC AL592188, is the first repeat after the distal junction, towards the telomere. This repeat is a high quality sequence generated by Sanger sequencing and is therefore highly representative of the rDNA gene.

Regarding the transcribed region of rDNA, the alignment-generated consensuses from RPE-1 454 reads, RPE-1 PacBio reads and HeLa Pacbio reads indicated new nucleotides at positions 3714 (A instead of G) and 7794 (T instead of C). Whilst the consensus generated from the CHM1 PacBio reads was ambivalent at these positions, it indicated A or G at position 3714 and T or C at position 7794. The most likely possibility is that both nucleotides at each position are true, and that there are differences between the first repeat in the array and the following repeats, or at least some of them. The mismatch at 3714 is located in the 18S sequence, which, unlike the other mismatch located in the ITS2, is not spliced out. This could result in a different but fully functioning isoform or it can impair ribosomal function and even its assembly.

Many other mismatched nucleotides were reported in the new consensuses in various positions. Some of these nucleotides overlapped in two or three generated consensuses but not in the others or other (tables 3.3, 3.4 and 3.5). This strengthens the possibility that there are small differences between the different repeats, or that repeats are dissimilar between the short arms of acrocentric chromosomes. The other reported mismatches could be from sequencing errors or polymorphisms between cell lines. The consensus from the

RPE-1 454 reads had the lowest query cover and sequence identity of the generated consensuses. This was due to the low coverage of these reads in the transcribed region.

Sequencing individual short arms of acrocentric chromosomes with reads that span the length of two or three repeats (longer than the reads currently available) would provide a better representation of the rDNA repeats and give a more suitable framework of these regions of the genome.

4 Spatial Organisation of the Distal Junction

4.1 Background

4.1.1 Functional relevance of genome spatial organisation

The human genome project aimed to identify the 3.2 billion base pairs of the human genome. This reference genome provided a framework to identify and localise genes as well as functional and regulatory elements (Consortium, 2004). However, sequence information and epigenetic mechanisms are not the only contributors to gene silencing and activation. The folding and disposition of chromatin within the nucleus also modulates gene expression (Bickmore and van Steensel, 2013; Finlan et al., 2008; Kurz et al., 1996). The folding of the genome can bring into close spatial proximity functional elements, such as enhancers and promoters, from distant locations (Carter et al., 2002; Giorgetti et al., 2014; Jin et al., 2013; Sanyal et al., 2012; Tolhuis et al., 2002). The condensation and decondensation of chromatin is also involved in DNA damage signalling and repair (Burgess et al., 2014). Synthetic transcription factors can induce repositioning of genes toward the interior of the nucleus in embryonic stem cells by activating transcription (Therizols et al., 2014).

Within the nucleus, chromosomes are organised into chromosomal territories (CTs) or, more recently, chromatin domain clusters (CDCs), of active

(ANC) or inactive (INC) nuclear compartments, which in turn consist of smaller subdomains (Chen et al., 2015; Cremer and Cremer, 2001; Cremer and Cremer, 2010; Parada et al., 2002; Tanabe et al., 2002). The ANC encompasses the transcriptionally active periphery of CDCs and the interchromatin compartment (IC). ICs are connected to nuclear pores through a network of channels mostly devoid of DNA. These DNA empty spaces contain the transcription, splicing, replication and repair complexes (Albiez et al., 2006; Markaki et al., 2010). Importantly, the location of genes within these territorial subdomains offers easier access to the transcription and splicing machineries (Markaki et al., 2010). Different techniques for chromosome conformation capture complemented by high-resolution FISH have revealed the existence of topologically associated domains (TADS) of sizes from 100 kb to 1 Mb (Giorgetti et al., 2014; Lieberman-Aiden et al., 2009). Their position in the genome is largely conserved between cell types and their sizes resemble replication domains (Dixon et al., 2012; Pope et al., 2014). The boundary regions of TADs are thought to constrain the proliferation of heterochromatin and are enriched for housekeeping genes, SINEs, tRNAs, and CTCF sites (Dixon et al., 2012). CTCF, an insulator binding protein, together with the cohesin complex, contributes to modulate chromatin organisation and gene expression (Phillips and Corces, 2009; Wendt et al., 2008; Zuin et al., 2014). Disruption of cohesin results in an overall loss of local chromatin interactions whilst TADs remain unaltered (Gosalia et al., 2014; Zuin et al., 2014). Depletion of CTCF however, increases interdomain interactions and reduces intradomain interactions (Zuin et al., 2014). Depletion of both CTCF and cohesin deregulates expression of different groups of genes (Gosalia et al., 2014; Zuin et al., 2014).

Currently, the three-dimensional organisation of the nucleus in space (and time) is denoted the 4D nucleome (Tashiro and Lanctot, 2015).

4.1.2 Techniques to observe genome folding

Establishing the three-dimensional folding of the genome will help us understand genomic processes such as replication and transcription regulation. At low resolution (100 - 200 nm), the folding of individual loci can be observed by light microscopy with green-labelled DNA-binding proteins (Schermelleh et al., 2010). FISH allows the visualisation of multiple loci but the experimental procedure might affect the way chromatin folds (Dekker et al., 2002; Lichter et al., 1988; Pinkel et al., 1986; Schermelleh et al., 2010). Electron microscopy enables visualisation of nuclei in fine detail (~10 nm) but it lacks connectivity to the DNA sequences.

Hi-C, chromosome conformation capture followed by high-throughput sequencing, enables the detection of unbiased long-range interactions in a genome-wide fashion (Belton et al., 2012; Lieberman-Aiden et al., 2009; van Berkum et al., 2010). The technique follows 3C (section 1.7.3.) quite closely, but before ligation, the digested fragments are labelled with a biotinylated nucleotide that then marks ligation junctions (Fig. 1.14). Subsequently, the resulting fragments are subjected to massively parallel paired-end sequencing. The reads are mapped to the genome enabling the detection of fragment contacts and their abundance. The aligned paired-ends can then be used to create heatmap matrices of the intrachromosomal interactions based on the alignment position of each read (Fig. 4.1).

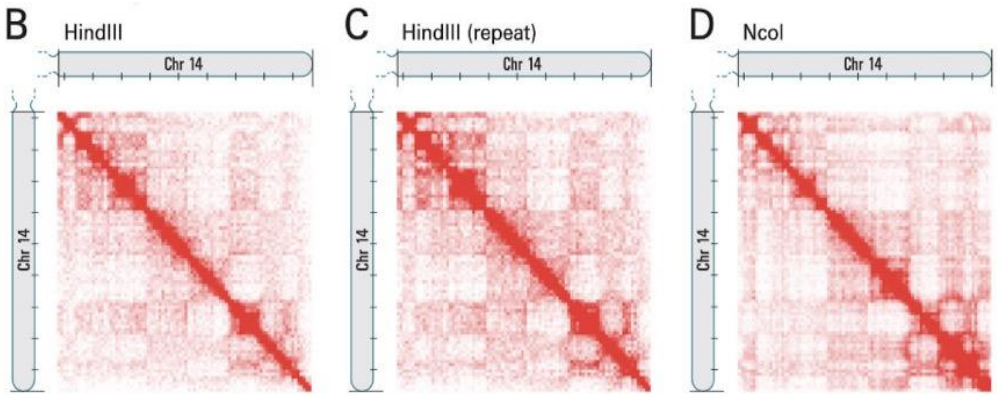

Figure 4.1 - Genome-wide contact matrices for chromosome 14 using HindIII and NcoI as restriction enzymes. Each pixel represents interactions between 1Mb loci. Dark red corresponds to higher number of interactions. The diagonals in all pictures represent the higher number of interactions between sequences in close proximity. From (Lieberman-Aiden et al., 2009). Reprinted with permission from AAAS.

The large square blocks that can be observed in the matrices can be interpreted as compartments of open and closed chromatin. Comparison with the distribution of genes, histone modifications indicative of transcribed gene bodies and DNase I Hypersensitive sites in the same genomic regions confirms the nature of the compartments (Fig. 4.2).

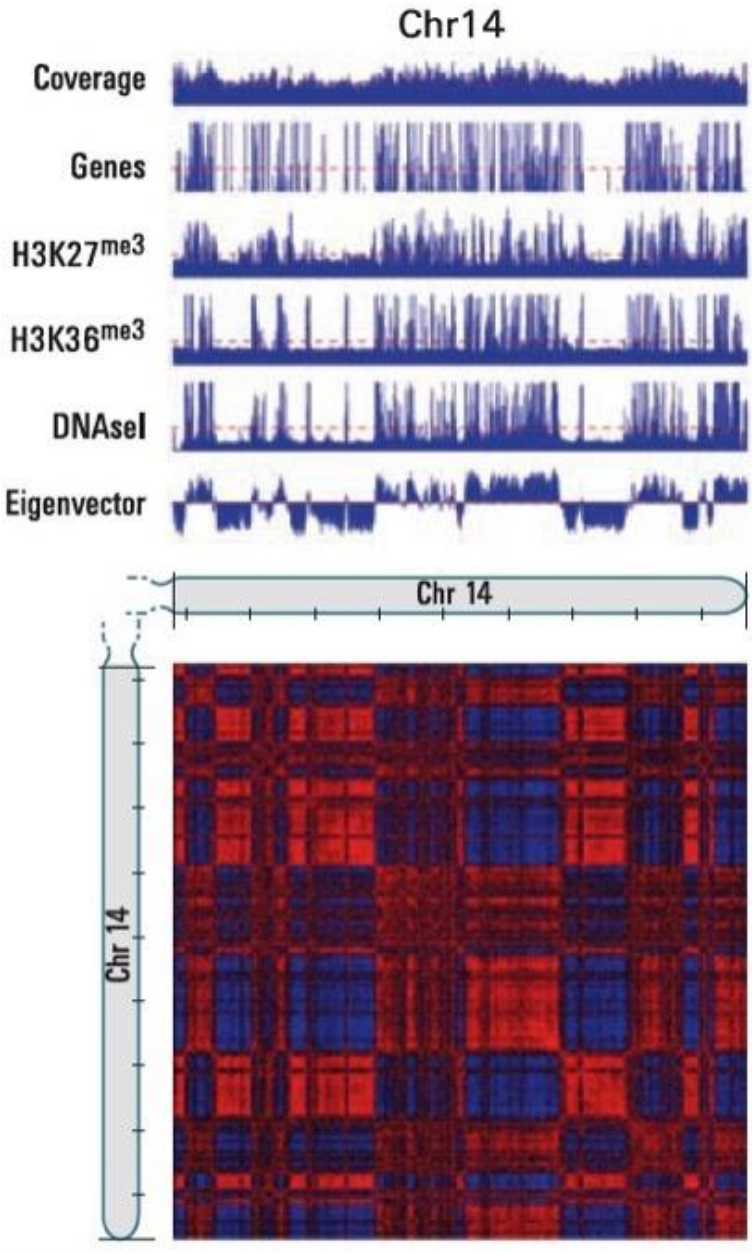

Figure 4.2 – Hi-C data shows the human nucleus is segregated into open and closed chromatin compartments. Map of chromosome 14 at 1 Mb resolution exhibits an intense diagonal of close-proximity sequences interacting and a constellation of large blocks. The plaid-pattern indicates the presence of two compartments within the chromosome and corresponds. These are compartments of open (red) and closed (blue) chromatin and correspond to the distribution of genes and with features of open chromatin.

4.1.3 Hi-C data sets

Hi-C paired-end reads from different human cells are publicly available on GEO (Barrett et al., 2013; Edgar et al., 2002). To study the 3D organisation of the DJ, numerous data sets with different human cell lines and restriction enzymes were analysed (Table 4.1).

Table 4.1 - List of Hi-C data sets employed to study the spatial organisation of the DJ

GEO accession number	Restriction enzyme	Cell lines	Study
GSE18199	HindIII (A'AGCTT), NcoI (C'CATGG)	GM12878, K562	(Lieberman-Aiden et al., 2009)
GSE37752	HindIII (A'AGCT)	RWPE1	(Rickman et al., 2012)
GSE41763	HindIII (A'AGCT)	HGPS fibroblasts	(McCord et al., 2013)
GSE43070	HindIII (A'AGCT)	IMR90, H1hesc	(Jin et al., 2013)
GSE44267	HindIII (A'AGCT)	HEK293	(Zuin et al., 2014)
GSE51687	HindIII (A'AGCT)	Breast cancer cells	(Mourad et al., 2014)
GSE63525	HindIII (A'AGCT), MboI ('GATC), NcoI (C'CATGG), DpnII ('GATC)	GM12878, HMEC, HUVEC, IMR90, K562, NHEK, KBM7	(Rao et al., 2014)
GSE56860	No RE, DNaseI	H1hesc, K562	(Ma et al., 2015)

Whereas most Hi-C studies explore the chromatin interactions of the human genome in untreated normal and cancer cells, the study from where the GSE43070 data set originated also treated cells with TNF- and flavopiridol (Jin et al., 2013). Interestingly, the authors observed that upon TNF-α signalling, TNF-α responsive enhancers were already in spatial proximity to their target promoters. This suggests that chromatin organisation is stable in a cell type and once established undergoes little change during signalling (Jin et al., 2013). Interactions between exons and their promoters (Mercer et al., 2013) also remain in place after treatment with flavopiridol. Flavopiridol stops transcription elongation by inactivating the elongation factor P-TEFb (Chao and Price, 2001) this blocks the action of RNA pol II. All this implies that contacts between promoters and gene bodies is independent of transcription. The GSE18199 set is the first Hi-C study and has a considerably lower number of reads (~100,000,000) compared to subsequent studies (at least 350,000,000).

The GSE41763 data set was developed for a study on the Hutchinson-Gilford Progeria syndrome (HGPS) (McCord et al., 2013). The study showed alterations genome-wide on loci associations with the nuclear lamina. The accumulation of a dominant lamin A protein, progerin, from a point mutation that causes the syndrome, leads to genomic disorganisation (McCord et al., 2013). Changes are also observed in the regulation of transcription and in H3K27me3 (indicative of transcription repression) marks in heterochromatin. The GSE63525 data set was produced to create a comprehensive three-dimensional interaction map of the human genome with 1 kb resolution (Rao et al., 2014). In this study, it was observed that the human genome is partitioned into regional domains, segregated into one of six subcompartments. These

subcompartments are associated with distinct patterns of histone marks. More than 10,000 conserved loops, linking promoters and enhancers were identified. The majority of loop anchors bind CTCF and occur at domain boundaries that separate active from inactive chromatin (Rao et al., 2014). An updated Hi-C protocol was used, which reduced the frequency of random ligations and enabled higher resolution through the usage of a 4-cutter restriction enzyme. Another experimental modification was the omission of the formaldehyde cross-linking step in some samples.

The use of restriction enzymes to fragment chromatin limits the resolution of three-dimensional maps due to their local distribution (Ma et al., 2015). Two DNA fragments need to have the same RE sequence (4 or 6 base cutter) to ascertain their contact probability with conventional Hi-C. DNase Hi-C achieves higher resolution in areas of open chromatin as it applies instead DNase I to randomly fraction DNA (Koohy et al., 2013). DNase Hi-C results were consistent with previous observations also depicting open and closed chromatin domains, higher frequency of intrachromosomal contacts and polymer-like structures (Koohy et al., 2013).

4.1.4 Distal Junction

The known distal junction (DJ) is a 380 kb sequence present in all human acrocentric chromosomes, adjacent to the rDNA repeats on the telomere side. It contains a large inverted repeat with 79.5% arm sequence identity and length ~109 kb and ~111 kb (Fig. 4.3).

Figure 4.3 - Location and arrangement of the large inverted repeat (white arrows) in the DJ contig (in green). The DJ orientation in this figure is from rDNA repeats, on the left, towards the telomere side, on the right. Sizes and positions of the arms of the inverted repeat are indicated in pairs. Figure from (Floutsakou et al., 2013).

On interphase cells, the DJ is located in the nucleolar periphery (Fig. 1.10). When cells are treated with actinomycin D (AMD), rDNA transcription is inhibited and rDNA retreats to the location of its corresponding DJ (Floutsakou et al., 2013). The DJ does not move towards rDNA, this suggests that the sequence composition and/or the spatial conformation of the distal junction is responsible for its localisation in the perinucleolar heterochromatin. The DJ also shows evidence for transcription activity and occurrence of histone modification (Floutsakou et al., 2013). Given that the folding of chromatin impacts on biological function, the DJ might possess a chromatin disposition that facilitates or controls the regulation of transcription.

In this chapter, I will describe the high-resolution analysis of publicly available human Hi-C data to identify chromatin interactions occurring within the distal junction (DJ). Interestingly the inverted repeat folds into a topological domain that is likely to have functional relevance. I will also describe the Hi-C analysis of other large inverted repeats present in the human genome that do not form a similar structure, suggesting that this large-scale structural feature in the DJ is unique in the human genome.

4.2 Results

4.2.1 Hi-C data quality control

Hi-C data sets were downloaded in SRA format and converted to paired-end FASTQ format. Trimmomatic was used to remove low quality bases and/or reads (section 2.8).

4.2.2 Hi-C data analysis for DJ

The Bowtie aligner (section 2.22) was used to align Hi-C reads, as single-end reads, to a custom genome comprised of GRCh37 and the sequences for DJ, PJ and an rDNA repeat, which was extracted from AL592188 (Fig. 4.4). The AL592188 BAC contains the last rDNA gene sequence before the DJ and is the most representative version to date (personal communication with Prof Brian McStay). Pairs where both mates mapped to the DJ were selected. Reads from each pair that mapped to the same restriction fragment (restriction maps generated according to enzyme used) were discarded. Given the small size of the DJ (380 kb) and that the majority of chromatin interactions occur between sequences in close proximity, pairs with reads that mapped within 5 kb of each other were also discarded. This was done for a better visualisation of long-range interactions within the DJ. The remaining pairs were used to create matrices representing DJ intramolecular interactions at 1 kb resolution. The matrices were

normalised by dividing the total number of interactions in each 1 kb block by the total number of reads that mapped to the DJ per set.

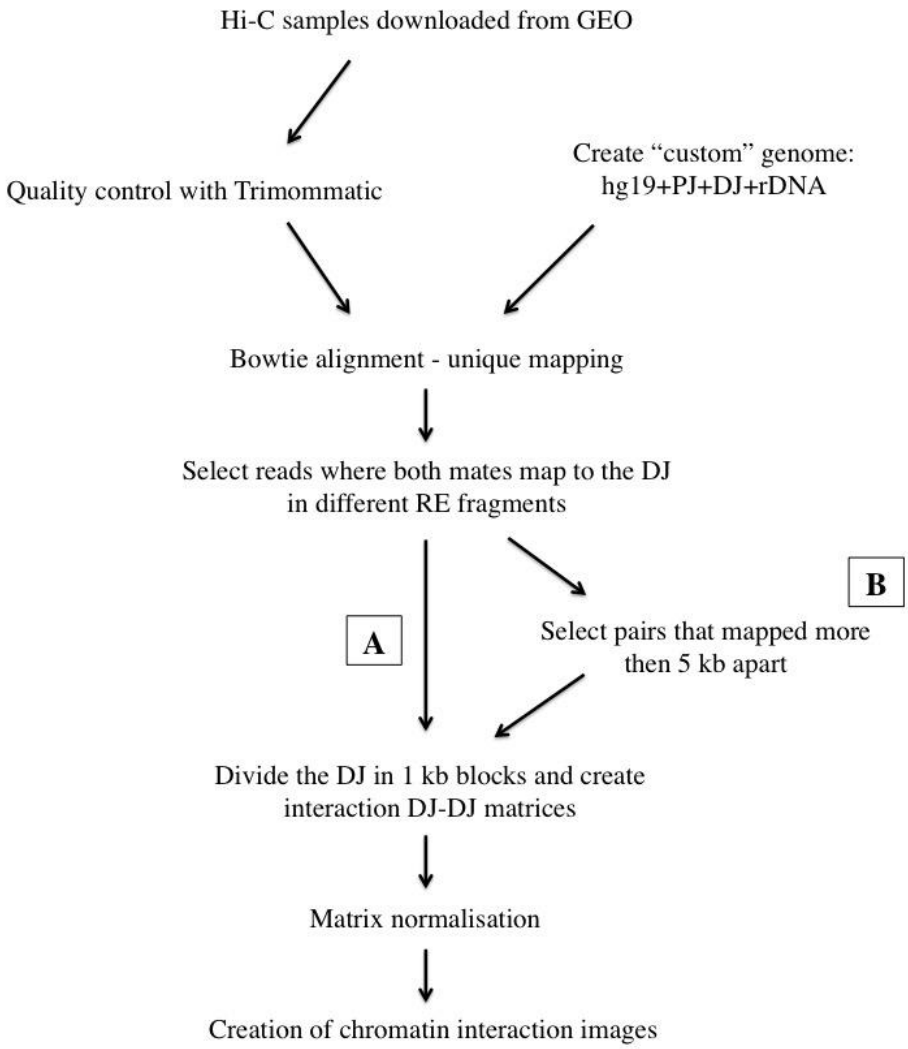

Figure 4.4 - Strategy to analyse the spatial conformation of long-range intramolecular interactions of the DJ using Hi-C sequencing reads. A – All read pairs that mapped to the DJ in different restriction fragments were used to create the interaction matrices. B – Read pairs that mapped to neighbouring regions (within 5 kb) were excluded from the interaction matrices.

4.2.3 DJ interaction maps

All data sets from table 4.1 were analysed. Although coverage of the DJ varied between studies, the same chromatin intramolecular interactions could be observed in all samples except in GSE18199. Results from studies GSE43070, GSE63525 and GSE56869 are described below. Results obtained from the other studies can be found in Appendix A (A1 – A5).

4.2.4 GSE43070

The GSE43070 data set was the first high-resolution genome-wide data available for the human genome. The high number of reads provided a 5 kb to 10 kb resolution and also allowed the identification of chromatin contacts over short distances. This study aimed to provide a high-resolution map of the interactions between enhancers and promoters in the human genome (Jin et al., 2013). Analysis of the GSE43070 data set for the distal junction was carried out as described above (Fig. 4.4-A). The first analysis was carried out with all paired-end reads that mapped to different interacting fragments. As with the full chromosome intramolecular matrices from this and other previous studies (Fig 4.1), a diagonal line could be observed in the DJ (Fig. 4.5). This stems from the higher number of interactions between sequences in close proximity. The plaid-like pattern characteristic of open and closed chromatin was not discernible. However, topological domains in genomes have been reported to be megabase-

sized (Dixon et al., 2012; Jin et al., 2013; Zhang et al., 2012b). The DJ is roughly 380 kb in length, possibly too small for this pattern to be observable.

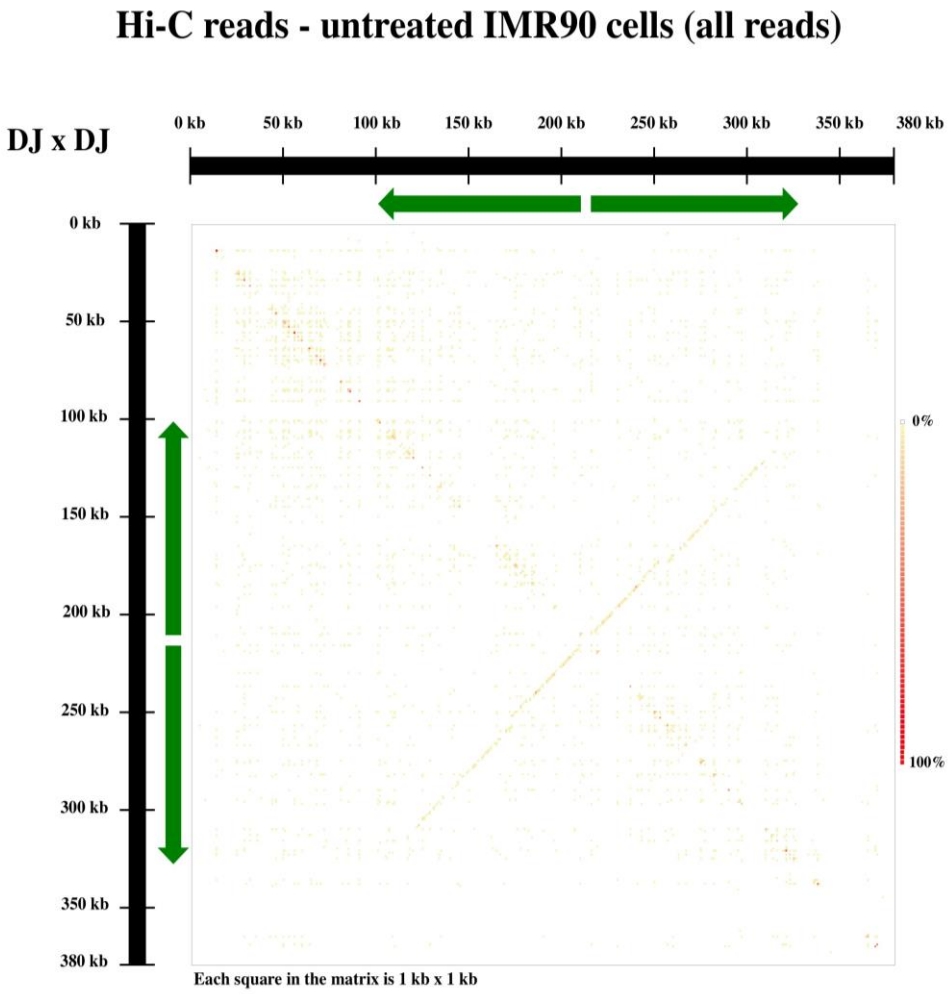

Figure 4.5 - Intrachromosomal interactions in the DJ captured by Hi-C data from IMR90 cells in normal conditions. All paired-end reads that mapped to the DJ in a unique manner and to different restriction fragments were used to construct this matrix. Although a clear diagonal line can be observed, DJ chromatin also folds into a tight loop shape at the large inverted repeats region. Scale from no observed interactions to highest observed number of interactions in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

Interestingly, another line, perpendicular to the expected diagonal line of interactions from closely placed regions could also be observed (Fig. 4.5). This line revealed that the arms of the large inverted repeat present in the DJ are in contact with each other. For better visualisation of this novel chromatin feature,

read pairs representing interactions from neighbouring regions (within 5 kb) were removed from all future analyses. The same untreated set and a replicate untreated sample were analysed using the strategy with the new spatial condition (Fig 4.4-B). Both samples showed the arms of the inverted repeat are in close spatial proximity to each other (Fig. 4.6 and 4.7).

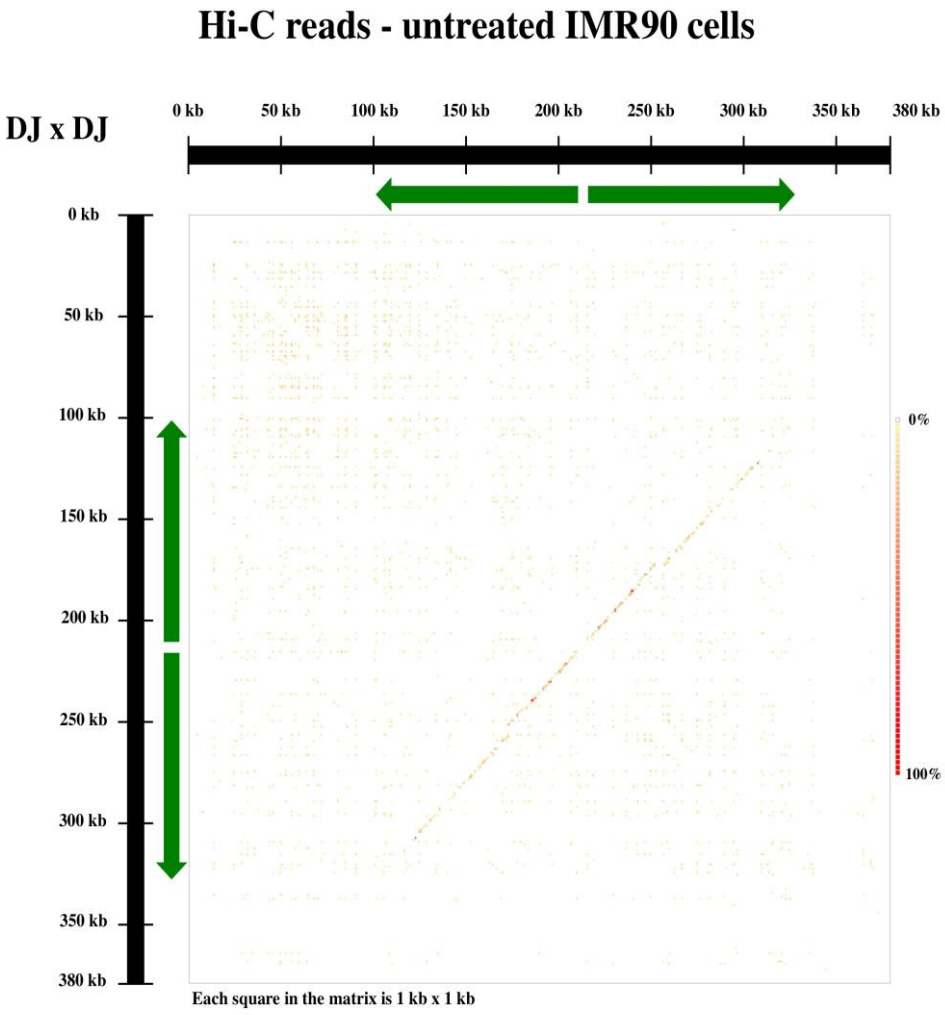

Figure 4.6 - Intramolecular interactions in the DJ using Hi-C reads from IMR90 cells in normal conditions. DJ chromatin folds into a tight loop shape at the large inverted repeats region. Scale from no observed interactions to highest observed number of interactions in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

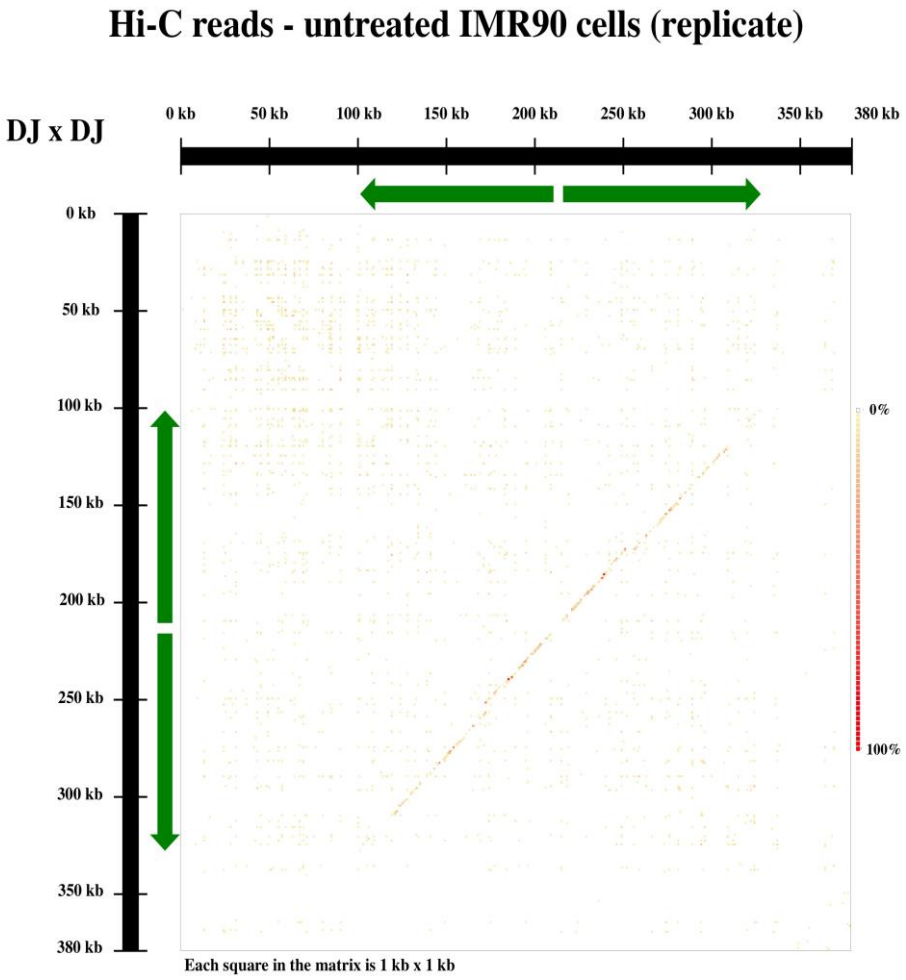

Figure 4.7 - Intramolecular interactions in the DJ using Hi-C reads from IMR90 cells in normal conditions (replicate sample). As with the previous IMR90 sample, the DJ chromatin folds into a tight loop shape at the large inverted repeats region. Scale from no observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

The inferred configuration is a tight loop structure, comprised of the two arms of the large inverted repeat folded at the spacer (Fig. 4.8). The number of interactions along the arms suggests a fold that produces a tower-like conformation.

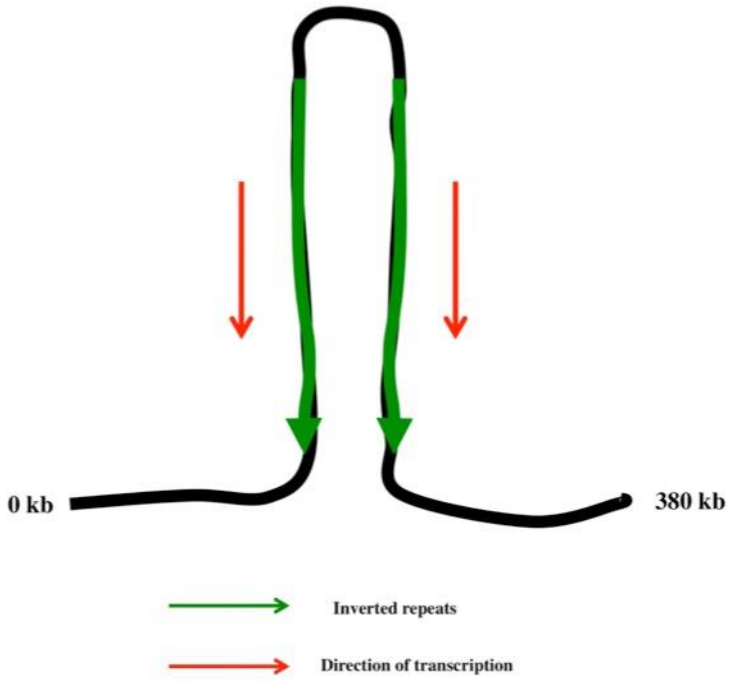

Figure 4.8 - DJ structural domain. Analysis of Hi-C reads revealed the presence of a topological domain in the DJ centred at the large inverted repeats.

Hi-C reads from IMR90 cells treated with Flavopiridol showed much fewer contacts (~66% decrease) between the inverted repeats (Fig. 4.9). Flavopiridol inhibits transcription by targeting and inactivating the positive elongation factor P-TEFb (Chao and Price, 2001). This stops RNA Polymerase II, leading to the loss of mRNA synthesis.

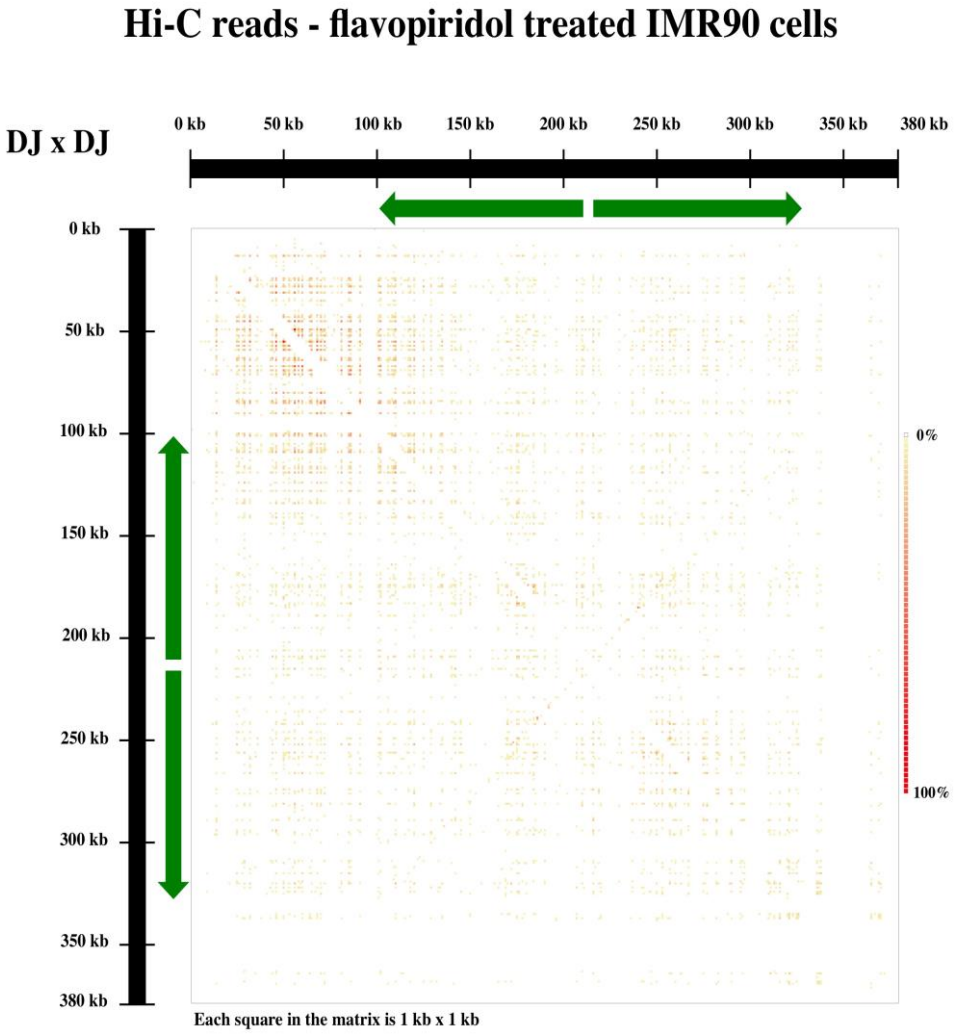

Figure 4.9 - Intramolecular interactions in the DJ after treatment of IMR90 cells upon flavopiridol treatment. Scale from no observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

4.2.5 GSE63525

The GSE63525 data set was produced to create a comprehensive three-dimensional interaction map of the human genome with 1 kb resolution (Rao et al., 2014). The DJ intrachromosomal domain was observed in data sets generated with the original dilution protocol (Fig. 4.10) and the Hi-C *in situ* protocol (Fig.

4.11). The in situ protocol sample showed 86% more contacts than the dilution sample. Another experimental modification was the omission of the formaldehyde cross-linking step. The DJ domain was also observed in this sample (Fig. 4.12). The frequency of contacts between the large inverted repeats had approximately the same number of contacts in the *in situ* protocol without the cross-linking step (3% more) as the set from the dilution protocol.

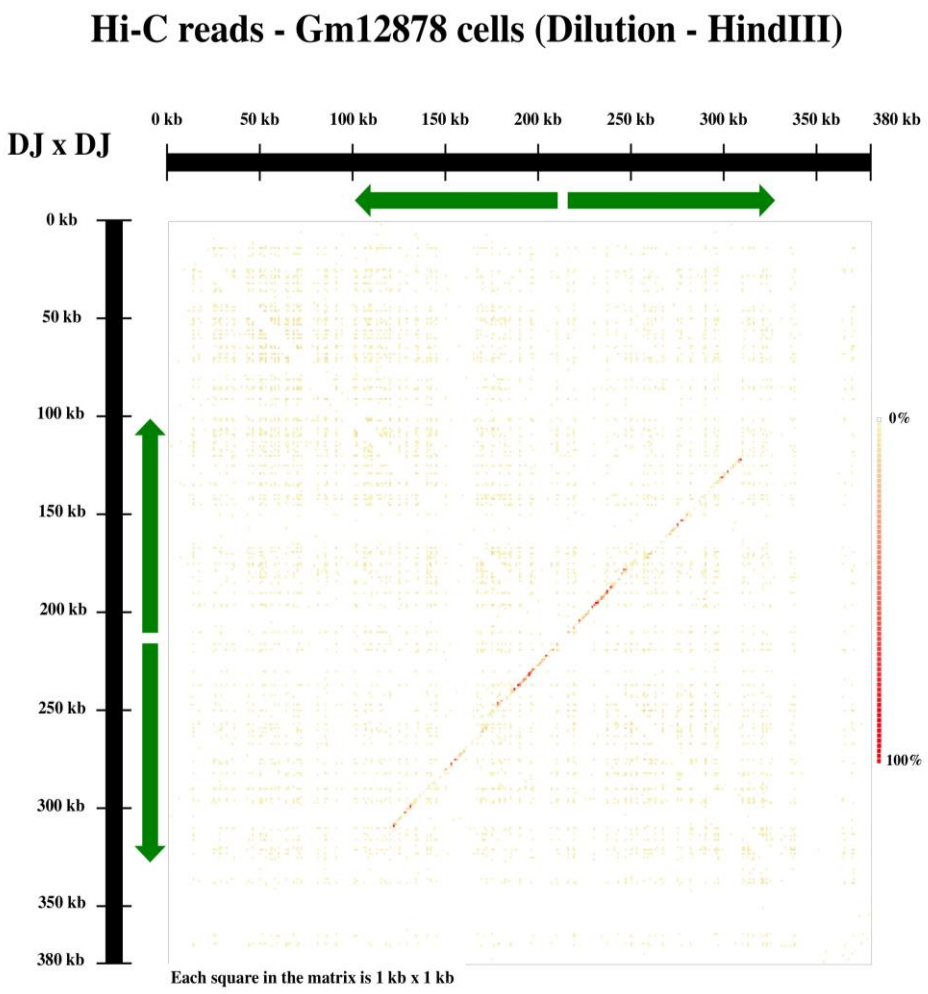

Figure 4.10 - Intrachromosomal interactions in the DJ obtained through analysis of Hi-C reads from the GSE63525 study. This sample used HindIII (A'AGCT) and original dilution protocol in Gm12878 cells. Green arrows indicate the position of the large inverted repeats. Scale from no observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

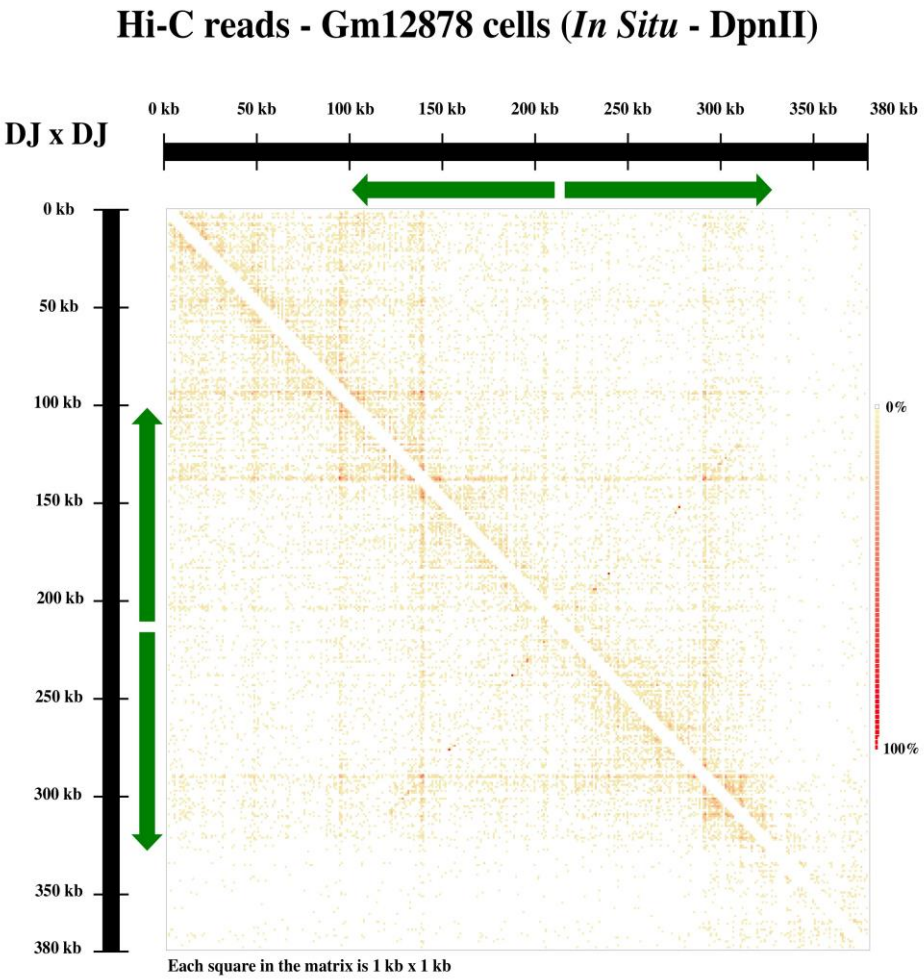

Figure 4.11 - Intrachromosomal interactions in the DJ obtained through analysis of Hi-C reads from the GSE63525 study. This sample used DpnII ('GATC) and *in situ* protocol in Gm12878 cells. Green arrows indicate the position of the large inverted repeats. Scale from no observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

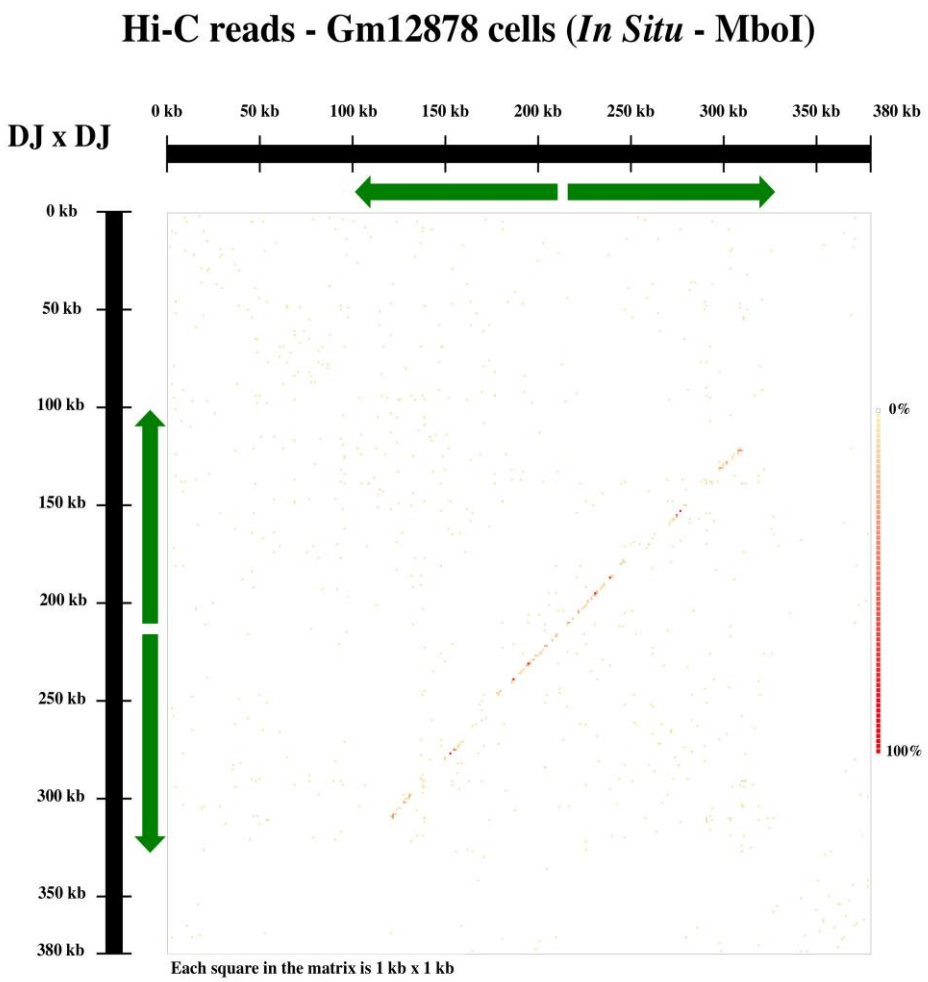

Figure 4.12 - Intrachromosomal interactions in the DJ with Hi-C reads from analysis of Hi-C reads from the GSE63525 study. This sample used MboI ('GATC) and *in situ* protocol with gentle handling and no cross-linking step in Gm12878 cells. Green arrows indicate the position of the large inverted repeats. Scale from no observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

4.2.6 GSE56869

Fragmenting chromatin with restriction enzymes can reduce the resolution of three-dimensional maps due to their local distribution (Ma et al., 2015). Instead, DNase I can be used to cut chromatin, which creates random

fragmentation in open chromatin (Koohy et al., 2013). DNase Hi-C results were consistent with previous observations such as open and closed chromatin domains, higher frequency of intrachromosomal contacts and polymer-like structures. A higher number of contacts (55% more) in the inverted repeat could be observed (Fig. 4.13) when compared to a data set cut with a restriction enzyme (GSE43070, HindIII, Fig. 4.6).

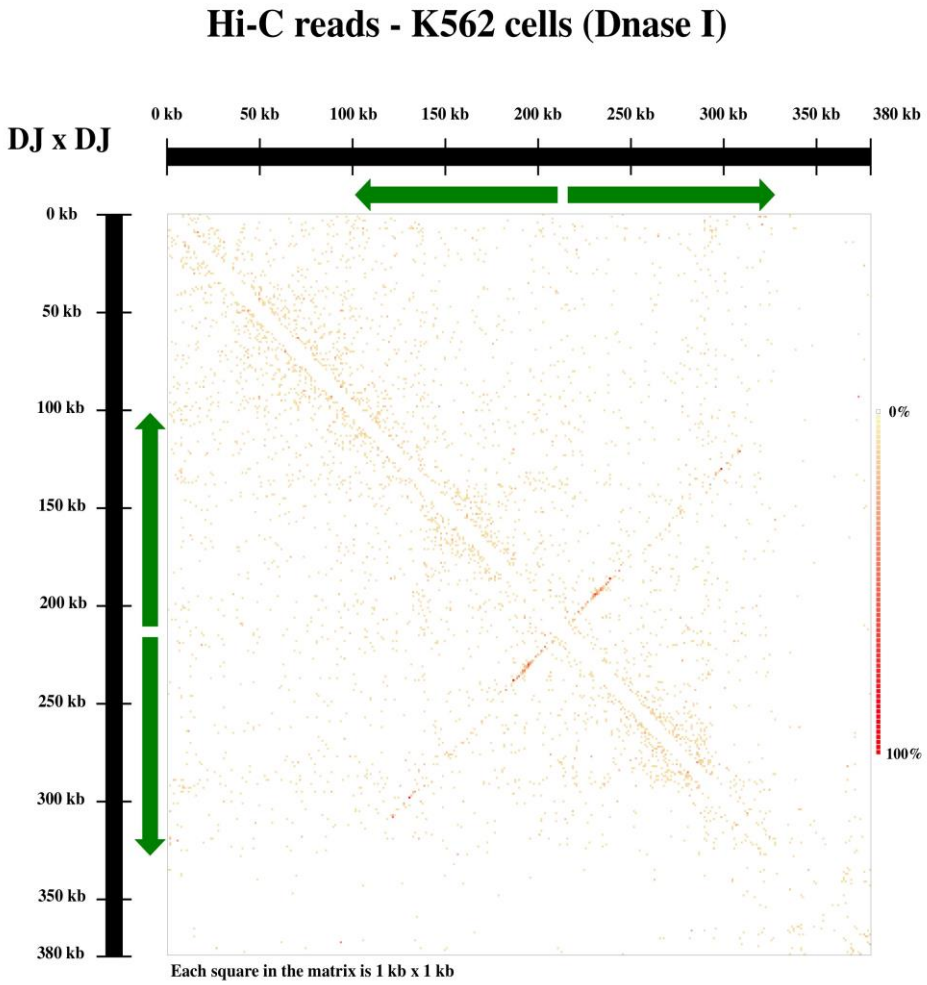

Figure 4.13 - Observation of intramolecular interactions between the DJ large inverted repeats in DNase Hi-C reads from K562 cells. Green arrows indicate the position of the large inverted repeats. Scale from no observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

4.2.7 Analysis of other large inverted repeats in the human genome

The large inverted repeat present in the DJ has ~79.5% identity, arm lengths of 109 kb and 111 kb and is centred around a 6 kb spacer. The data set treated with flavopiridol suggests that this chromatin feature is not an artefact of the analysis of Hi-C data. To ascertain this, and to check if a similar structure occurs, the GSE43070 Hi-C data set was used to analyse other large inverted repeats (>75% identity, arms lengths > 8kb and spacer length up to 100 kb) present in the assembled human reference genome (Warburton et al., 2004). None of the analysed repeats had a similar structure to the DJ domain (Fig. 4.14 – two of the repeats with similar percentage identity to DJ. Results obtained for the remainder inverted repeats can be found in Appendix A Figure A6). The inverted repeats analysed in Fig. 4.14 showed no contacts at all between its arms. Analysis was carried out using the hg17 (GRCh35) augmented with the sequences for PJ, DJ, rDNA and AL591856 (for this last BAC pleaser refer to the next chapter of this thesis). A region of 380 kb centred at the inverted repeats spacer was used to mimic the size of the DJ.

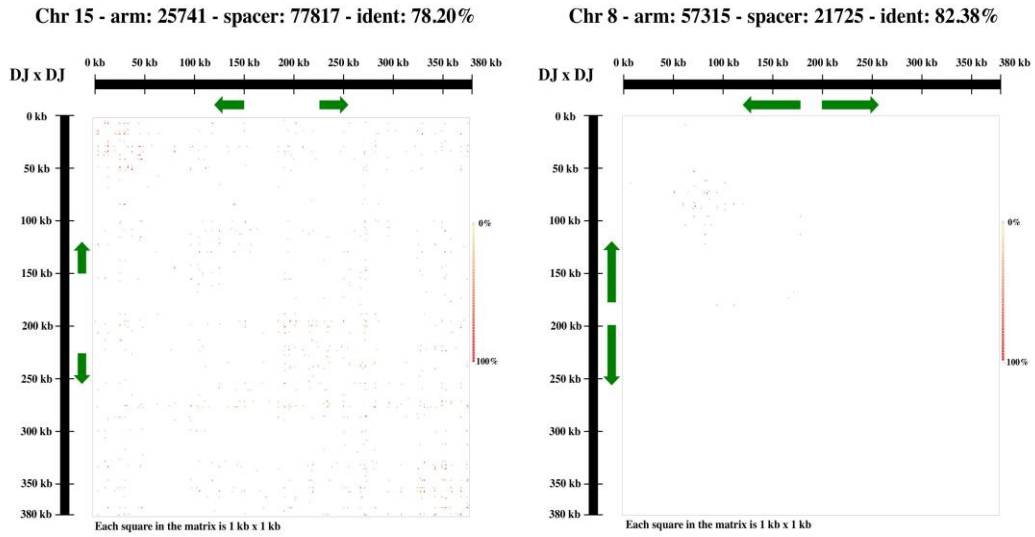

Figure 4.14 - Analysis of intrachromosomal interactions in two large inverted repeats present in the human genome. Arm and length sizes are presented in bp. Scale from no observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

4.3 Discussion

The DJ, present in all short arms of acrocentric chromosomes, is located on the nucleolar periphery and anchors the rDNA repeats. When Pol I transcription is inhibited, rDNA clusters move to the edge of the nucleolus forming caps adjacent to their corresponding DJ. This occurrence is possibly aided by the spatial organisation of the DJ. To investigate this, publicly available Hi-C data was analysed using a compound genome of GRCh37 and sequences for DJ, PJ and rDNA gene repeat.

Hi-C data revealed the presence of a chromatin feature centred at the DJ large inverted repeats (Fig. 4.15). All Hi-C data sets analysed revealed the presence of this chromatin feature, except GSE18199 (Jin et al., 2013;

Lieberman-Aiden et al., 2009; Ma et al., 2015; McCord et al., 2013; Mourad et al., 2014; Rao et al., 2014; Rickman et al., 2012; Zuin et al., 2014). However, this could be due to the lower number of reads of this study. The DJ domain was clearly visible in all other studies, including a sample where cells were immobilised in agar but the chromatin cross-linking step was not employed (Fig. 4.12).

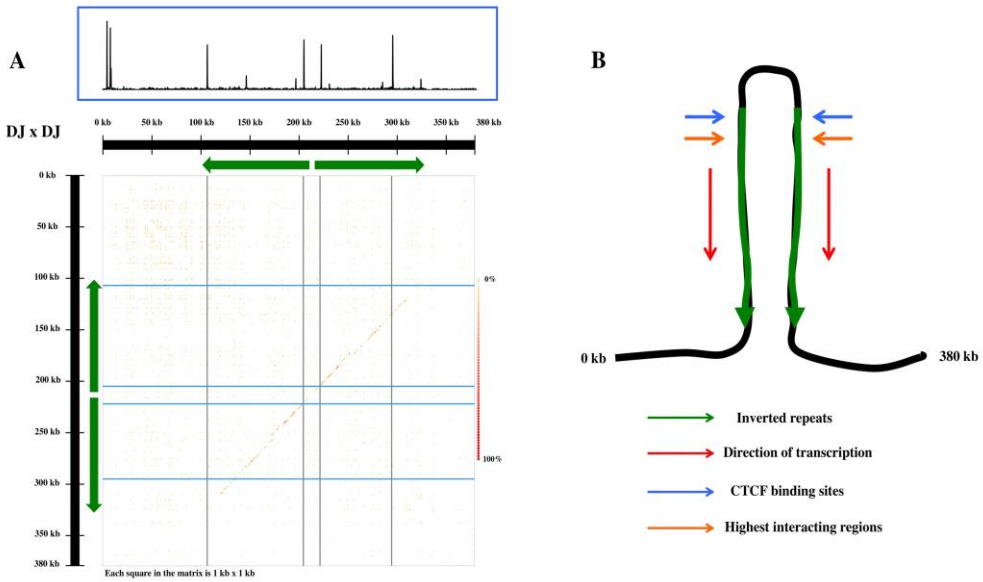

Figure 4.15 – The DJ chromatin intramolecular contacts form a loop structure centred at the large inverted repeat. A- CTCF binding sites from ChIP-seq data (top panel) overlap with the large inverted repeat at the end and beginning of left and right arm respectively. Position of the large inverted repeat in the DJ is shown in green.

CTCF binding sites were previously identified in the DJ (Floutsakou et al., 2013). CTCF and Cohesin could be responsible for maintaining the structure (Gosalia et al., 2014; Zuin et al., 2014). The highest interacting regions of the inverted repeats do not overlap with the CTCF peaks. However, DJ regions with CTCF binding sites are clearly in contact in all samples. Perhaps CTCF has an alternative function in the DJ. Besides regulating chromatin architecture, CTCF also acts as an insulator, limiting the interaction between promoters and

enhancers (Cuddapah et al., 2009; Wendt et al., 2008). These types of regulatory elements have also been characterised in the DJ (Floutsakou et al., 2013). CTCF binding sites are also more common in boundary domains of open and closed chromatin where they prevent the spread of heterochromatin (Gosalia et al., 2014; Zuin et al., 2014). The position of CTCF within the DJ inverted repeats seems to also follow this premise. The histone modification H3K36me3 is indicative of transcribed gene bodies, defining actively transcribing chromatin (Ernst et al., 2011). CTCF peaks in the DJ are found outside these areas (Fig. 4.16). It is possible that CTCF and cohesin are responsible for creating and maintaining the loop structure by clamping the inverted repeat loop at the top (Fig. 4.15 - B). In this case, ChIP-seq data for cohesion should be analysed for the DJ. The CTCF sites at the bottom of this domain do not overlap in this conformation. This differs from loops reported in the human genome where CTCF is responsible for maintaining the loop conformation (Guo et al., 2015). Importantly, looped chromatin also loses its configuration when inversion of the CTCF binding sites in enhancer regions is carried out with CRISPR (Guo et al., 2015). This sheds some light on how local architecture can be encoded by the linear sequences of DNA.

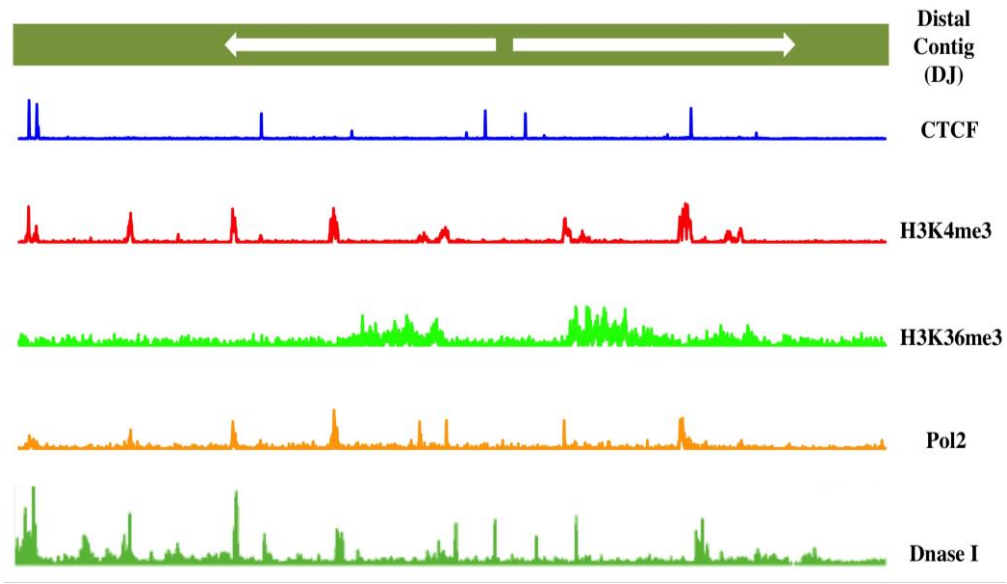

Figure 4.16 - ChIP-seq peaks or CTCF, H3K4me3, H3K36me3 and Pol II and DNase-seq in the DJ. Data Analysis from (Floutsakou et al., 2013).

The structural domain created by this loop places transcription of the DJ long non-coding RNAs occurring in the same direction. The frequency of contacts between the large inverted repeats that form this structure decreased considerably when Pol II transcription was inhibited by flavopiridol (Fig. 4.9). The DJ seems to possess plasticity of spatial configuration. This contrasts with the reported behaviour of the rest of the genome where the three-dimensional chromatin landscape is established for a cell type specific manner and stable (Jin et al., 2013). In general, enhancers and their target promoters are already in contact regardless of activated or repressed signalling and inhibition of transcription (Jin et al., 2013). Therefore, this topological domain might be formed as a cause or consequence of transcription in the distal junction. Whether it forms to facilitate transcription or it also happens as a structural role to help shape the nucleolus if the DJ transcripts contribute to nucleolar formation/maintenance should be further pursued. It is important to note, however, that the DJ supports a certain degree of intra and interchromosomal

segmental duplications and that this observed intramolecular organisation is an average of 10 DJs per cell. This has implications on the number of real interactions that are lost due to the unique read mapping approach. It is also possible that not all DJ have this configuration at the same time throughout the cell cycle. If indeed this structure is needed for transcription to occur or for nucleolar structural purposes, then the DJs from inactive short arms of acrocentric chromosomes that are not part of nucleoli might not have this conformation. The structure reported by Hi-C data is an average of the spatial configuration of all DJs in a cell.

The perpendicular line observed in the DJ Hi-C matrices is not continuous (Fig. 4.6, 4.7, 4.10, 4.12 and 4.13). Around [250, 260] kb and [160, 170] kb on both sides of the inverted repeats there is a small deviation from the perpendicular line. This probably means that there is another layer of spatial organisation in the DJ that is not entirely visible in Hi-C data and will require a different technique with higher resolution. These regions do not overlap with ChIP-seq peaks reported in (Floutsakou et al., 2013) and therefore their spatial organisation might be relevant to the function of the DJ domain.

Importantly, analysis of intrachromosomal interactions in other large inverted repeats in the human genome (Warburton et al., 2004) revealed the DJ domain to be a unique feature in the human chromosomes. This reinforces the possibility that it contributes to nucleolar maintenance. To further validate the existence of the spatial conformation of the DJ experimentally, chromosome conformation capture should be carried out. Functional analysis of the DJ long non-coding RNAs and their connection to the DJ conformation and vice-versa should also be pursued.

5 Extension and characterisation of sequences along the distal side of acrocentric short arms

5.1 Background

5.1.1 Nucleolar Organiser Regions

The entire short arms of the five human acrocentric chromosomes (13, 14, 15, 21 and 22) are missing from the current reference genome (GRCh38). NORs, comprised of rDNA arrays and located in the p-arms of acrocentric chromosomes, control the assembly and regulation of nucleoli (McStay and Grummt, 2008; Pederson, 2011). The sequences surrounding the rDNA repeats were also shown to contribute to the formation and maintenance of the nucleolus (Floutsakou et al., 2013). Therefore currently, the definition of an NOR includes the rDNA genes and the known adjacent sequences to rDNA. Importantly, the NORs are not only shared and highly conserved across all acrocentric chromosomes but also possess a complex sequence composition (Floutsakou et al., 2013). The known proximal side of NORs (Fig. 5.1, in orange) is almost entirely segmentally duplicated (92.4%), for the most part to peri-/centromeric regions, with long segments (ranging from 1 kb to over 100 kb and more than 85% identity) that occur frequently in the rest of the genome and possibly at other sites on acrocentric p-arms.

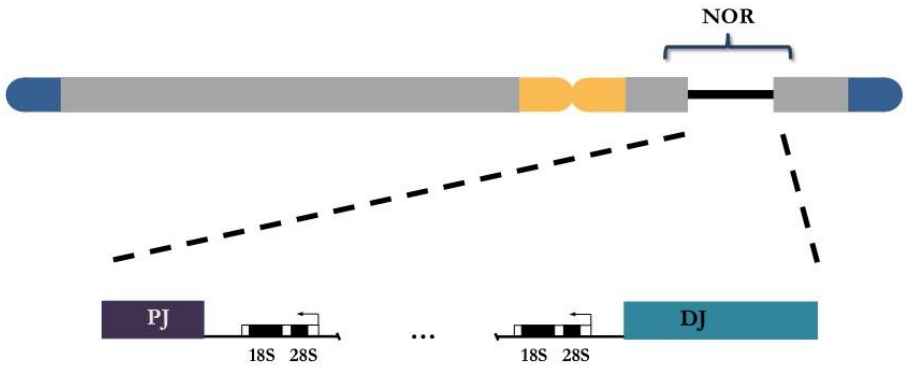

Figure 5.1 - Schematic of an acrocentric chromosome. Telomeres are shown in blue and the centromere in yellow. The NOR, located on the short arm, is comprised of tandem rDNA repeats flanked by the proximal junction towards the centromere and the distal junction towards the telomere.

The known distal region (DJ), towards the telomere (Fig 5.1, in blue), has a low degree of segmental duplication (7.3%), mainly to euchromatic and telomeric regions (restricted to no more than 5 kb segments with more than 85% identity), and shows evidence of functionality (Floutsakou et al., 2013). The DJ displays distinct transcriptional activity and chromatin organisation, such as promoter markers and CTCF binding sites (ChIP-seq data). Also, RT-PCR and RNA-seq data confirmed the existence of spliced polyadenylated transcripts originating from the majority of the reported promoters. The polyadenylated transcripts strengthen the evidence for transcriptional activity by RNA Polymerase II (Hirose and Manley, 1998; McCracken et al., 1997). These transcripts may function as long non-coding RNAs. In active nucleoli, the DJ is located in the perinucleolar heterochromatin anchoring the transcribed rDNA arrays. When transcription by Pol I is inhibited, the rDNA withdraws from the

nucleolus to nucleolar caps that form adjacent to its corresponding DJ (Floutsakou et al., 2013; Schofer et al., 1996). The DJ of all acrocentric chromosomes also contains a 38.6 kb block of CER repeats (Fig. 1.9-A in dark blue). CER is a 48 bp satellite repeat also localised near the centromeres of chromosomes 14 and 22 (Jurka et al., 2005). Contrary to the distal junction however, the proximal junction (Fig 5.1, in dark purple) is unlikely to hold functional elements relevant to nucleoli activity due to its high degree of chromosomal duplication. This characteristic renders unique mapping of sequencing reads to the PJ difficult to accomplish. For this reason, efforts to expand the short arms were focused on the distal side of rDNA repeats, towards the telomere. It is likely that the DJ contains regulatory elements relevant to the maintenance and formation of the nucleolus. However, there is limited sequence information beyond the distal junction. To expand it, there is a need to differentiate if candidate sequences are located in acrocentric chromosomes. Furthermore, it is important to distinguish between the individual acrocentric chromosomes as well as sequences from the proximal or distal side. This can be achieved with monochromosomal cell hybrids.

5.1.2 Monochromosomal hybrids for human chromosomes 13, 14, 15, 21 and 22

Monochromosomal somatic hybrids are rodent/human cells lines that contain a single human chromosome (Inoue et al., 2001). These cells were created by transferring single human chromosomes into A9 mouse cells through microcell fusion (Cuthbert et al., 1995; Tanabe et al., 2000; Warburton et al., 1990). Monochromosomal cell hybrids were originally created for gene studies and gene mapping (Athwal et al., 1985; Warburton et al., 1990), and for this project, were useful for assessing sequence information in the individual acrocentric chromosomes. In order to discern between sequences on the PJ or DJ side of the short arms, cells with individual chromosomes (GM09142 and GM10063) containing X and 21 reciprocal translocation products (Fig. 5.2) from Coriell Cell Repositories were used.

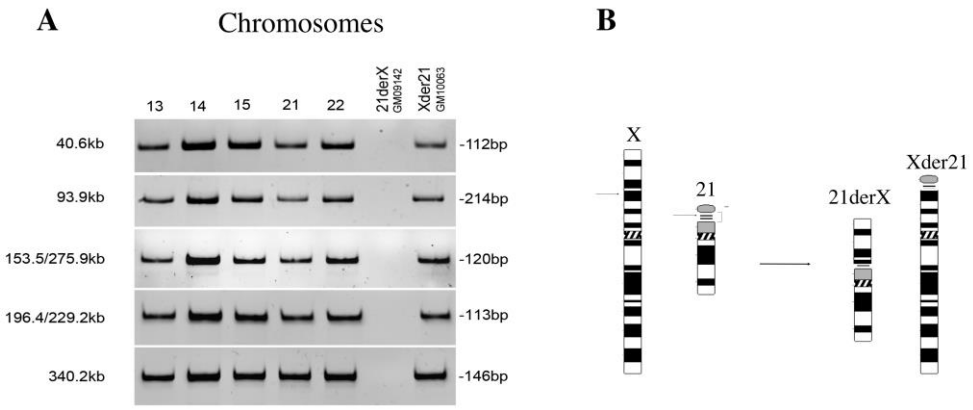

Figure 5.2 - PCR with monochromosomal hybrids searching for regions of the DJ. Panel A – PCR on monochromosomal cells with primer pairs located on the DJ in the areas indicated on the left and product lengths on the right. Panel B - Schematic of the reciprocal translocation between chromosomes X and 21. The break occurs in the rDNA on chromosome 21 and in a region of the short arm of chromosome X. The PJ will be located in 21derX and the DJ in Xder21. 21derX does not contain sequences from the distal side of the short arm of chromosome 21. Figure from Prof Brian McStay.

The derivative chromosome 21derX has the proximal side (including the PJ sequence) of chromosome 21 but not the distal side, as it was swapped for a

portion of the short arm of chromosome X. On the other hand, Xder21 contains the distal side (including the DJ sequence) of the short arm of chromosome 21 and is missing the proximal side, having instead the long arm and a portion of the short arm of chromosome X (Fig. 5.2, panel B).

In this chapter I describe how I extended the known distal region. This was achieved with bioinformatics approach using large assembled contigs (454 sequencing) from nucleolar DNA and Hi-C sequencing data. I will also describe experiments that confirm the placement of the BACs in the short arms of acrocentric chromosomes. I will describe the chromatin and transcription analysis performed on the new identified BAC to assess its chromatin landscape.

5.2 Results

5.2.1 Search for BACs from the short arms of acrocentric chromosomes

The unknown distal region of the acrocentric short arms starts 380 kb after the last rDNA repeat and continues towards the telomere. The known 380 kb is the DJ, which localises to the nucleolar periphery. Previously, we sequenced DNA extracted from purified nucleoli using 454 technology and Roche/454 delivered a first pass attempt at *de novo* assembly of the nucleolar reads. Given that the 454 assembled contigs originated from nucleolar DNA and

that these did not contain rDNA sequences, they were potentially a good representation of the missing regions of acrocentric chromosomes. Also, the majority of chromatin interactions are intrachromosomal and occur between sequences that are located within a few megabases (Lieberman-Aiden et al., 2009). These interactions can be studied and identified by chromosome conformation capture followed by high-throughput sequencing (Hi-C) (For detailed description please see Thesis introduction and chapter 4). Therefore, unplaced Hi-C reads with DJ mates can be used to single out contigs, from the 454 nucleolar data, that might extend the short arms of acrocentric chromosomes (Fig. 5.3).

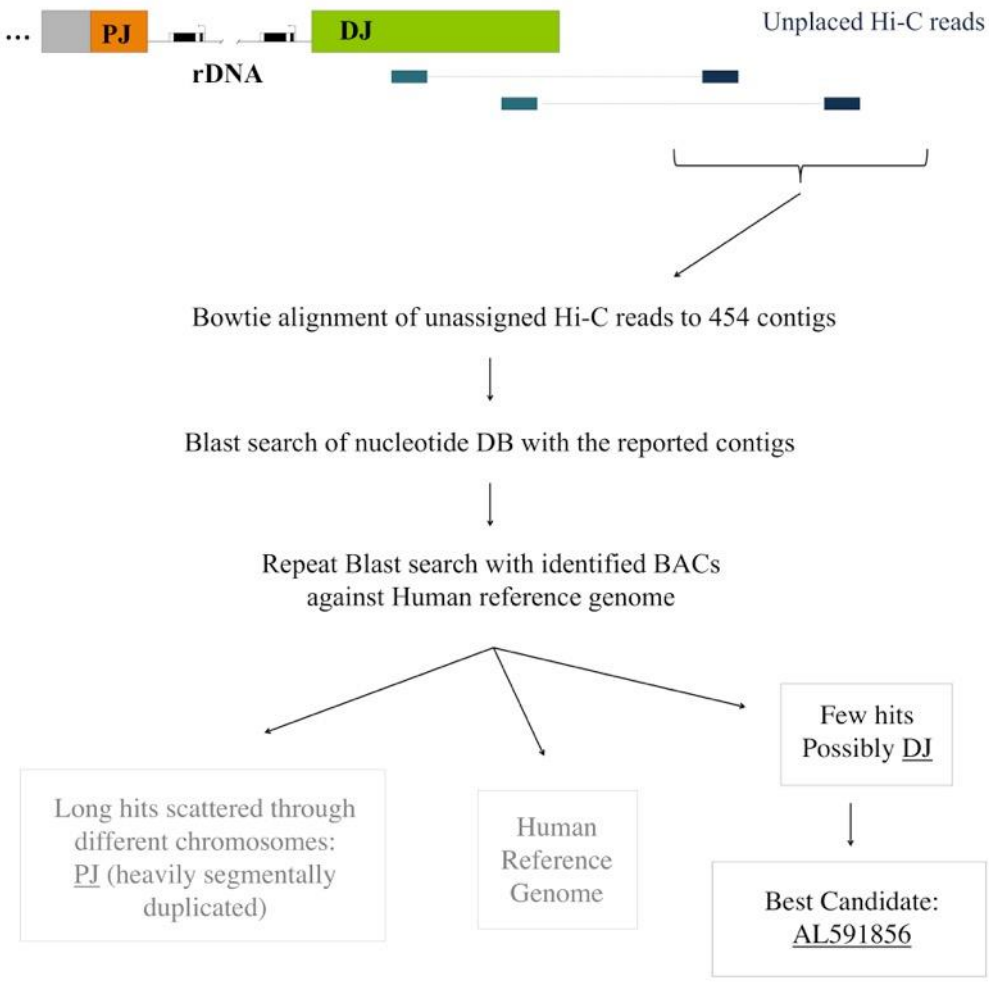

Figure 5.3 - Strategy to identify novel sequences in the short arms of acrocentric chromosomes using Hi-C sequencing data. Hi-C identifies chromatin interactions between closely located regions. Reads that do not map to the human reference genome or to any other known sequence (rDNA, PJ or DJ) but with mates that mapped to the DJ can be used to search for BACs.

The unassigned Hi-C reads were mapped against the nucleolar contigs using bowtie (section 2.22). The outputted contigs were used to search the human sequences in the nucleotide (nt) database from NCBI with BLAST (section 2.16). Many BAC sequences were reported and to further refine this search, a second BLAST search was carried out against the human genome reference (GRCh37) to assess their localisation. Any BACs that localised to the current human genome reference or to many chromosomes were discarded as

belonging to either the human reference genome (ie. non short arms of p-chromosomes) or the proximal junction due to the high segmental duplication of that region (Fig. 35). Numerous BACs were reported but the 3 most promising ones were AC013640, AC103988.7 and AL591856.

5.2.2 Primer design and PCR on monochromosomal hybrids

Of all the analysed BACs, AC013640, AC103988.7 and AL591856 had the fewest hits to the reference genome (GRCh37). PCR was performed (section 2.17) on DNA from monochromosomal hybrids to assess and confirm their position on the human chromosomes. Primer pairs (table 5.1, 5.2 and 5.3) were designed after using RepeatMasker (Smit, 2013-2015) to remove and avoid interspersed repeats and low-complexity DNA sequences that could lead to erroneous products.

Table 5.1 - Primer pairs for BAC AC013640 and expected product lengths

Primer	Primer sequence	Product size
ac013640_1f	TGGGCCCGTGCTTATTTTTATG	-
ac013640_1r	TAAGCACTTAGCTGCCTACTGAA	224 bp
ac013640_2f	TTGGGGTCACATTACTGCCC	-
ac013640_2r	TGTCGCCGAACCAATCTCAA	323 bp
ac013640_3f	CCAGGTTCCTGGCCTTCTTT	-
ac013640_3r	TTTTCGCTGCCTTCACAAGC	479 bp

Table 5.2 - Primer pairs for BAC AC1039887.7 and expected product lengths

Primer	Primer sequence	Product size
ac103988.7_1f	GGAGCTCTTGCCTGCCTAAT	-
ac103988.7_1r	ATGTTTAGCGTTCCTAACACGA	439 bp
ac103988.7_2f	ATTTGGGAGGGGTGGGGAATTATTA	-
ac103988.7_2r	AGCCACACAGTTAGATGCTGTTA	227 bp
ac103988.7_3f	TTGCCACTTTGTGAACGGCT	-
ac103988.7_3r	AGCTGGAGTTCGGTGAGAGA	377 bp

Table 5.3 - Primer pairs for BAC AL591856 and expected product lengths

Primer	Primer sequences	Product size
Primer_105_F	GTAGTGAAATAGTTAATGCAGCTGA	-
Primer_105_R	CCCACTGAAGATGTACCCACAA	212 bp
Primer_102_F	TGGCCGGCATTCGAATTTTCCC	-
Primer_102_R	CCGTCCTGTGCTGGTGACGT	211 bp
Primer_84_F	CACTGCTGCTTGAAGGGCAACA	-
Primer_84_R	GCTCCTCTGCCGTGGGTACTG	174 bp

Gel electrophoresis (see section 2.5) was performed on monochromosomal hybrids containing the 5 human acrocentric chromosomes and the X/21 reciprocal translocation to depict the PJ and DJ side of the short arms. For AC013640, all primer pairs indicate this BAC is located in chromosome 15 (Fig. 5.4). Only primer pair 2 created a product for the AC1039887.7 BAC, also located in chromosome 15 (Fig. 5.5).

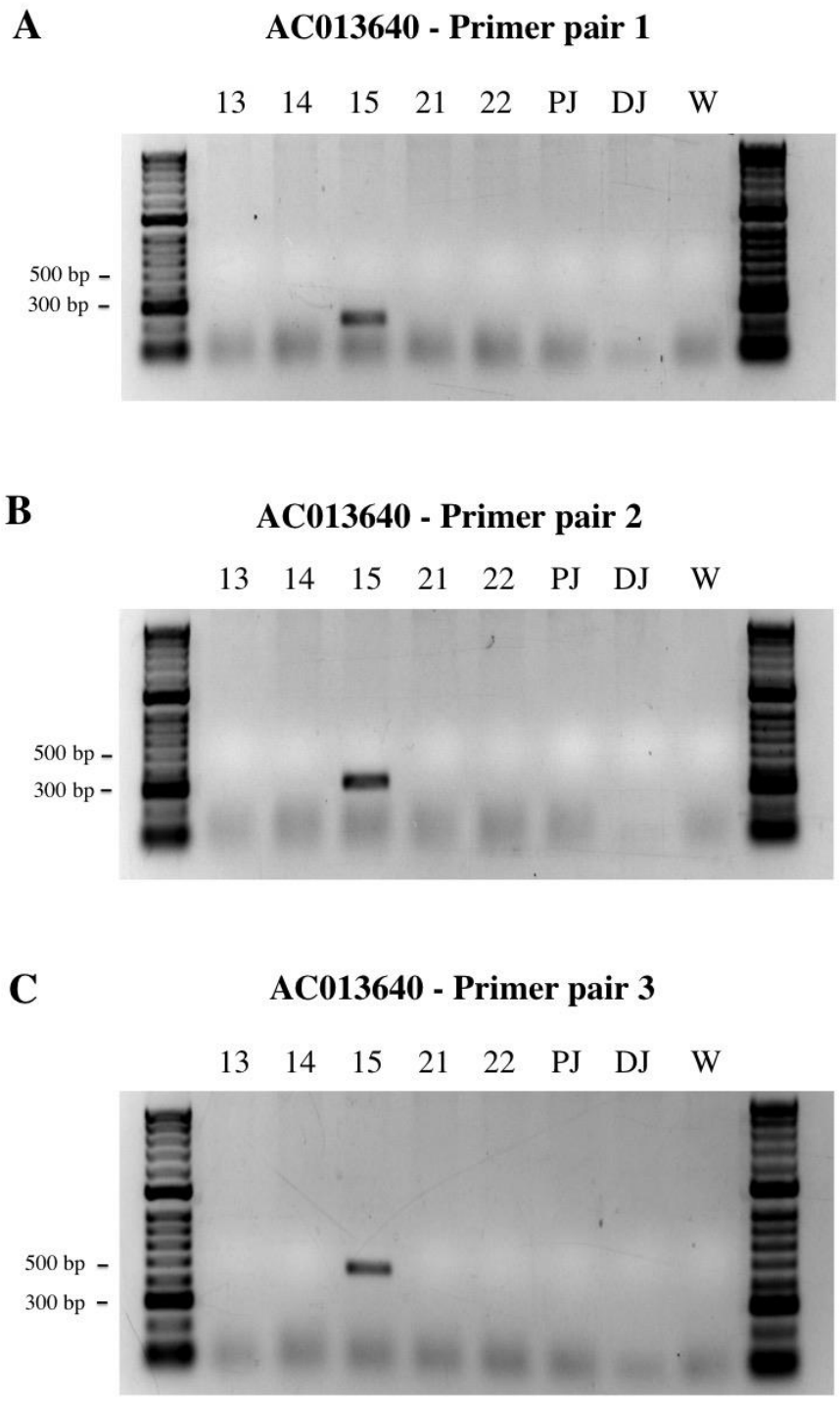

Figure 5.4 - Gel electrophoresis of PCR product from AC013640 using DNA from monochromosomal somatic cell hybrids (mouse/human) as template. Hyperladder II was used. Lanes show hybrids for human chromosomes 13, 14, 15, 21, 22, 21derX (proximal side of short arm of chromosome 21), Xder21 (distal side of short arm of chromosome 21) and water control. A – Primer pair 1. BAC is placed on chromosome 15 with a 224 bp product. B – Primer pair 2. BAC is located on chromosome 15 with a 323 bp length product. C – Primer pair 3. BAC is located on chromosome 15 with a 479 bp length product.

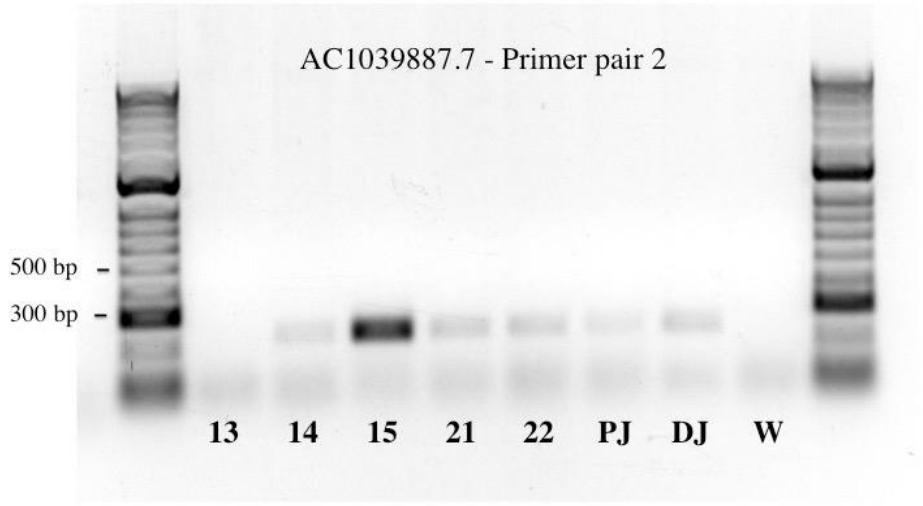

Figure 5.5 - Gel electrophoresis electrophoresis of PCR product (primer pair 2) from AC1039887.7 using DNA from monochromosomal somatic cell hybrids (mouse/human) as template. Hyperladder II was used. Lanes show hybrids for human chromosomes 13, 14, 15, 21, 22, 21derX (proximal side of short arm of chromosome 21), Xder21 (distal side of short arm of chromosome 21) and water control. BAC is located on chromosome 15 with a 479 bp length product.

PCR and gel electrophoresis for BAC AL591856 showed placement of BAC on chromosomes 14, 15, 22 and in the case of primer pair 84 on Xder21 (Fig. 5.6).

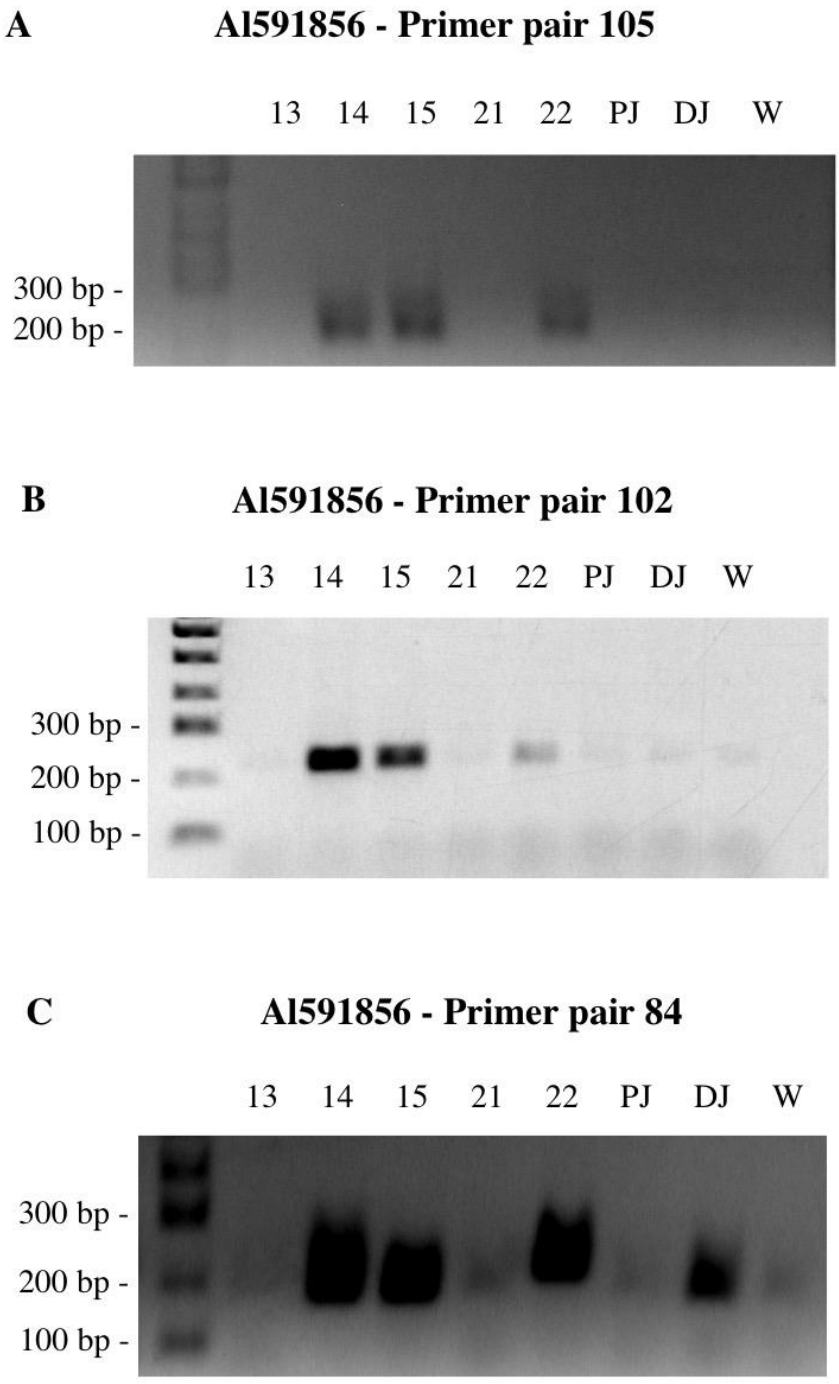

Figure 5.6 - Gel electrophoresis of PCR product from AL591856 using DNA from monochromosomal somatic cell hybrids (mouse/human) as template. Hyperladder 100 bp was used. Lanes show hybrids for human chromosomes 13, 14, 15, 21, 22, 21derX (proximal side of short arm of chromosome 21), Xder21 (distal side of short arm of chromosome 21) and water control. A – Primer pair 105. BAC is located on chromosomes 14, 15 and 22 with a 212 bp length product. B – Primer pair 102. BAC is located on chromosomes 14, 15, 22 with the expected product of length 211 bp. C – Primer pair 84. Products on chromosomes 14, 15 and "DJ" have the expected product of length 174 bp, chromosome 22 seems to have a longer product.

5.2.3 Confirmation of placement of BACs with FISH

Definite placement of each BAC along the chromosomes was confirmed by fluorescent in situ hybridisation (see sections 2.6, 2.12, and 2.14). PCR suggested that AC013640 and AC103988.7 were located on chromosome 15. To identify this chromosome in FISH, an alpha satellite probe specific to this chromosome was used (Choo et al., 1990). FISH was performed on human male metaphase slides from normal male PHA-stimulated lymphocytes from Applied Genetics Laboratories *inc*. Location of AC013640 and AC103988.7 was confirmed to be in the long arm of chromosome 15 near the centromere (Fig. 5.7 and 5.8). AL591856, however, is located on all short arms of acrocentric chromosomes on the distal side (Fig. 5.9 and 5.10).

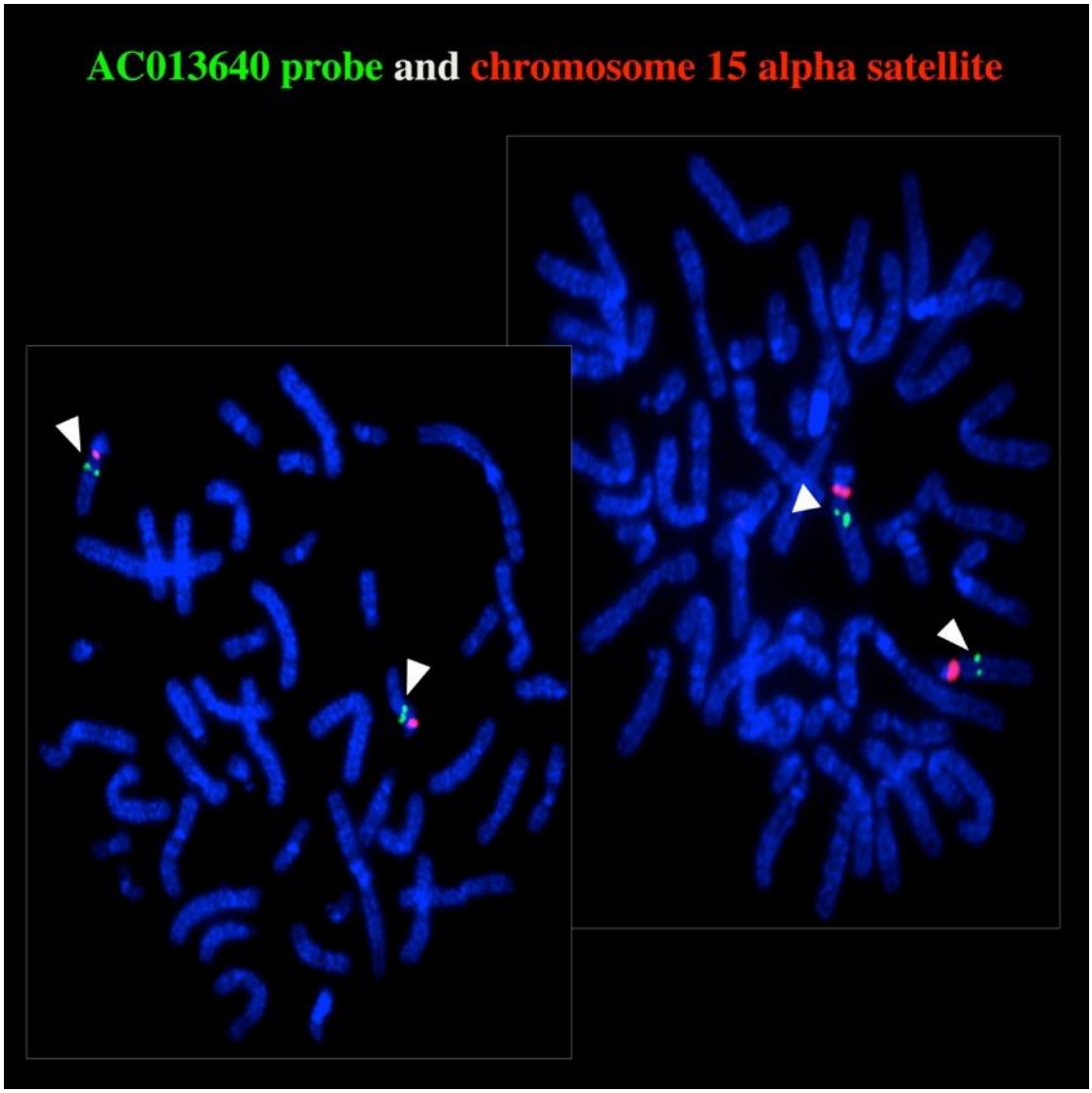

Figure 5.7 - FISH of AC03640 BAC on human male metaphase slides. A centromeric probe specific for chromosome 15 is shown in red. AC03640 probe is shown in green. There is hybridisation of BAC to the long arm of chromosome 15.

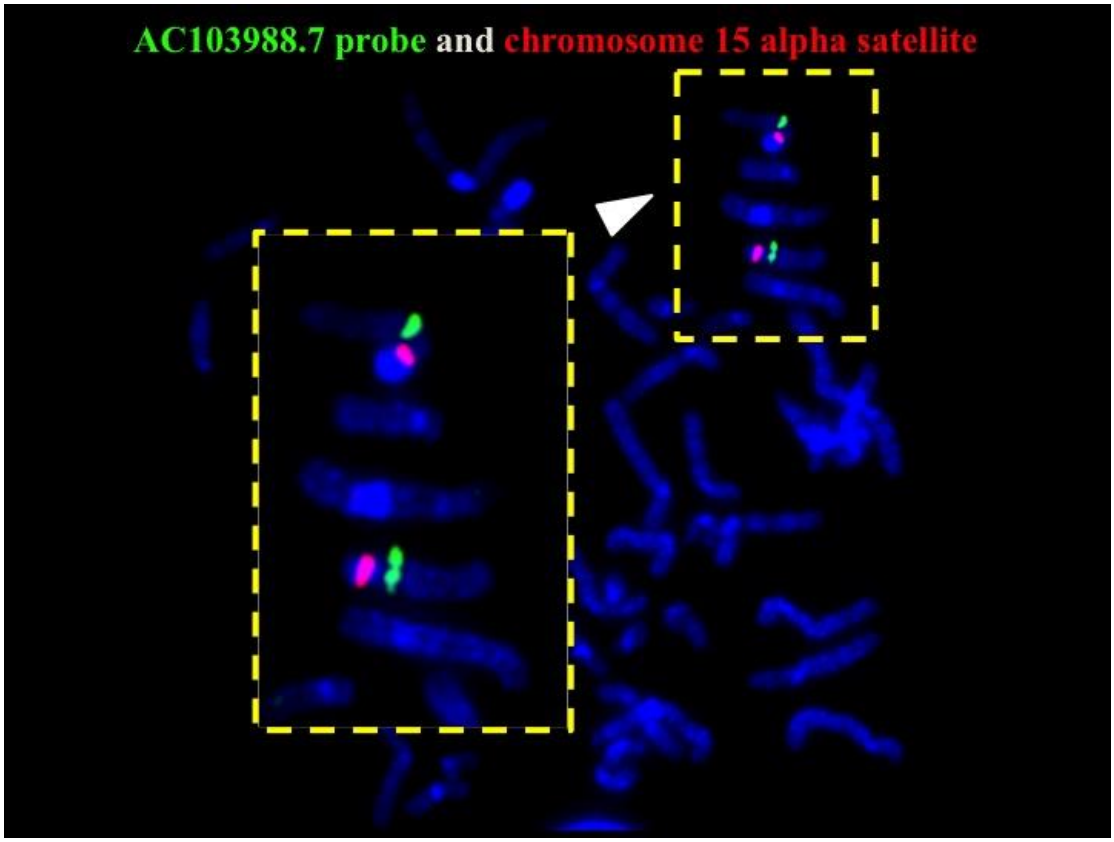

Figure 5.8 – FISH of AC103988.7 BAC on human male metaphase chromosomes. A centromeric probe specific for chromosomes 15 is shown in red. AC103988.7 probe is in green. FISH shows localisation of BAC in the long arm of chromosome 15.

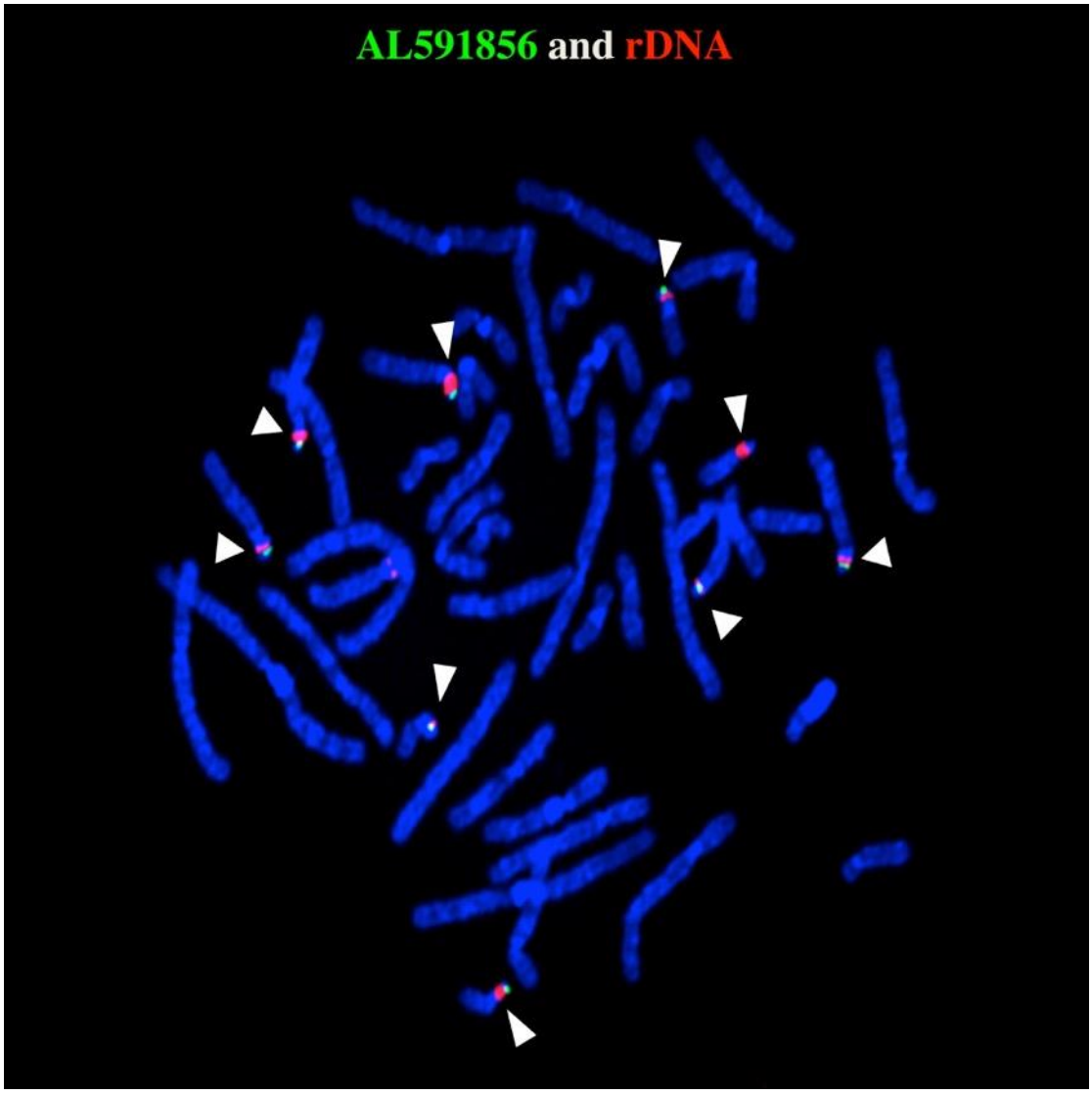

Figure 5.9 - FISH of AL591856 on male metaphase chromosomes. An rDNA probe is shown in red. AL591856 probe is shown in green. FISH confirmed localisation to the short arms of all acrocentric chromosomes on the distal side.

Chromosome specific alpha satellite probes were used to differentiate between each acrocentric chromosome (Fig. 5.10 - Brian McStay, Ioanna Floutsakou and Sofia Barreira). As in the previous figure, FISH shows different degrees of hybridisation, suggesting that there is sequence variability between the arms.

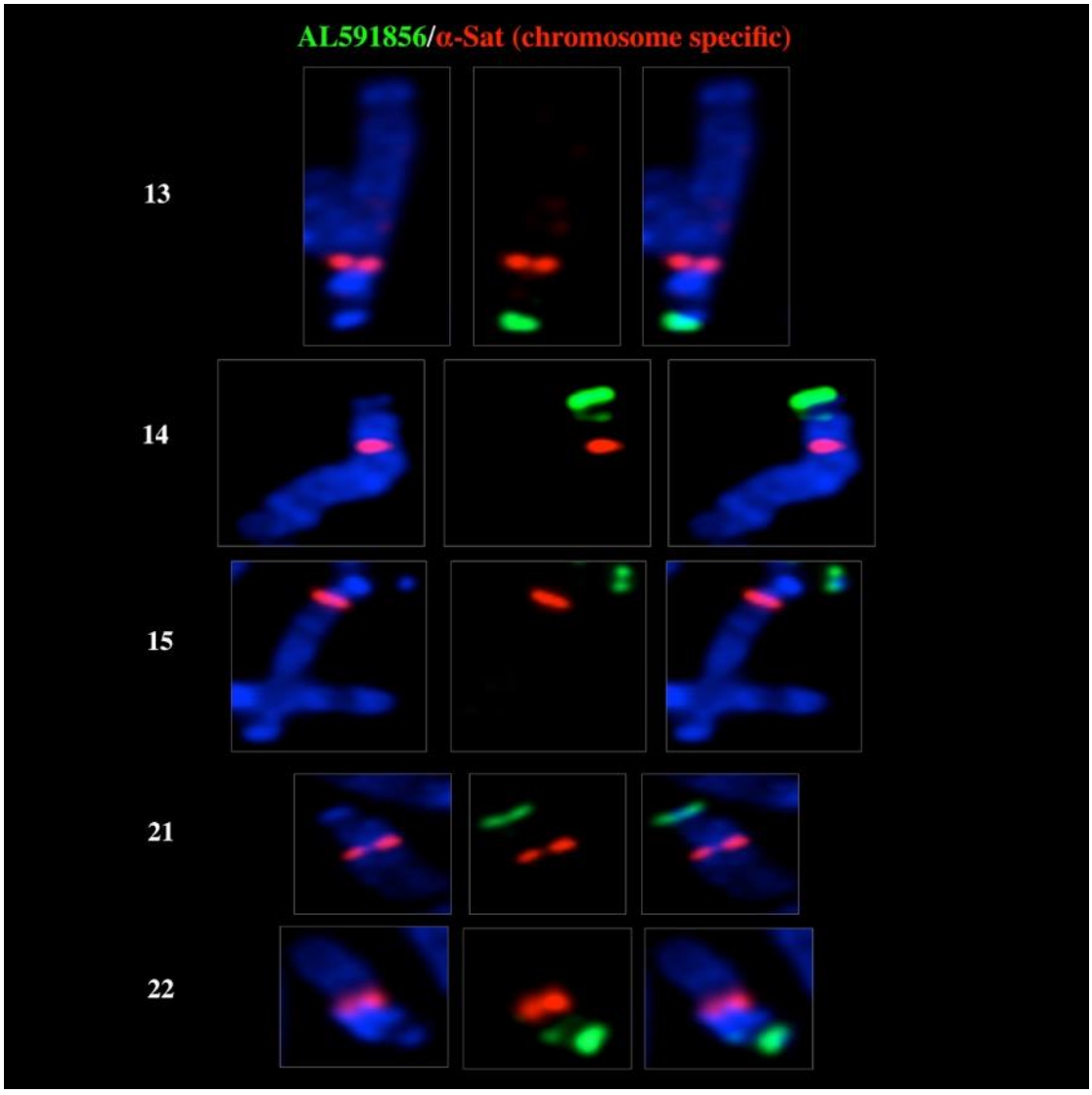

Figure 5.10 - FISH of BAC AL591856 and alpha satellite probes specific for individual acrocentric chromosomes on male metaphase chromosomes. AL591856 is in green and in red the alpha probes. AL591856 is located on the distal side of all acrocentric short arms.

5.2.4 Sequence composition of AL591856

Sequence analysis of AL591856 BAC revealed similarities to PJ (BACs AC145212, AL3548822, CR381535 and CR392039) and DJ (BAC AC011841) and a region between 40,352 bp and 110,982 bp that is distinct from the rest of the genome (Fig. 5.11). The PCR primer pairs for the AL591856 BAC were

designed within this unique region. Analysis of segmental duplications was carried out by dividing the BAC into 1 kb blocks with a sliding window of 500 bp and using BLAST to align these segments to the human genome (Fig. 5.11-B). AL591856 has a GC content of 43.3%.

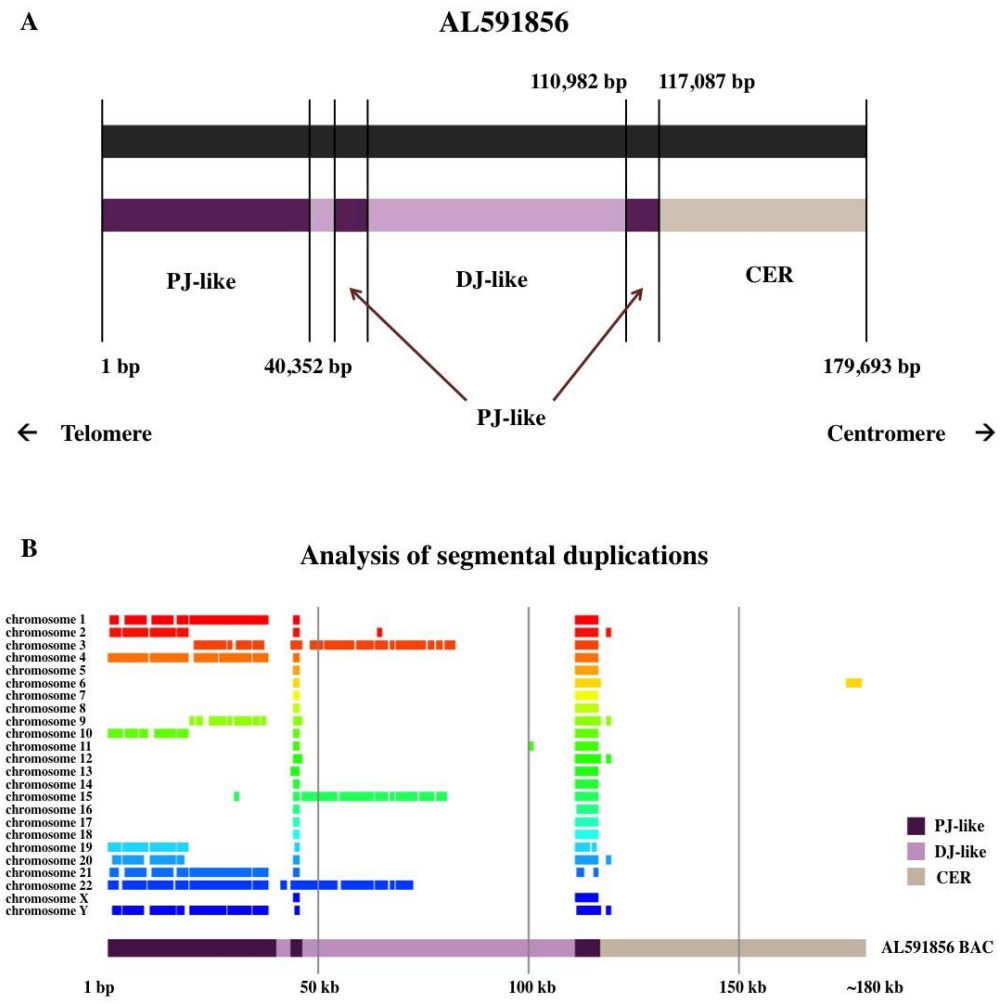

Figure 5.11 - Schematic of sequence homology for AL591856. First 40 kb and the region between 110,982 bp and 117,087 bp have similarities to the proximal junction. The last 63 kb have 77% similarity to the DJ BAC AC011841.7.

Closer inspection of the first 40 kb of AL591856 revealed a fragmented alignment to the PJ (96% sequence identity), with numerous segments within this region mapping to the same segments in the PJ. Similarly, segments of the

AL591856 sequence from 110,982 to 117,087 bp align to the PJ in different locations and orientations (13003 to 6914 bp, 58030 to 55300 bp and 193859 to 196168 bp). Inspection of FISH results confirms the cross-hybridisation between PJ and AL591856 (Fig. 5.12). The sequence at the end of the DJ, between 337837 to 379046 bp, has 77% identity to the last segment of AL591856 (117,087 to 179693 bp), placing this side of the new BAC towards the centromere and the other end, which has similarities to the PJ on the telomere side.

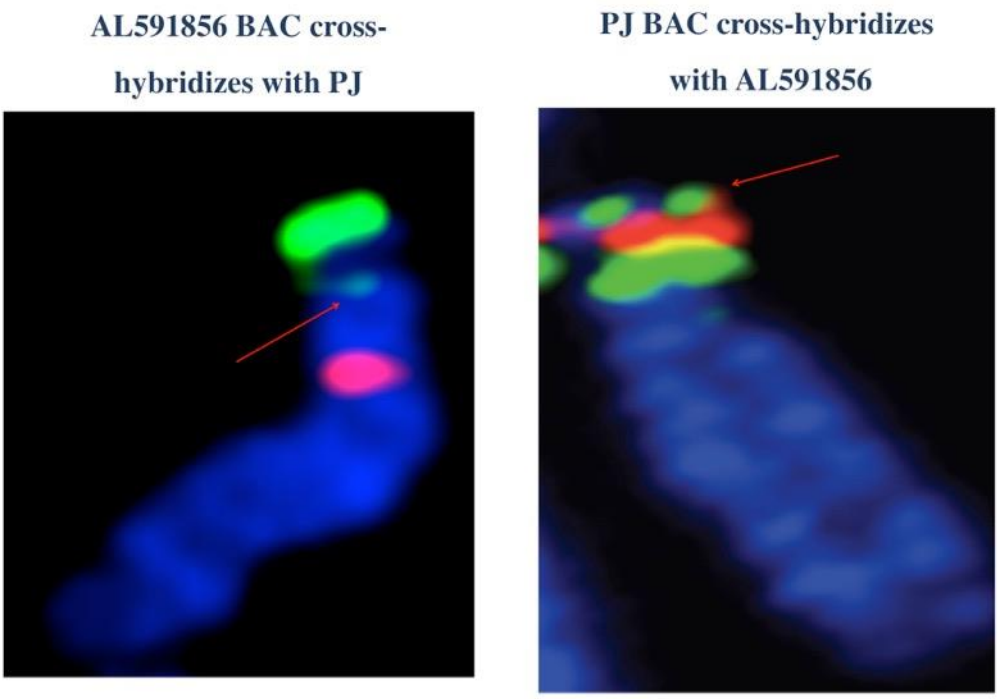

Figure 5.12 - FISH of AL591856 shows cross-hybridisation with the proximal side of rDNA. PJ BACs cross-hybridise with the distal side of rDNA, further confirming positioning of AL591856. Image of PJ BAC from (Floutsakou et al., 2013).

5.2.5 Analysis of the chromatin and gene expression profile of AL591856

ChIP-seq data sets (36 bp reads), for histone modifications and CTCF binding sites, from the ENCODE project (GEO accession number GSE29611) were aligned against a custom genome containing GRCh37 and the sequences for rDNA (extracted from BAC AL592188), DJ, PJ and AL591856 with bowtie (see section 2.20). Histone modifications studied included H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H4K20me1, H3K36me3 and H3K27me3 for numerous cell lines, including Gm12878, H1hesc, Hepg2, Hmec, Hsmm, Huvec, K562, Nhek and Nhlf. Previous to alignment, reads were subjected to quality control using Trimmomatic (see section 2.8). Alignment files were converted to sorted and indexed BAM files and duplicates removed with Picard Tools. Aligned reads were extended by 250 bp to mimic the estimated length of ChIP-DNA fragments. Data were normalized by subtracting the alignments from the control experiment for each cell line after normalising the data with the ratio of antibody/control (number of aligned reads in the genomic region of interest). RNA polymerase II ChIP-seq data for 5 cell lines, Gm12878, HepG2, Huvec, K562 and Nhek from ENCODE (GEO accession number GSE30226) were also analysed using the same strategy. Two cell lines, K562, a myelogenous leukaemia cell line, and Nhek, normal human epidermal keratinocytes, depicted high peaks for promoter and enhancer markers, H3K4me3, H3K4m2, H3K4me1, H3K9ac, H3K27ac, Pol II and CTCF binding sites, indicative of insulator activity and regulation of chromatin architecture (Figs 5.17 and 5.18 - H3K4me3, Pol II and CTCF -, Appendix C, Figures C1 to C12 - all histone markers). Markers for actively transcribed chromatin,

H3K36me3 and H4K20me1, were also analysed (Fig. 5.17 and 5.19 - H3K36me3 - Appendix C, Figures C1 to C12). In addition, the same ChIP-seq peaks were also called when using MACS (Zhang et al., 2008) with two additional peaks called for H3K4me3 and CTCF for the Huvec cell line.

Following the analysis of ChIP-seq data for AL591856, RNA-seq data sets from the ENCODE project were analysed to look for RNA originating in the new BAC. Data sets for normal cell lines (H1hesc, Gm12878, Huvec, Hsmm, Nhek and Nhlf) and cancer cell lines (HepG2 and K562) were aligned to a custom genome (GRCh37+rDNA+PJ+DJ+AL591856) using Tophat (see section 2.15) after quality control with Trimmomatic (see section 2.8). Duplicates were removed with Picard Tools and Cufflinks was used to assemble the transcriptome (see sections 2.15). Cuffmerge was employed to merge the assembled transcripts from the replicates data sets. Data from H1hesc, Huvec and Hsmm revealed no transcripts, however, K562 and Nhek showed presence of transcription activity in AL591856 as expected. Combined ChIP-seq data and RNA-seq data revealed evidence of transcriptional activity in K562 and Nhek (Fig. 5.13 and 5.14) but not in Huvec (Fig. 5.15).

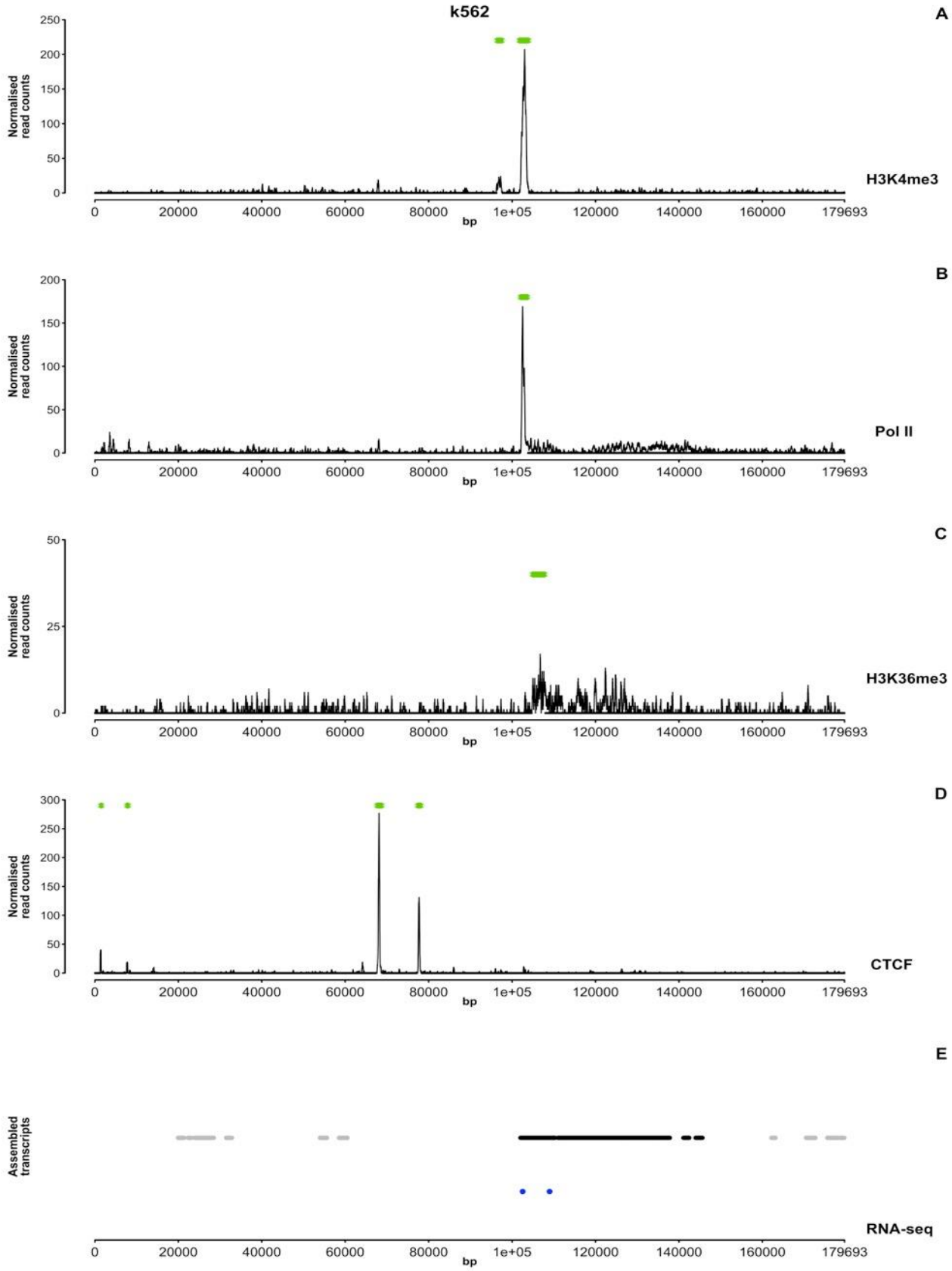

Figure 5.13 - ChIP-seq peaks for H3K4me3, Pol II H3K36me3 and CTCF and assembled RNA-seq transcripts for K562. Green dots specify the position of peaks called by MACS. A - ChIP-seq peaks for H3K4me3, indicative of transcription start sites and promoter regions of actively transcribed genes. B - ChIP-seq peaks for Pol II, indicative of gene promoters of actively transcribed genes. C - ChIP-seq peaks for H3K36me3, indicative of transcribed gene bodies. D - CTCF ChIP-seq peaks, indicative of boundaries of histone methylation domains. E - Assembled RNA-seq transcripts. Transcripts outside the ChIP-seq peaks for H3K36me3 are shown in grey. Blue dots represent the position of the PCR primers used to perform RT-PCR to confirm the existence of the transcript *in vivo*.

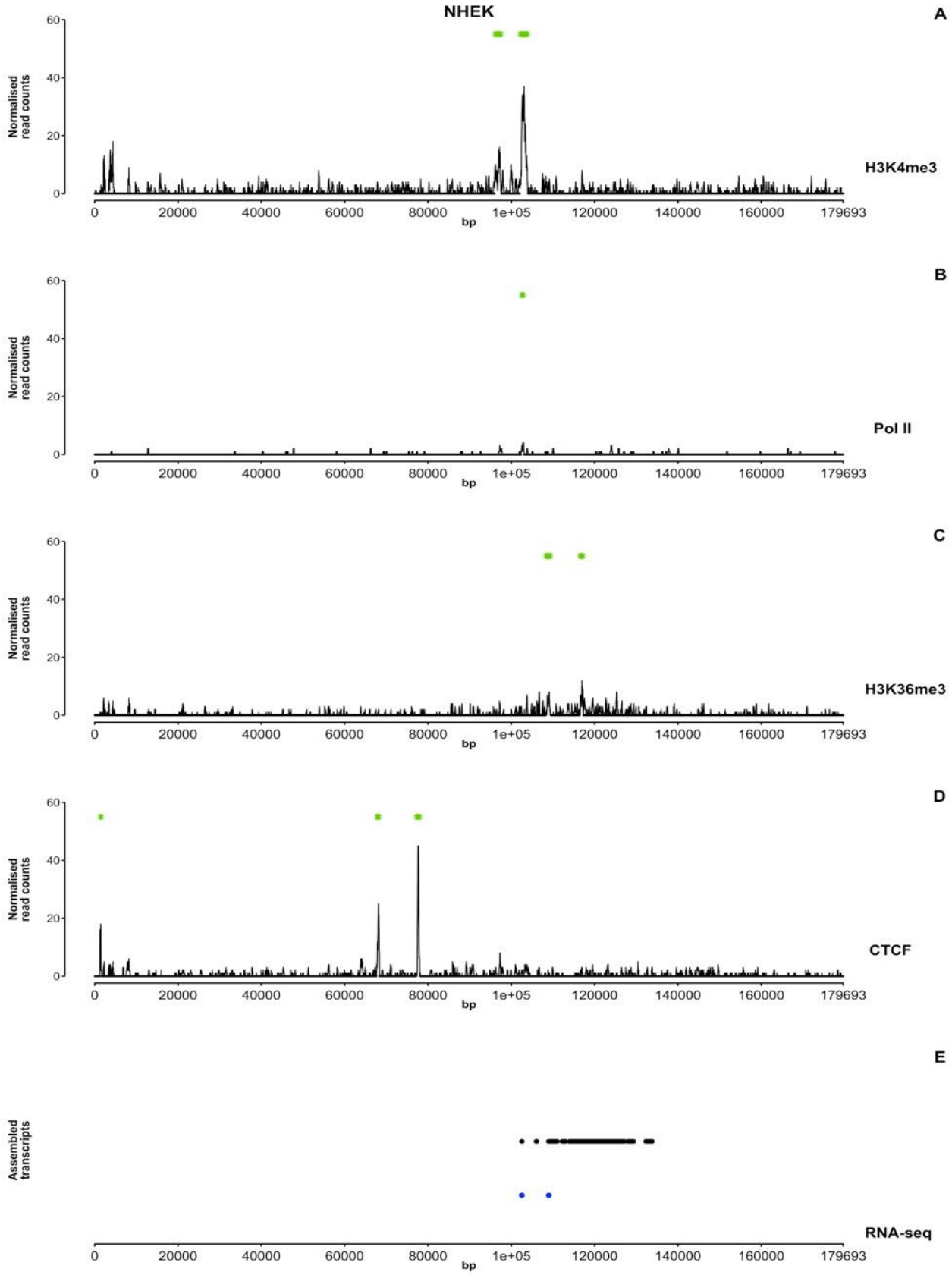

Figure 5.14 - ChIP-seq peaks for H3K4me3, Pol II H3K36me3 and CTCF and assembled RNA-seq transcripts for Nhek. Green dots specify the position of peaks called by MACS. A - ChIP-seq peaks for H3K4me3, indicative of transcription start sites and promoter regions of actively transcribed genes. B - ChIP-seq peaks for Pol II, indicative of gene promoters of actively transcribed genes. C - ChIP-seq peaks for H3K36me3, indicative of transcribed gene bodies. D - CTCF ChIP-seq peaks, indicative of boundaries of histone methylation domains. E - Assembled RNA-seq transcripts. Transcripts outside the ChIP-seq peaks for H3K36me3 are shown in grey. Blue dots represent the position of the PCR primers used to perform RT-PCR to confirm the existence of the transcript *in vivo*.

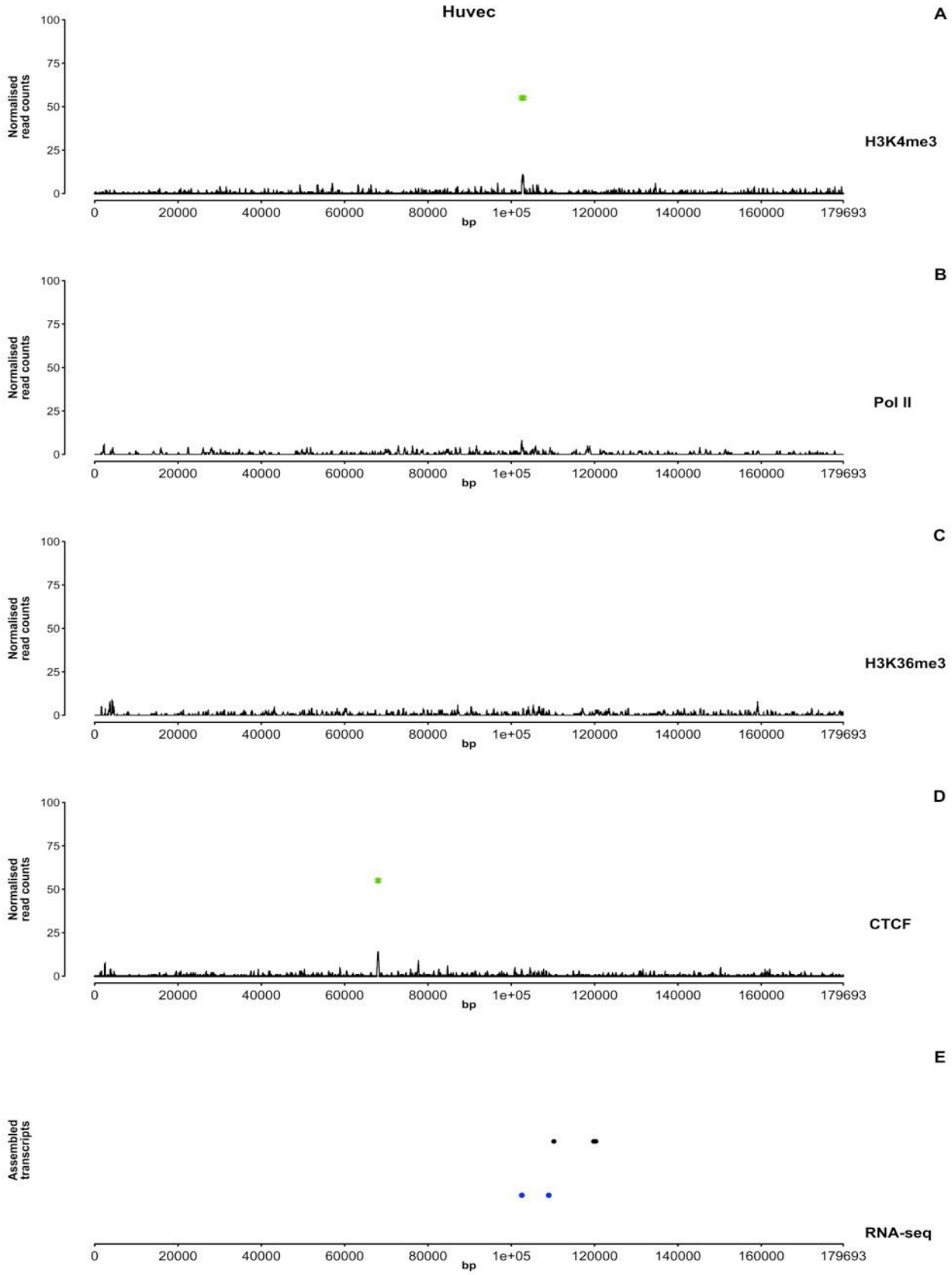

Figure 5.15 - ChIP-seq peaks for H3K4me3, Pol II H3K36me3 and CTCF and assembled RNA-seq transcripts for Huvec. Green dots specify the position of peaks called by MACS. A - ChIP-seq peaks for H3K4me3, indicative of transcription start sites and promoter regions of actively transcribed genes. B - ChIP-seq peaks for Pol II, indicative of gene promoters of actively transcribed genes. C - ChIP-seq peaks for H3K36me3, indicative of transcribed gene bodies. D - CTCF ChIP-seq peaks, indicative of boundaries of histone methylation domains. E - Assembled RNA-seq transcripts. Transcripts outside the ChIP-seq peaks for H3K36me3 are shown in grey. Blue dots represent the position of the PCR primers used to perform RT-PCR to confirm the existence of the transcript *in vivo*.

5.2.6 Confirmation of transcripts from AL591856 through Reverse Transcriptase PCR

Reverse transcriptase PCR (RT-PCR) was performed on HT1080 cells, a fibrosarcoma cell line, to acquire wet-lab evidence of the occurrence of transcription (See section 2.17). Oligo dT primers were first used to synthesize cDNA from the HT1080 RNA. This kind of primer hybridizes to the Poly(A) tail of mRNA, only converting mature mRNA, that was transcribed by Pol II (McCracken et al., 1997), to its complementary DNA. A specific primer pair was designed for the transcript identified in the unique region of AL591856 (Fig. 5.13 – last row, Table 5.4).

Table 5.4 - Primer pair sequence and expected product length for transcript from AL591856.

Primer Pair	Sequence	Product length
XR1109f	GAGAATCCCGGCTCCTGCTGC	-
XR1109r	GCAGGAAGAAGTCTCTCATCAGG	198 bp

RT-PCR revealed the presence of two transcript variants (Fig. 5.16). The amplified cDNA that resulted from the RT-PCR was cloned (see section 2.18 and 2.19) and sent for sequencing. The sequenced clones (sequences in Appendix B, table B1) revealed the two transcript variants captured by RT-PCR differed in the size of exon 2 (Fig. 5.17). Exon 2 from the 287 bp product is 159 bp and 61 bp in the 198 bp product.

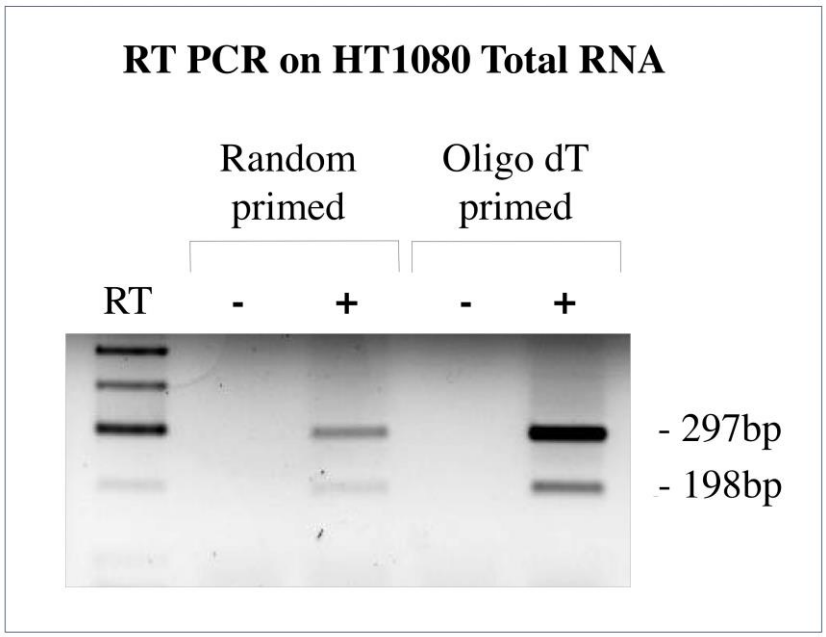

Figure 5.16 - Reverse transcriptase PCR to confirm the occurrence of transcription in AL591856.

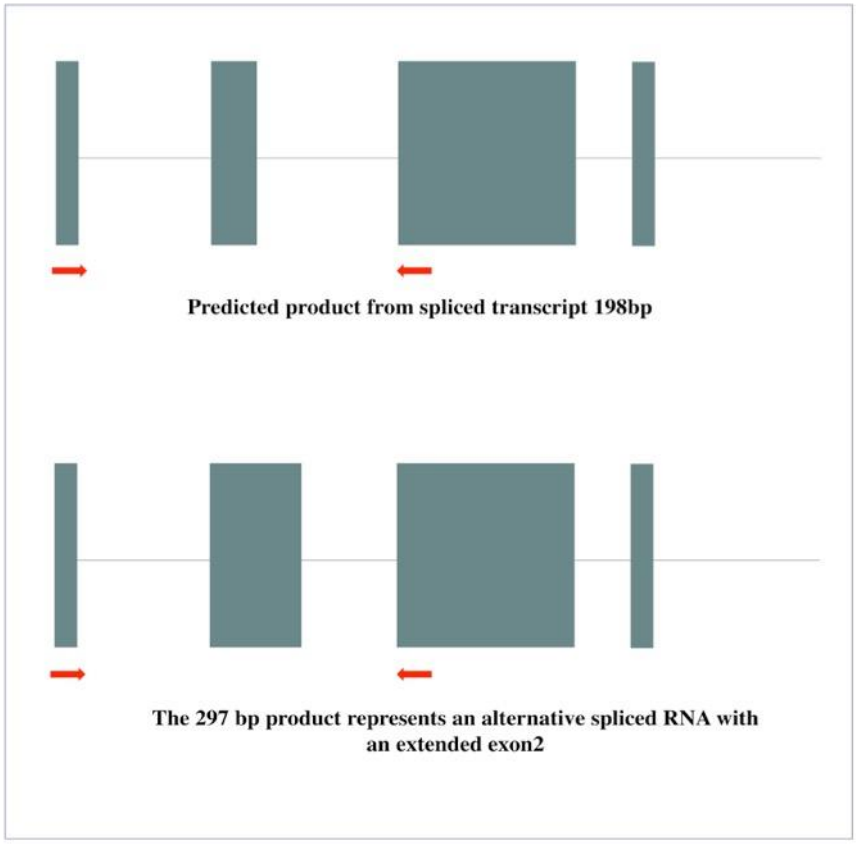

Figure 5.17 - Schematic of the two spliced variants identified by RT-PCR. The two transcripts differ in the length of exon 2.

5.3 Discussion

Repetitive regions are difficult to sequence and assemble with high-throughput sequencing data. Repeats sequenced with short reads lead to ambiguities that introduce errors and biases in alignment and assembly results. Due to this effect, numerous regions in the human genome are not included in the current reference assembly (GRCh38).

Previously sequenced 454 reads did not yield sufficient coverage whilst attempting to assemble the rDNA repeats. However, the DNA used for sequencing was extracted from isolated nucleoli and therefore, there was the possibility that the remainder of the short arms of acrocentric chromosomes were represented in the reads. Assembled contigs from 454 nucleolar data were searched with BLAST using the Hi-C reads that were not aligned to the known genome but with mates that mapped to the DJ. The reported contigs were then used to search for unplaced BACs from the human genome project that could expand the distal side of the rDNA repeat clusters. Several potential novel sequences (BACs) were identified. Primer pairs were generated to assess their presence in acrocentric chromosomes using monochromosomal human/rodent hybrids in PCR. PCR and FISH revealed two of the pursued BACs were located in the long arm of chromosome 15. Previously identified as belonging to a genomic contig from chromosome 15 in GRCh37, these BACs are currently part of GRCh38 and CHM1 primary assemblies.

One BAC, AL591856 (clone RP11-426M5, 179,693 bp), had positive results in some of the chromosomal hybrids (Fig. 5.6). Fluorescent *in situ* hybridisation showed sequences that hybridise to AL591856 were present in all

p-arms of acrocentric chromosomes (Fig. 5.9 and 5.10). However, further evidence of sequence variation is also observed. The intensity of the hybridisation signals was lower in chromosomes 15 and 21 (Fig. 5.10). This is in part concordant with the lack of PCR products in chromosomes 13, 21 and 21derX (PJ). Interestingly, although all primers produced products in chromosomes 14, 15 and 22, only the primer originating from the sequence closest to the telomere side showed a product in Xder21 (DJ). This could be due to sequence variability in the different sources for the human cells used to create the monochromosomal hybrids. The monochromosomal for chromosome 21 has a human chromosome from a normal adult male fibroblast strain 1BR.2 (Cuthbert et al., 1995). Whereas the derivatives Xder21 and 21derX have a female donor (Jacobs et al., 1981). All this implies that the AL591856 BAC marks the start of interchromosomal divergence on the distal side of acrocentric short arms. Importantly, this contrasts with the shared sequence conservation observed in the DJ in all acrocentric chromosomes. AL591856 was then localised in the short arms after the DJ towards the telomere (Fig. 5.18). The CER blocks from each sequence either overlap or follow one another.

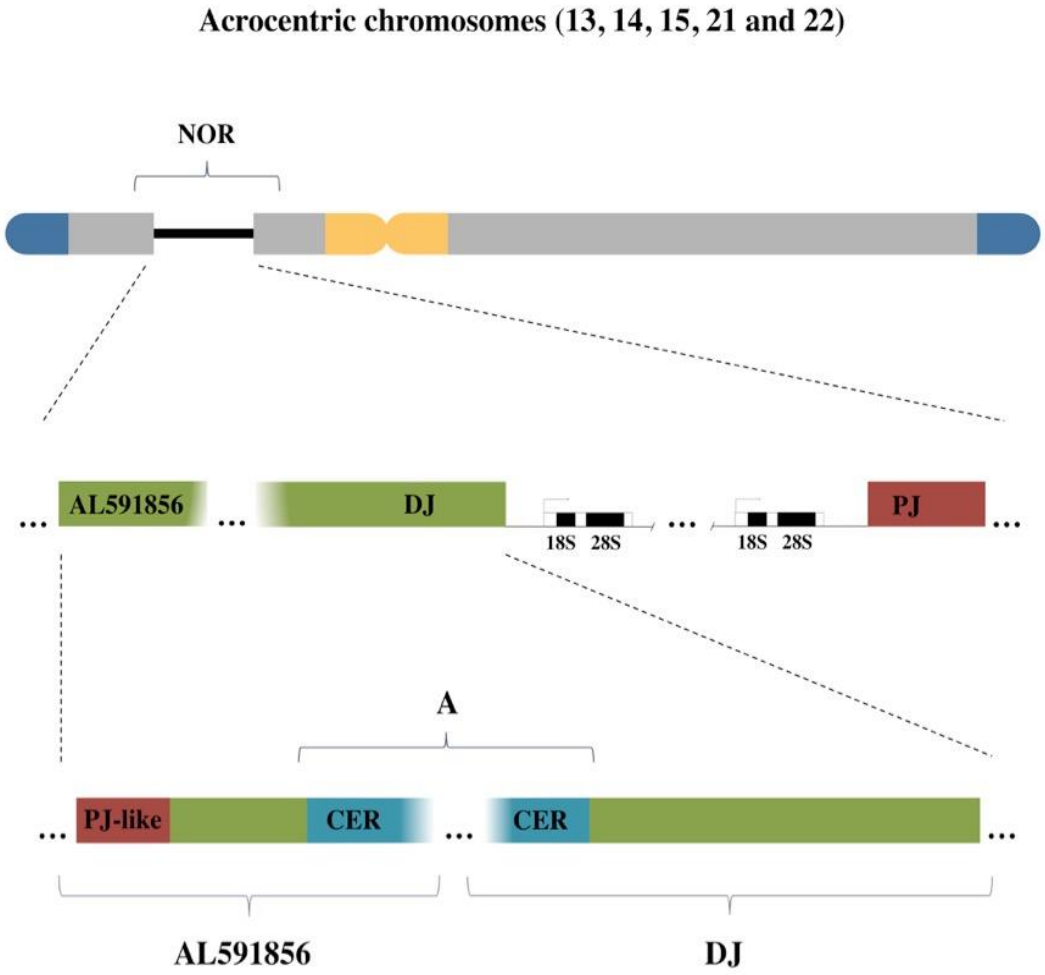

Figure 5.18 - Positioning of AL591856 in the short arms of acrocentric chromosomes. BAC is located on the distal side of the rDNA clusters immediately after the distal junction sequence. The DJ and AL591856 have CER (A) repeat blocks sharing 77% identity. It is not known if these repeats overlap to form an uninterrupted contig or if there is a sequence gap between them.

The human reference genome is a well-reviewed and annotated sequence that was generated through sequencing reads and optical/physical maps including FISH analysis. Experimental data generated in this project supports the existence and placement of AL591856 in the human genome and its functional activity.

Sequence analysis of this BAC showed that the last ~62 kb, from 117087 bp to 179693 bp, have 77% identity with the DJ, particularly to the CER satellite repeats. The CER repeat is a 48 bp sequence (5'-TTCCAGAACACTGCTRCKRGGGTCTGAATGTTTGTCCCTCACATAGGA

-3') located in all acrocentric short arms on the distal side of rDNA that also co-localises to the centromere region on chromosomes 14 and 22. The first 40 kb of AL591856 have 96% identity with the PJ and also depicted the high level of segmental duplication to other chromosomes as the proximal junction. Importantly, there is cross-hybridisation between AL591856 and PJ (Fig. 5.12), which confirmed the sequence composition analysis of this new BAC (Fig. 5.11). However, this makes unique mapping of other BACs or high-throughput sequencing reads to this region of AL591856 difficult, complicating further extension of the sequence information of the short arms on either side of NORs.

Analysis of the chromatin profile of AL591856 revealed a second level of variation. Whereas the known DJ has a consistent chromatin structure across all cell lines studied (in both cancer and healthy cells), the new BAC depicts patterns of functional variation. ChIP-seq peaks revealed the presence of histone modification markers associated with transcription activation, promoters, transcription start sites and active enhancers (H3K4me3, H3K4me2, H3K4me1, H3K9ac and H3K27ac) and actively transcribed gene bodies (H3K36me3) in two cell lines, K562 and Nhek (Fig. 5.13 and 5.14, and C1 – C12). However, some markers for active or repressed genes and poised enhancers (H4K20me1, H3K9me3 and H3K27me3) had lower signals in most of the cell lines studied (Fig. C1 – C12). Nevertheless, H3K9me3 (Fig. C5), indicative of repressed genes and H3K36me3 (Fig. C8), indicative of actively transcribed regions, depict faint peaks that complement each other. Peaks for CTCF (Fig. 5.13 and 5.14), an insulator binding protein, were also present in k562 and Nhek. CTCF has as primary function the prevention of unwanted interactions between regions of the genome (Cuddapah et al., 2009) and is involved in unravelling closed chromatin

(Weth et al., 2014). Pol II peaks are also observed in K562 (Fig. 5.13). Peak calling performed with MACs confirmed my ChIP-seq peaks and called an extra peak for Pol II in Nhek (Fig. 5.14) that my analysis did not reveal. Importantly, ChIP-seq peaks occur in both normal (Nhek) and cancer (K562) cell lines but are also absent in cells from both types (Gm12878, H1hesc, Hsmm, Huvec, Nhlf and HepG2). Despite the lack of strong ChIP-seq evidence for transcribed gene bodies in the majority of cells, transcript assembly and analysis of RNA-seq reads in AL591856 was carried out. Significantly, RNA-seq reads from K562 and Nhek revealed the presence of RNA originating from AL591856 (Fig. 5.13 and 5.14). Reverse transcriptase PCR performed on RNA from HT1080 cells (from a fibroblastic sarcoma), confirmed the existence of two RNA transcripts (Fig. 5.16). Subsequent cloning and sequencing of the cDNA from the RT-PCR revealed differences in the size of exon 2 (Fig. 5.17). Without further analysis of these transcripts it is not possible to ascertain if this new BAC contributes to nucleolar regulation and function. However, given its degree of conservation across the short arms, its proximity to the rDNA arrays and the DJ, and the presence of some transcriptional activity in specific cells types including cancer cells (K562 and HT1080. More primer pairs for RT-PCR spanning the entire RNA-seq transcript should be designed and tested to account for all potential spliced transcripts. It is possible these transcripts are translated into proteins and further analysis should be carried out, including *in silico* translation and search for hydrophobic amino acids that allow proper protein folding. A search for potential proteins in Swiss-Prot or even UniProtKB should be performed, as any peptides originating from AL591856 might have been identified but not annotated or assigned to a genomic location.

The conserved region of the short arms of acrocentric chromosomes starts and ends with PJ and PJ-like sequences (Fig. 5.18). Towards the telomere side, the new added sequence AL591856 showed a lower degree of interchromosomal conservation than the DJ. This was observed in the varying degrees of hybridisation intensity and in the lack of PCR products in some acrocentrics. This further reinforced by the lack of product in the chromosome 21 hybrid and existence of product in the Xder21 translocation chromosome. The human chromosomes used in both hybrids come from human fibroblasts. Functional variability was observed in different cell lines and further analysis to determine if all acrocentric chromosomes are actively transcribing AL591856 should be carried out.

Expanding the sequence information of the short arms of acrocentric chromosomes towards the telomere will require a different approach, with much longer and precise sequencing reads. Likewise, sequencing and assembling the arms separately, with a high-accuracy sequencing technology, aided by visual guides such as optical maps, would be the best strategy to complete the short arms of the human acrocentric chromosomes.

6 Conclusions and Future Work

The human reference genome is not complete. This thesis has focussed on one of the missing regions. Nucleoli organiser regions, containing the rDNA repeats and around which the nucleolus form, are located in the p-arms of acrocentric chromosomes. The entire short arms of these chromosomes are missing from the genome draft. The overall aim of my thesis was to improve our understanding of the organisation and function of these regions.

Previous work had hinted that as many as 30% of rDNA repeats are rearranged, possibly impacting on nucleolar and ribosome formation and protein synthesis (Caburet et al., 2005). In chapter 3 of this thesis, I describe attempts to confirm the presence of rearranged rDNA by directly sequencing DNA from purified nucleoli and searching for paired-end reads indicative of rearrangements. Single molecule real-time sequencing reads were also employed in the search for rearrangements. Although a considerable number of reads mapped to the rDNA repeat none of the data sets contained reads indicating valid rearrangements.

Considering the high number of sequencing reads that mapped to the rDNA, particularly in the CHM1 set, it is possible the problem lies with the experimental data on which claims of inverted rDNA repeats have been based. The combing technology relies on DNA molecules being stretched across a solid surface by a receding air-water interface (Bensimon et al., 1994). This method does not guarantee isolation of single DNA strands. The apparent rDNA rearrangements in the combing data could result from multiple hybridised fibres

localised in the same stretch. Also, the observed rearrangements could be replicating strands that collapsed together to the same spot during the extension procedure. A different experimental procedure, using non-replicating cells, which can be achieved by serum starvation, and ensuring individual DNA molecules for hybridisation of 18S and 28S sequences could help clarify the issue. The Irys system, developed by Bionano Genomics, uses nanochannel arrays to separate single DNA molecules and search for sequence motifs and structural variation with fluorescent labels (Lam et al., 2012). This technology could be used to visualise entire rDNA repeats from single DNA molecules using fluorescent probes for 18S and 28S and establish the organisation of rDNA.

Sequencing an acrocentric chromosome on its own would give a detailed map of the number of rDNA genes per chromosome. It is possible that the number of repeats varies between chromosomes, cell lines, and even individuals.

Chapter 4 explored the spatial organisation of the distal junction employing numerous Hi-C data sets. The arms of the inverted repeat that comprises most of the DJ appear to fold at their spacer to form a stem-like loop (Fig. 4.7). This loop, also called the DJ domain, makes the sequences of the long non-coding RNAs facing in the same direction. CTCF peaks from both arms overlap at the top of the fold, near the regions with the highest number of contacts between the two arms. CTCF is a conserved zinc-finger protein associated with, among other functions, insulator activity and regulation of chromatin architecture (Cuddapah et al., 2009; Guelen et al., 2008). It is possible CTCF binds together the two DNA strands, either maintaining the loop structure or being responsible for its appearance in the first place. Cohesin often associates functionally with CTCF in fact, CTCF is required for cohesin ligation (Parelho et

al., 2008). Localisation of cohesin in the DJ should be analysed through ChIP or ChIP-seq to test for this functional association in this case. Cohesin not only contributes to enhancing CTCF insulator binding activity but also forms chromosomal cis-interactions (Hadjur et al., 2009; Mehta et al., 2013). The association of CTCF and cohesin would explain how the fold is maintained. CTCF peaks also occur near the bottom of the loop but do not overlap in the contacts between the two arms. In this instance, CTCF probably acts as an insulator to the transcription of the DJ (Guo et al., 2015). However, this is an analysis of intrachromosomal interactions. These CTCF sites might interact with CTCF sites located in other chromosomes. In interphase cells, the DJ locates to the periphery of nucleoli, where it seems to anchor the rDNA repeats being transcribed in the centre of the nucleolus. Therefore, the DJ domain might also have a structural role in helping shaping nucleoli by acting as a tether to other chromatin strands in the nucleolar heterochromatin shell. In order to pursue this, however, it will be necessary to devise ways to study the chromatin interactions of individual DJs.

A Hi-C data set from Jin *et al* (Jin et al., 2013), derived from cells treated with flavopiridol. Flavopiridol inhibits mRNA production by inactivating the elongation factor P-TEFb which results in the blockage of Pol II (Chao and Price, 2001). The DJ domain is not observable in this sample, suggesting that the folding of the DJ is maintained by DJ transcription. This is also evidence that the observation of this structure in the DJ is not an artefact of Hi-C data.

In the future, a different experimental technique, such as 3C, should be implemented to confirm the existence of this structure *in vitro*. CRISPR-Cas9 could be employed to address the implications of not having this structure in

active cells with different strategies. To check if CTCF is needed to maintain the structure, guide RNAs could be used to block the CTCF sites followed by 3C to quantify the number of interactions between the two arms. To confirm that the domain is maintained by transcription, guide RNAs could also be used to block promoter sites in the DJ. CRISPR-Cas9 could also be used to delete or add a few nucleotides in the sequence of the long non-coding RNAs to help gauge their function. Analysis of the role of the DJ long non-coding RNAs should be carried out, as these together with the domain might be important for nucleolar regulation.

Initially, chapter 5 had as the main objective to extend the sequence information on the telomere side of NORs. The region immediately following the DJ was further extended by 180 kb through the identification of a BAC mapping to this region. This BAC, AL591856, was identified by combining Hi-C reads and contigs assembled from nucleolar high-throughput sequencing. The principle that regions in close proximity have a higher degree of interaction was applied to find contigs that contained a Hi-C read whose mate mapped to the DJ. These contigs were used to search for unplaced BACs that could be placed in the short arms of acrocentric chromosomes.

AL591856 not only was located after the distal junction but denotes the end region that is conserved and shared among NORs. The beginning of the BAC, on the DJ side contains CER blocks that have around 77% identity to the CER blocks at the end of the DJ. The last 40 kb on the telomere side has 96% sequence identity with the proximal junction, and like the PJ, is heavily segmentally duplicated to other chromosomes. This signifies that NOR sequence conservation starts and ends with PJ and PJ-like sequences. The sequence of this

BAC also varies between acrocentric chromosomes and its chromatin profile changes dramatically between cell lines. Primer pairs located in the centre region of AL591856, which is mostly unique to this BAC, did not generate products in all acrocentric chromosomes. The internal variability of the new BAC was confirmed by FISH on metaphase chromosome spreads. Although AL591856 hybridised to the region after the DJ in all acrocentric chromosomes, the hybridisation intensity varied. ChIP-seq, RNA-seq and RT-PCR analyses revealed AL591856 possesses histone modifications associated with modulation of transcription. It is transcribed in only a few of the cell lines studied. This poses a new question, regarding the purpose of the transcripts from this BAC and why they are generated only in certain cell lines.

Completing the distal side of the short arms of acrocentric chromosomes will require a different approach than chromosome walking using BACs or sequencing reads generated from whole cells. Individual chromosomes should be sequenced for a comprehensive assembly of the p-arms of acrocentric chromosomes. Isolating the region after the last rDNA repeat with a combination of I-Ppo1 enzyme, which cuts once per repeat, or using CRISPR-Cas9 to cut in the unique region of AL591856, with pulsed-field gel electrophoresis to separate this fragment followed by sequencing with long and accurate reads should yield better results. This method would have to be carried out in individual acrocentric chromosomes, attained using monochromosomal hybrids, to complete sequences separately.

In summary, I have found that there is no evidence from high throughput sequence data to support the existence of rearranged rDNA repeats. Other experimental techniques should be employed to shed light on why these could be

observed through molecular combing. I have characterised the internal organisation of the DJ and identified a structural domain that possibly contributes to nucleolar maintenance. I identified additional sequences distal to the known DJ and have shown that these vary between acrocentric chromosomes. Key questions to be addressed in the future are to ascertain the role of the transcripts generated in the DJ and AL591856, how they contribute to nucleolar function and whether the remaining unassembled sequences are functionally relevant to the nucleolus or the cell.

Appendix A Hi-C figures

Figures from Hi-C data sets GSE18199, GSE37752, GSE41763, GSE44267 and GSE51687 are shown below. Interactions between the arms of the inverted repeat present in the DJ can be observed in all sets except GSE18199. Other large inverted repeats found in the human genome did not show interactions between the two arms of each repeat.

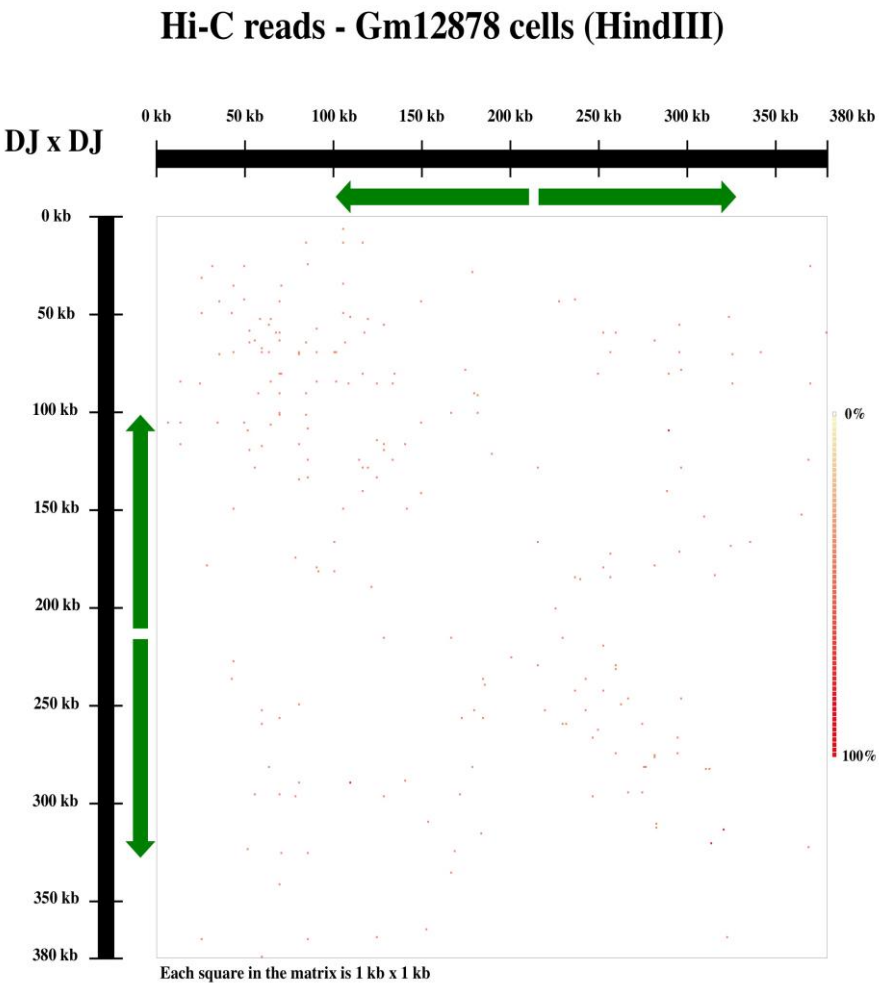

Figure A 1 - Intramolecular interactions in the DJ using Hi-C reads from Gm12878 cells in normal conditions (replicate sample). HindIII (A'AGCTT) restriction enzyme was used. Scale from no

observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

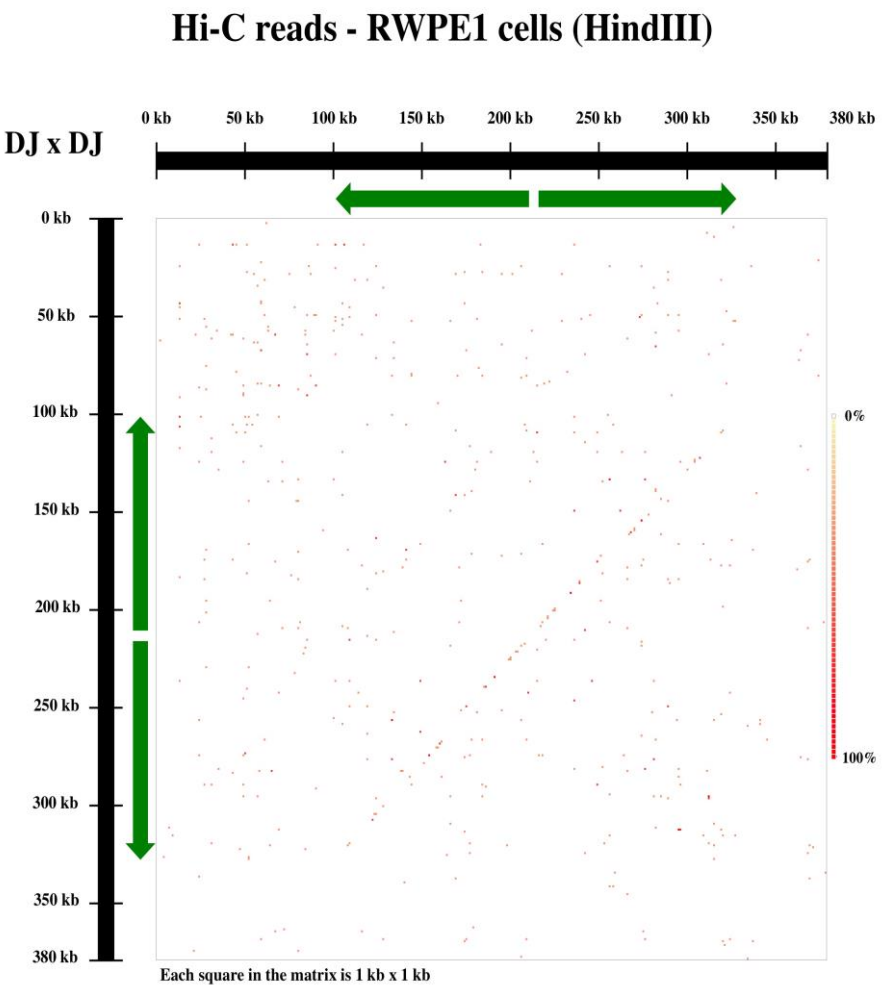

Figure A 2 - Intramolecular interactions in the DJ using Hi-C reads from RWPE1 cells in normal conditions. HindIII (A'AGCTT) restriction enzyme was used. The two arms of the inverted repeat present in the DJ show contact with each other. Scale from no observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

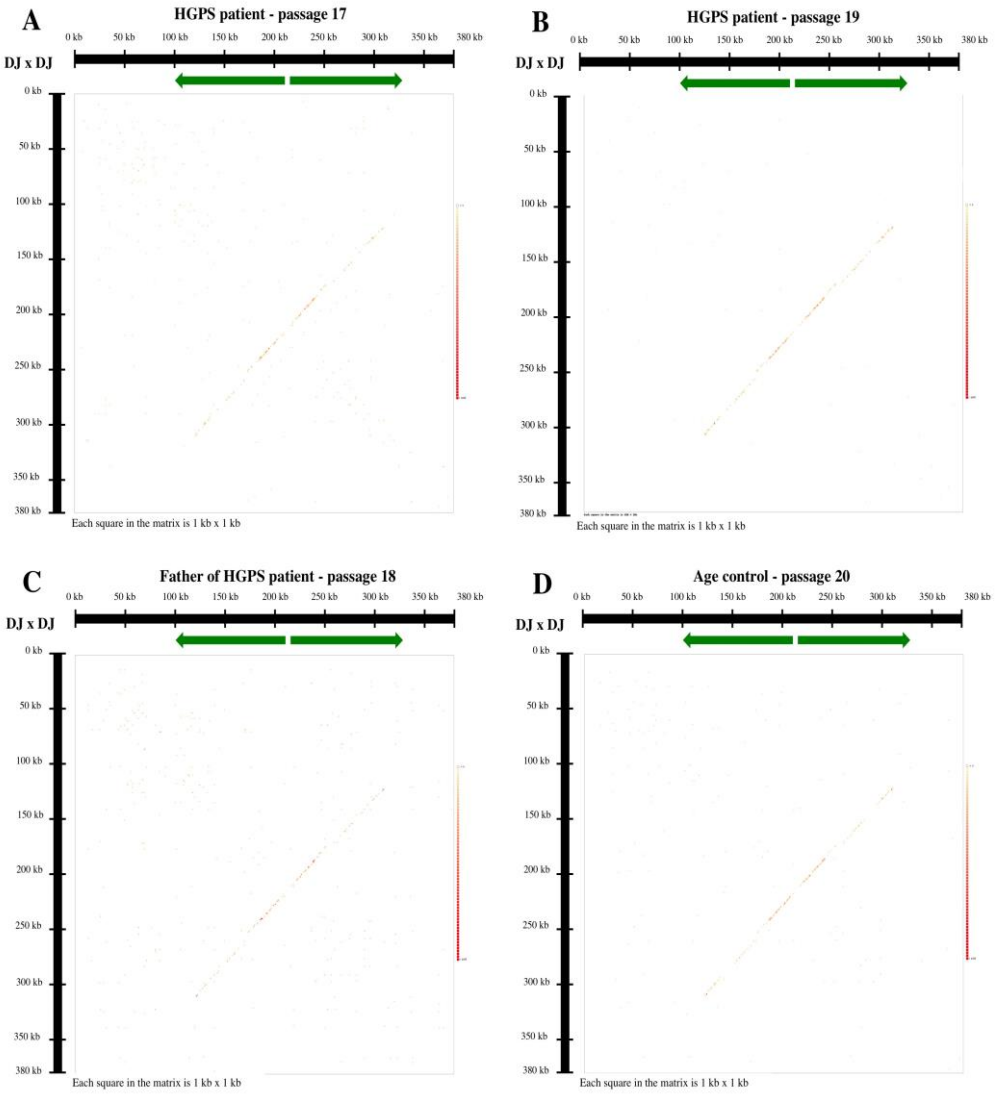

Figure A 3 - Intramolecular interactions in the DJ using Hi-C reads from Huntington-Guilford Progeria Syndrome (HGPS) fibroblasts in normal conditions. HindIII (A'AGCTT) restriction enzyme was used. All samples show high number of contacts between the two arms of the large inverted repeat present in the DJ. Scale from no observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb. A – Sample from HGPS patient (cell passage 17). B – Sample from HGPS patient (cell passage 19). C – Sample from father of HGPS patient (cell passage 18). D – Age control sample, normal fibroblasts (passage 20).

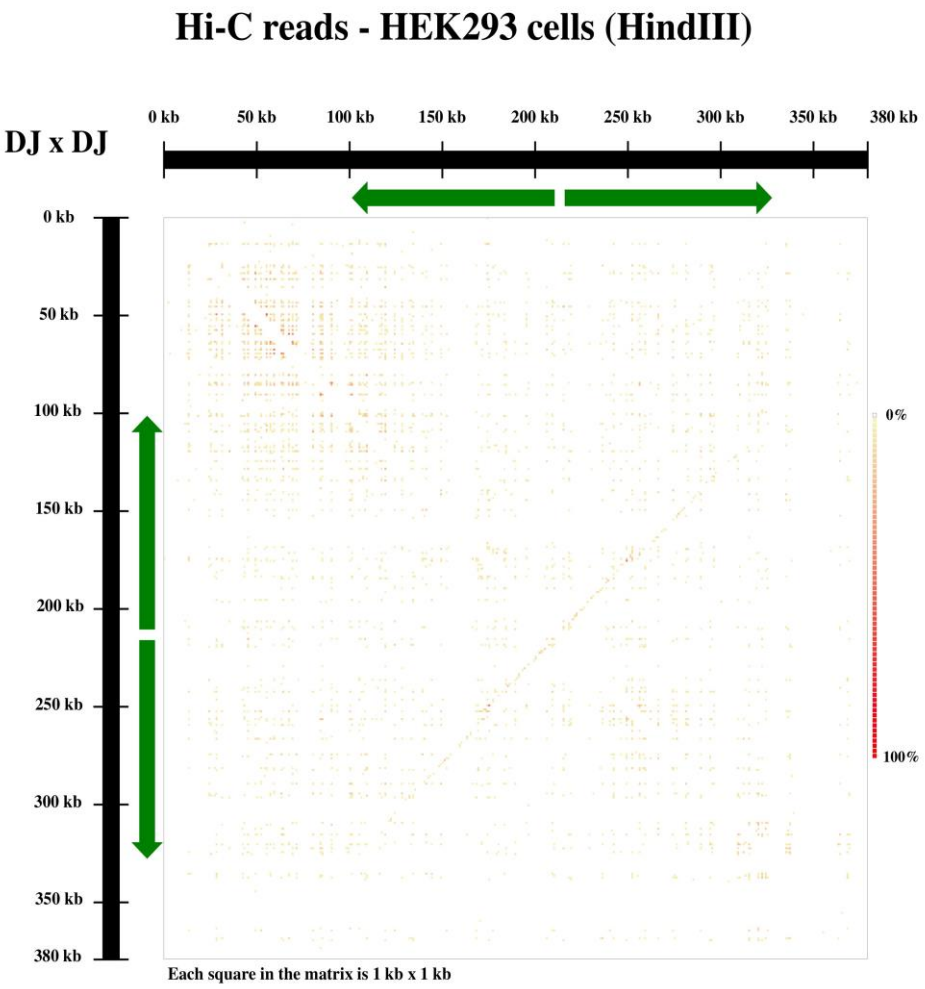

Figure A 4 - Intramolecular interactions in the DJ using Hi-C reads from HEK293 cells in normal conditions. HindIII (A'AGCTT) restriction enzyme was used. The two arms of the inverted repeat present in the DJ show contact with each other. Scale from no observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

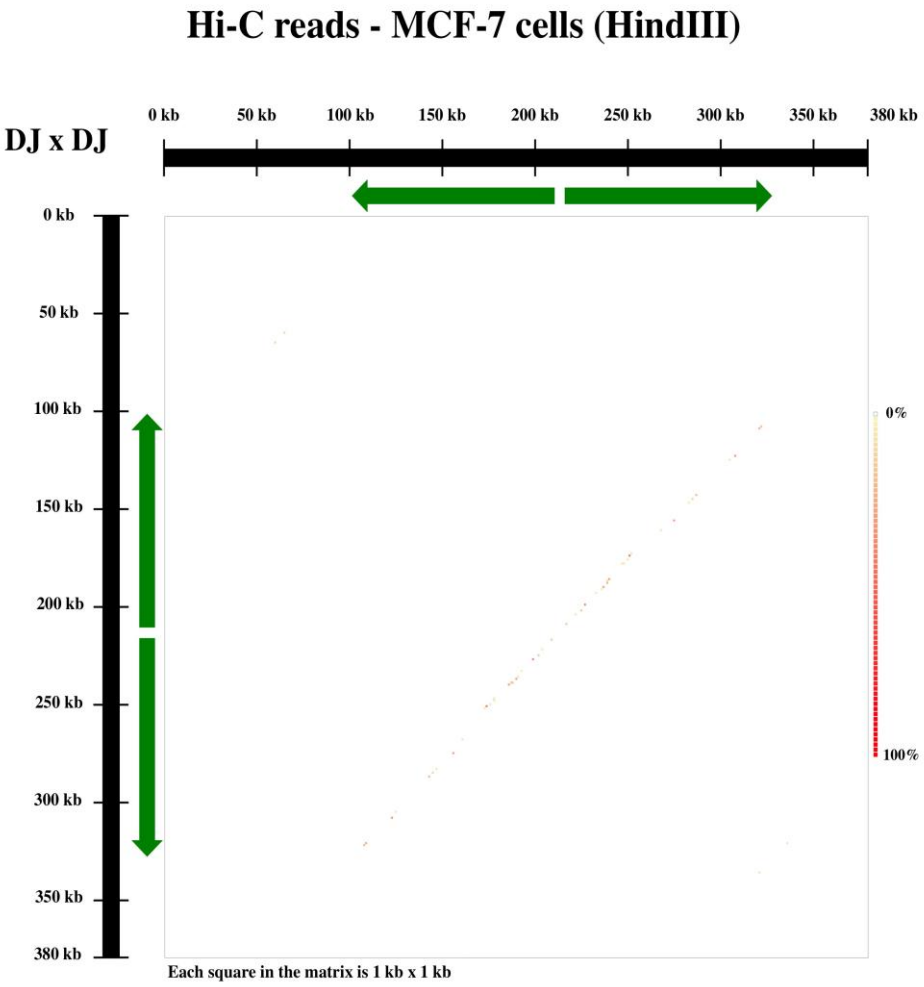

Figure A 5 - Intramolecular interactions in the DJ using Hi-C reads from MCF-7 cells in normal conditions. HindIII (A'AGCTT) restriction enzyme was used. The two arms of the inverted repeat present in the DJ show contact with each other. Scale from no observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

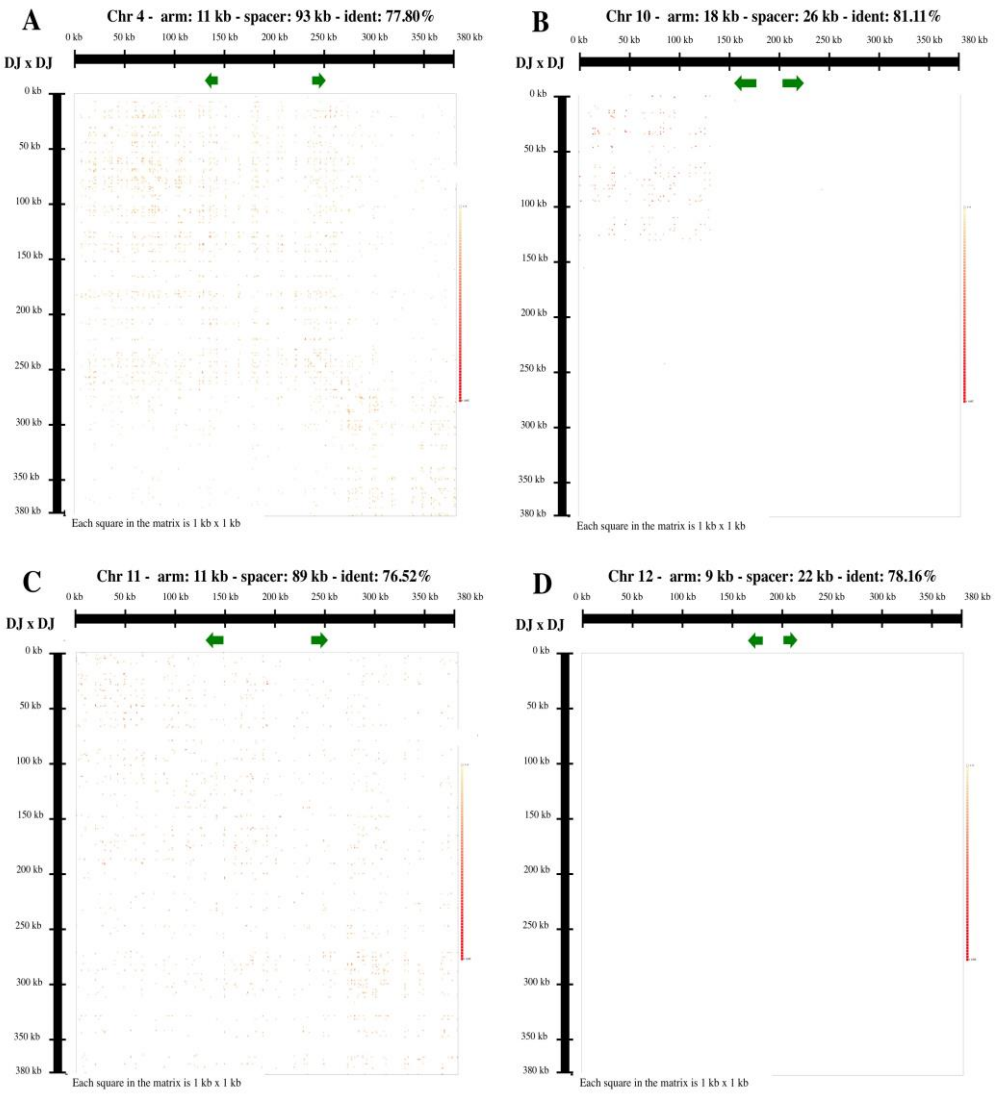

Figure A 6 - Intrachromosomal interactions in 4 inverted repeats found in chromosomes 4, 10, 11 and 12. Hi-C data set GSE43070 was mapped against human genome reference hg17. Unlike the DJ, the arms from each repeat do not interact with each other. Green arrows indicate the position of the large inverted repeats. Scale from no observed interactions to highest observed number of interaction in all normalised data sets. Each square in the interaction matrix is 1 kb x 1 kb.

Appendix B Sequenced clones from AL591856

Table B 1 – Sequenced cDNA clones from AL591856

AL591856 cDNA	Sequence
Clone 2	GAATTCGATTGCAGGAAGAAGTCTCTCATC AGGTACCAGATGACAGTGCCTTGATCCTGA ACTTCCCAGCTTCCAGAACAGAGATTTATT CATTGTTTGCTGATCGTGAAAATGCACAGA GCACTCTGGAAGCATAGCTGTGATAACAGT CGAATGGAATGGGTTAAGGAGAGAACTGAT GTGGCTTGGTCCTAGATTCTCCTGCCCAAA ATGCTCTACACAATAGCTGGAAAAGACTGC AGGGACAAAGACCCGAACCCCAGAGCTTCC AGAACAGCCTCTTATTGCAGCAGGAGCCGG GATTCTCAATCACTAGTGAATTC
Clone 3	GAATTCGATTGCAGGAAGAAGTCTCTCATC AGGTACCAGATGACAGTGCCTTGATCCTGG ACTTCCCAGCTTCCAGAACAGAGATTTATT CATTGTTTGCTGATCGTGAAAATGCACAGA GCACTCTGGAAGCATAGCTGTGATAACAGT CGAATGGAGTGGGTTAAGGAGAGAACTGAT GTGGCTTGGTCCTAGATTCTCCTGCCCAAA ATGTTCTACACAATAGCTGGAAAAGACTGC AGGGACAAAGACCCGAACCCCAGAGCTTCC AGAACAGCCTCTTATTGCAGCAGGAGCCGG GATTCTCAATCACTAGTGAATTC
Clone 4	GAATTCGATTGAGAATCCCGGCTCCTGCTG CAATAAGAGGCTGTTCTGGAAGCTCTGGGG TTCGGGTCTTTGTCCCTGCAGCTATGCTTC CAGAGTGCTCTGTGCATTTTCACGATCAGC AAACAATGAATAAATCTCTGTTCTGGAAGC TGGGAAGTCCAGGATCAAGGCACTGTCATC TGGTACCTGATGAGAGACTTCTTCCTGCAA TCACTAGTGAATTC
Clone 5	GAATTCGATTGAGAATCCCGGCTCCTGCTG CAATAAGAGGCTGTTCTGGAAGCTCTGGGG TTCGGGTCTTTGTCCCTGCAGCTATGCTTC CAGAGTGCTCTGTGCATTTTCACGATCAGC AAACAATGAATAAATCTCTGTTCTGGAAGC TGGGAAGTCCAGGATCAAGGCACTGTCATC TGGTACCTGATGAGAGACTTCTTCCTGCAA TCACTAGTGAATTC

Appendix C AL591856 Chromatin profile figures

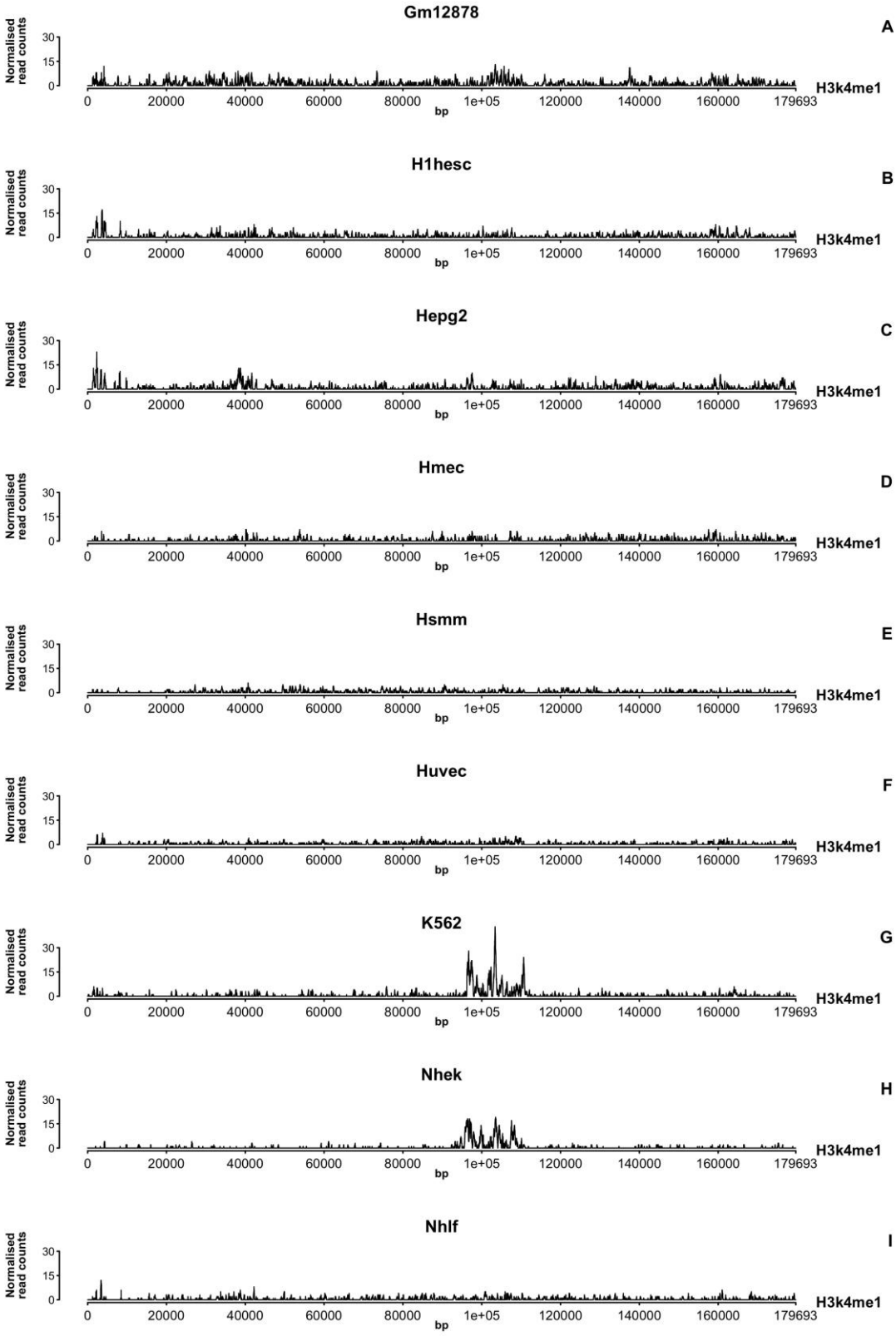

Figure C 1 ChIP-seq peaks for histone modification H3K4me1 for BAC AL591856. Cell lines are indicated above each graph.

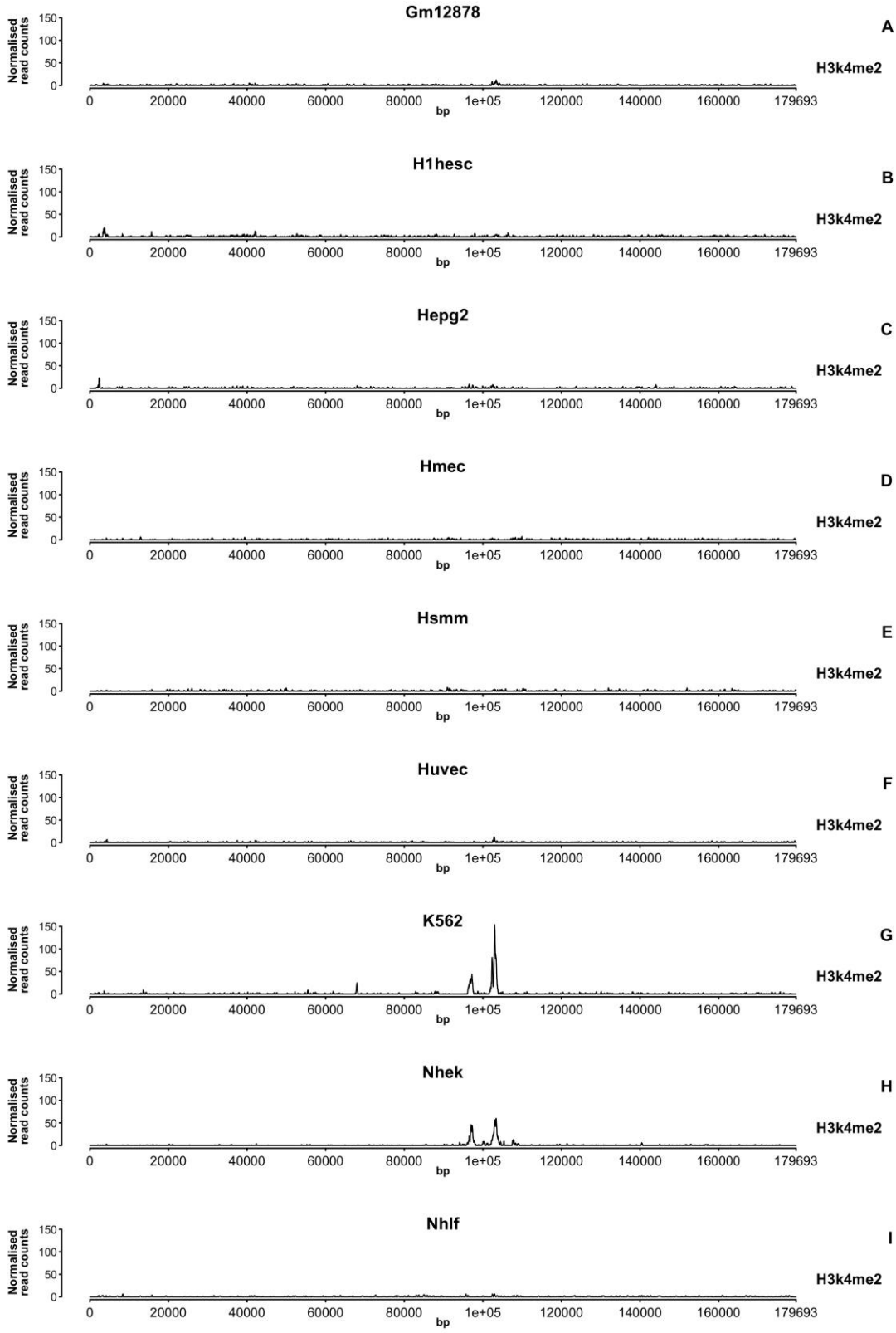

Figure C 2 - ChIP-seq peaks for histone modification H3K4me2 for BAC AL591856. Cell lines are indicated above each graph.

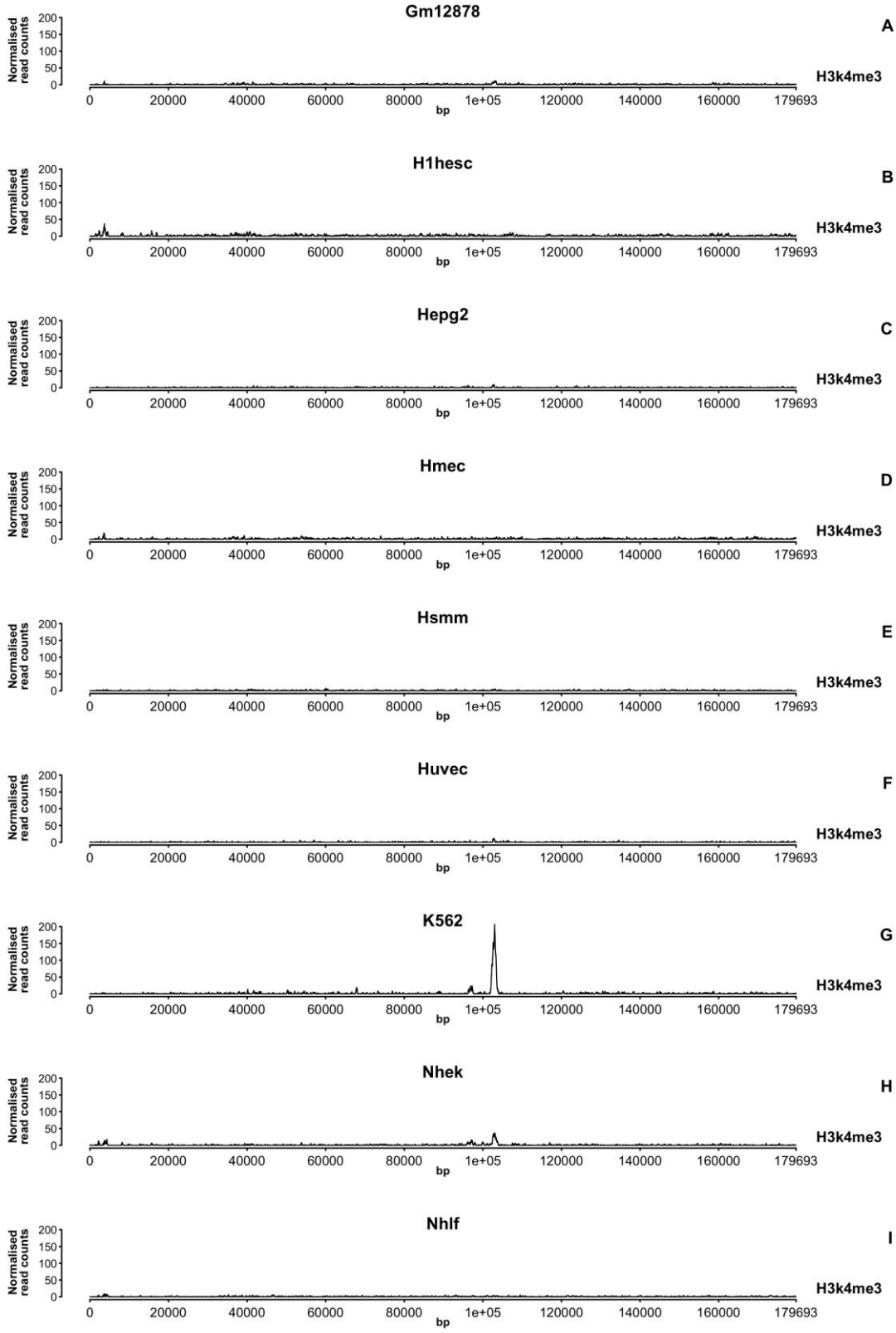

Figure C 3 - ChIP-seq peaks for histone modification H3K4me3 for BAC AL591856. Cell lines are indicated above each graph.

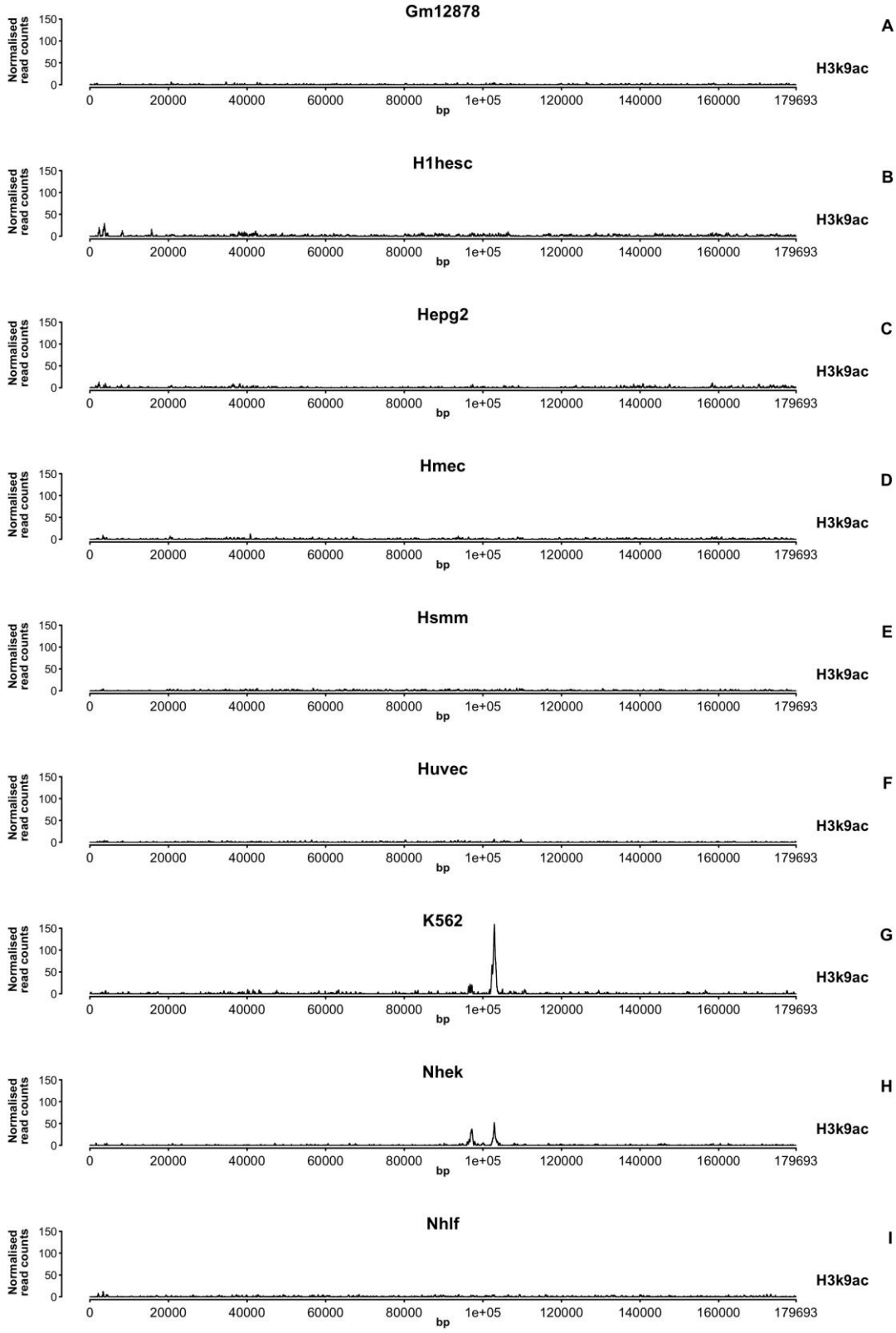

Figure C 4 ChIP-seq peaks for histone modification H3K9ac for BAC AL591856. Cell lines are indicated above each graph.

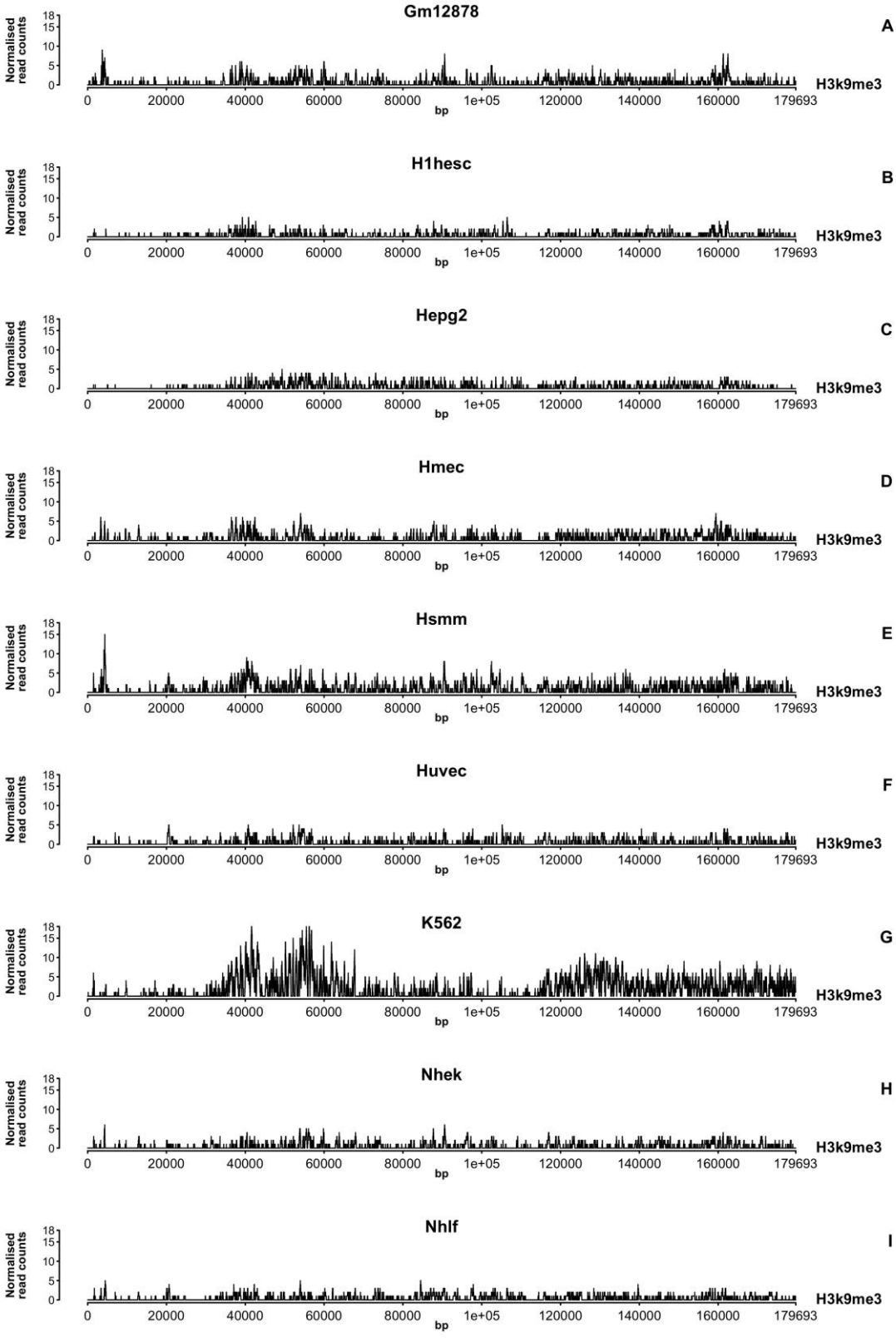

Figure C 5 - ChIP-seq peaks for histone modification H3K9me3 for BAC AL591856. Cell lines are indicated above each graph.

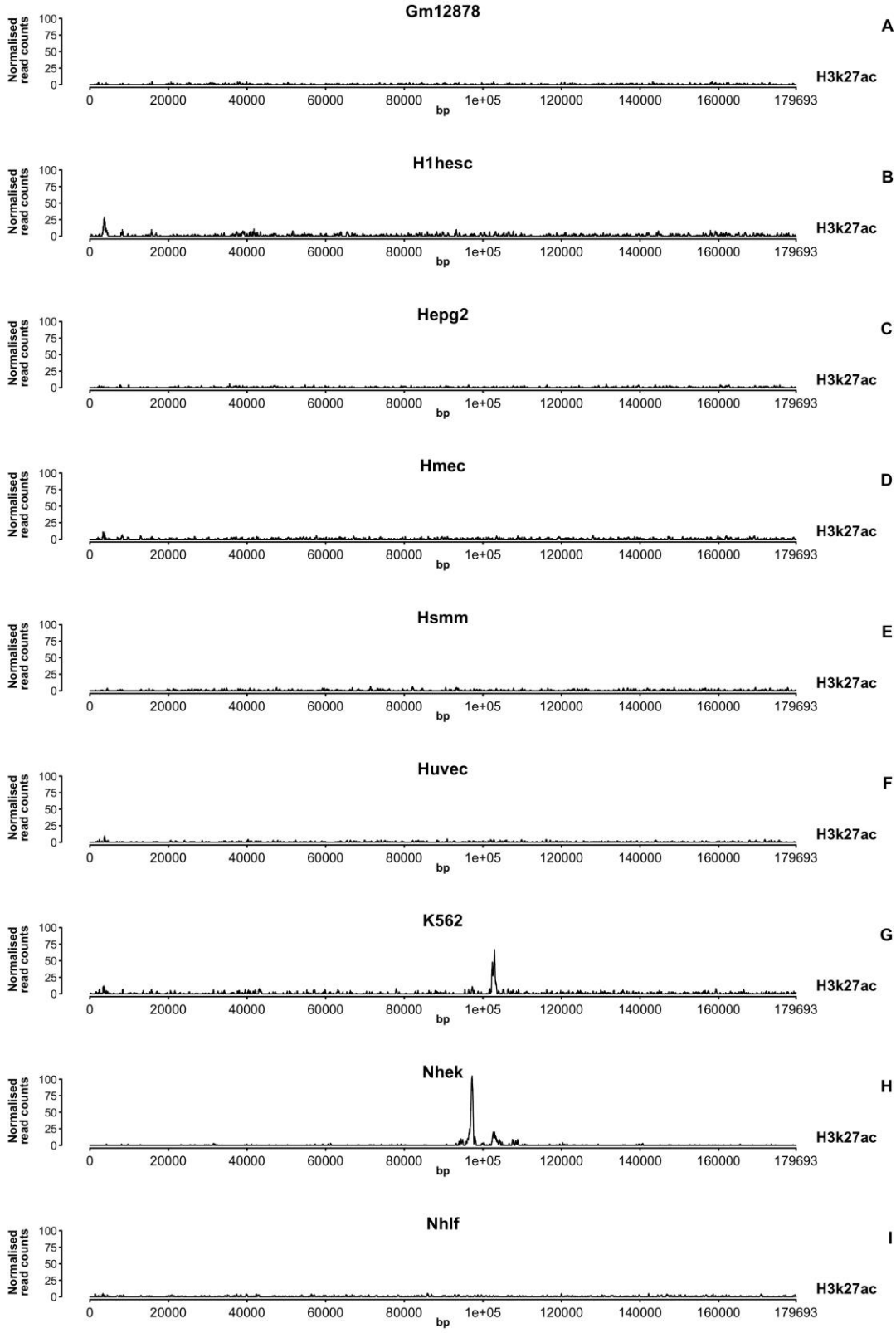

Figure C 6 - ChIP-seq peaks for histone modification H3K27ac for BAC AL591856. Cell lines are indicated above each graph.

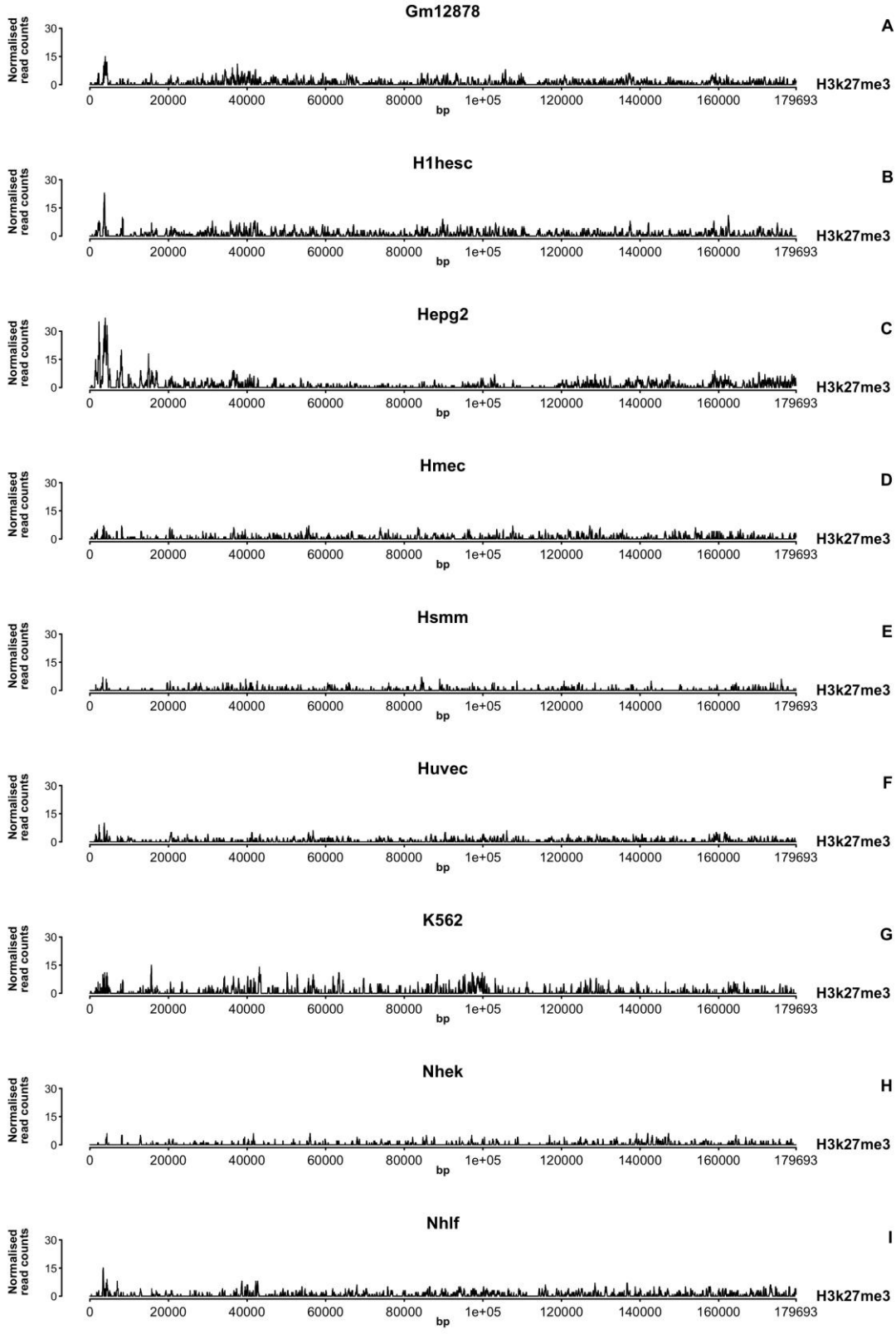

Figure C 7 - ChIP-seq peaks for histone modification H3K27me3 for BAC AL591856. Cell lines are indicated above each graph.

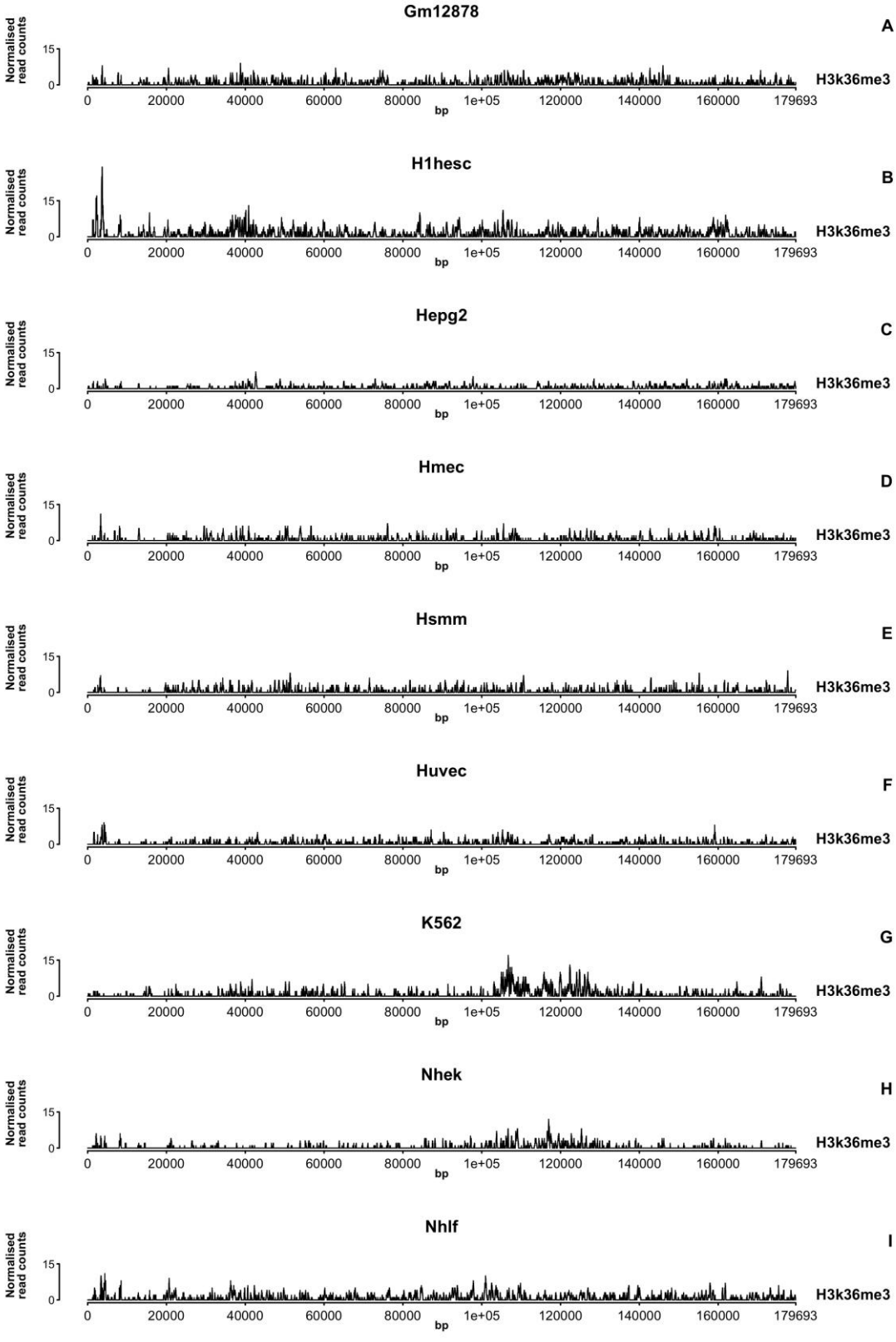

Figure C 8 - ChIP-seq peaks for histone modification H3K36me3 for BAC AL591856. Cell lines are indicated above each graph.

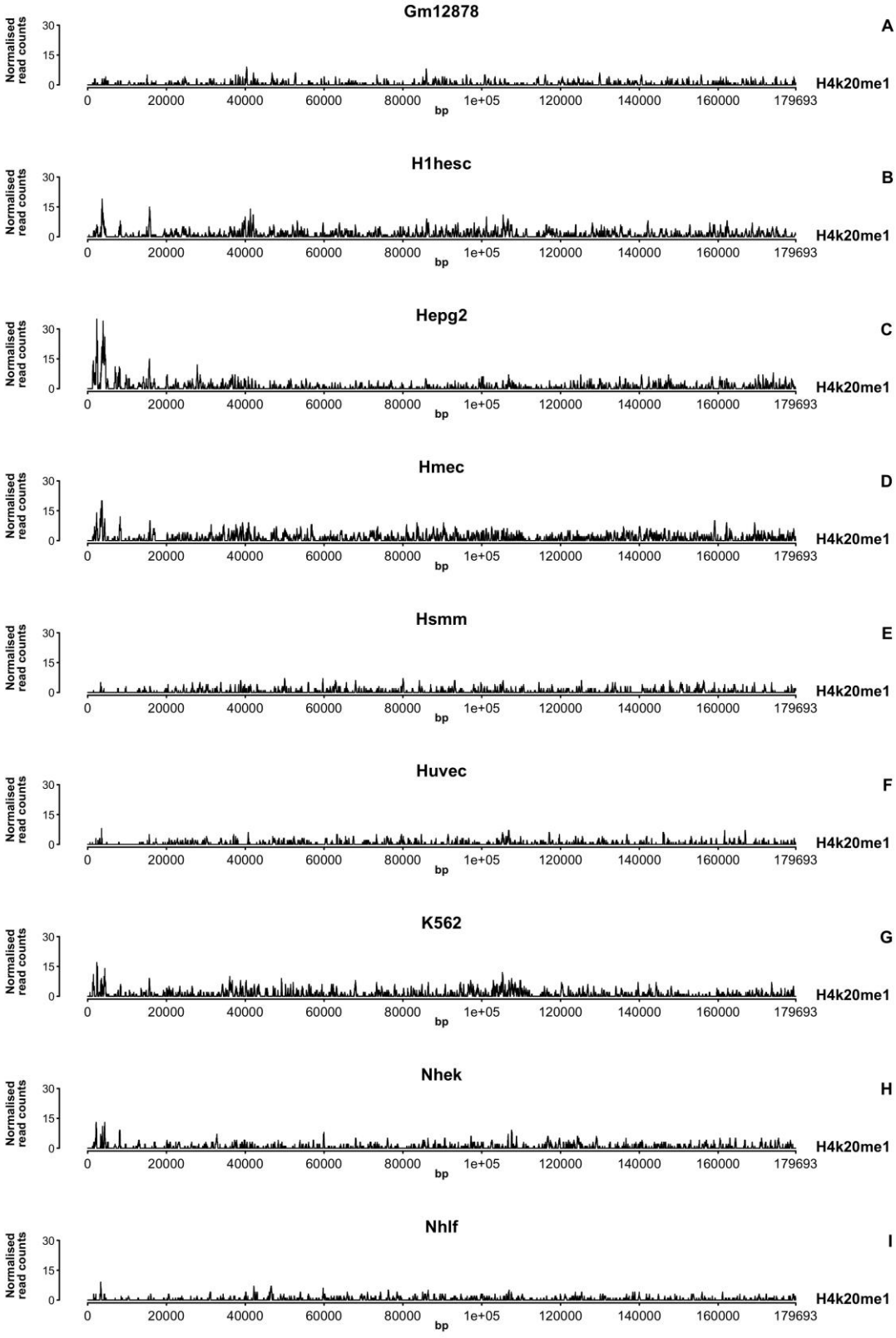

Figure C 9 - ChIP-seq peaks for histone modification H4K20me1 for BAC AL591856. Cell lines are indicated above each graph.

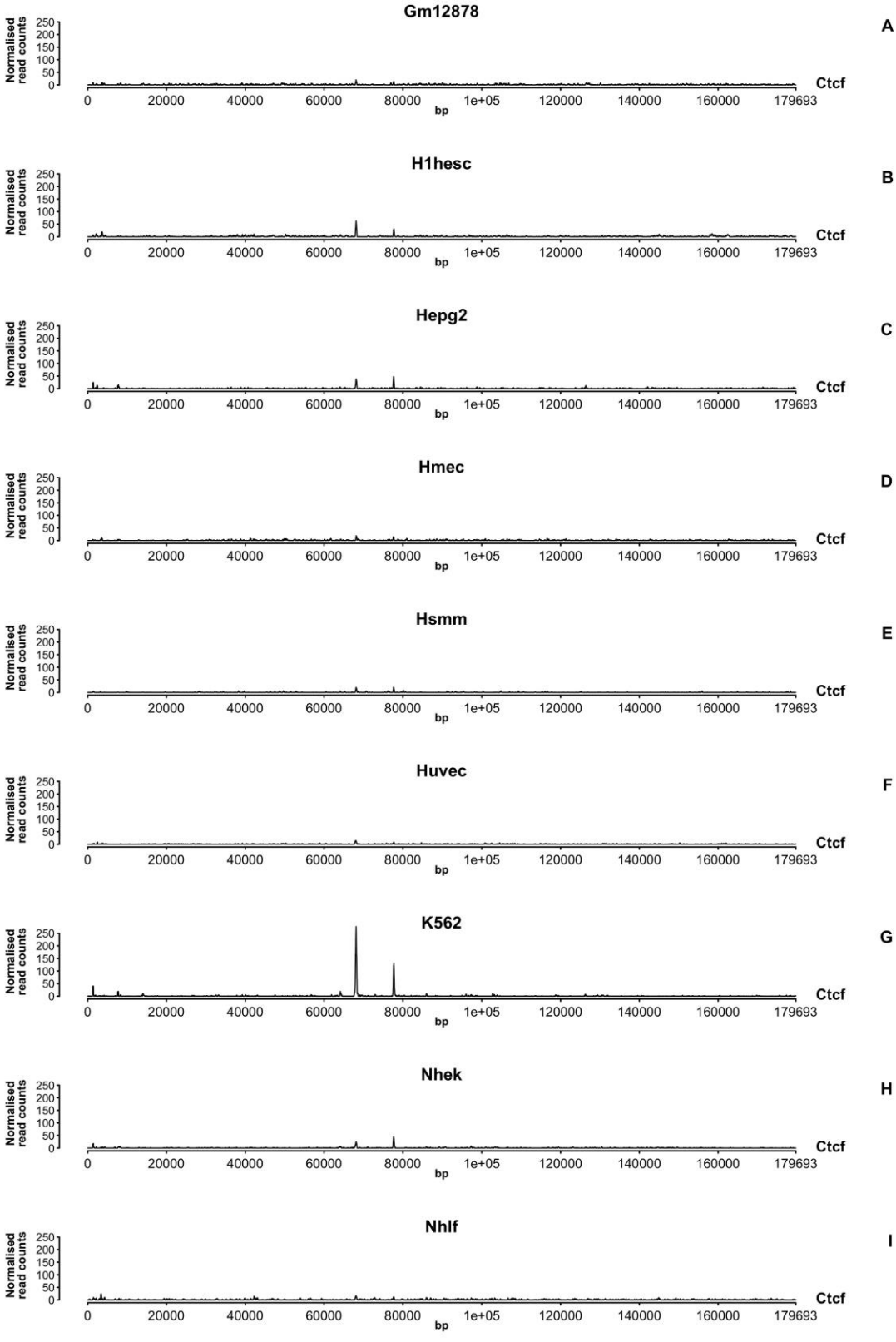

Figure C 10 - ChIP-seq peaks for histone modification CTCF for BAC AL591856. Cell lines are indicated above each graph.

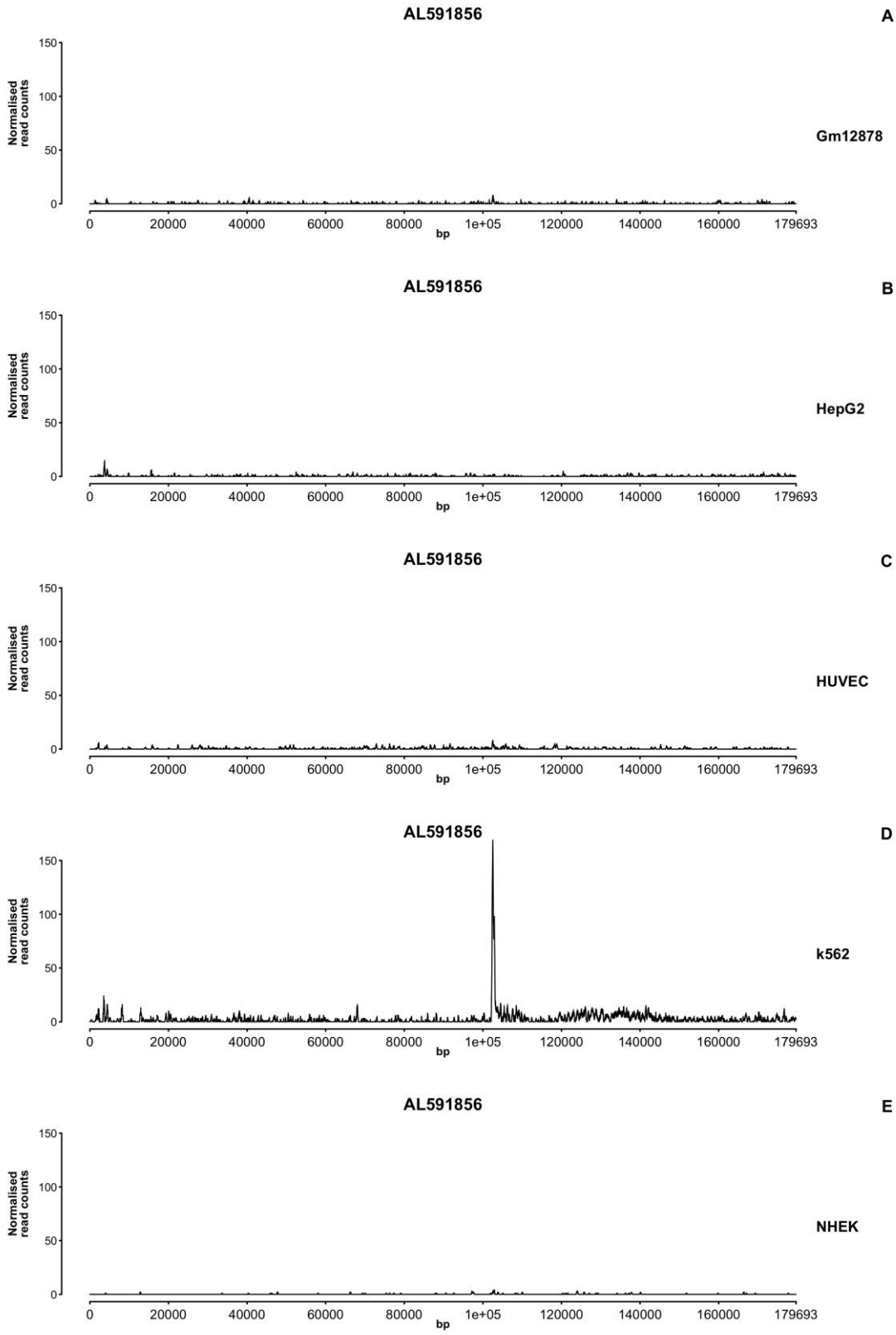

Figure C 11 - ChIP-seq peaks for histone modification Pol II for BAC AL591856. Cell lines are indicated above each graph.

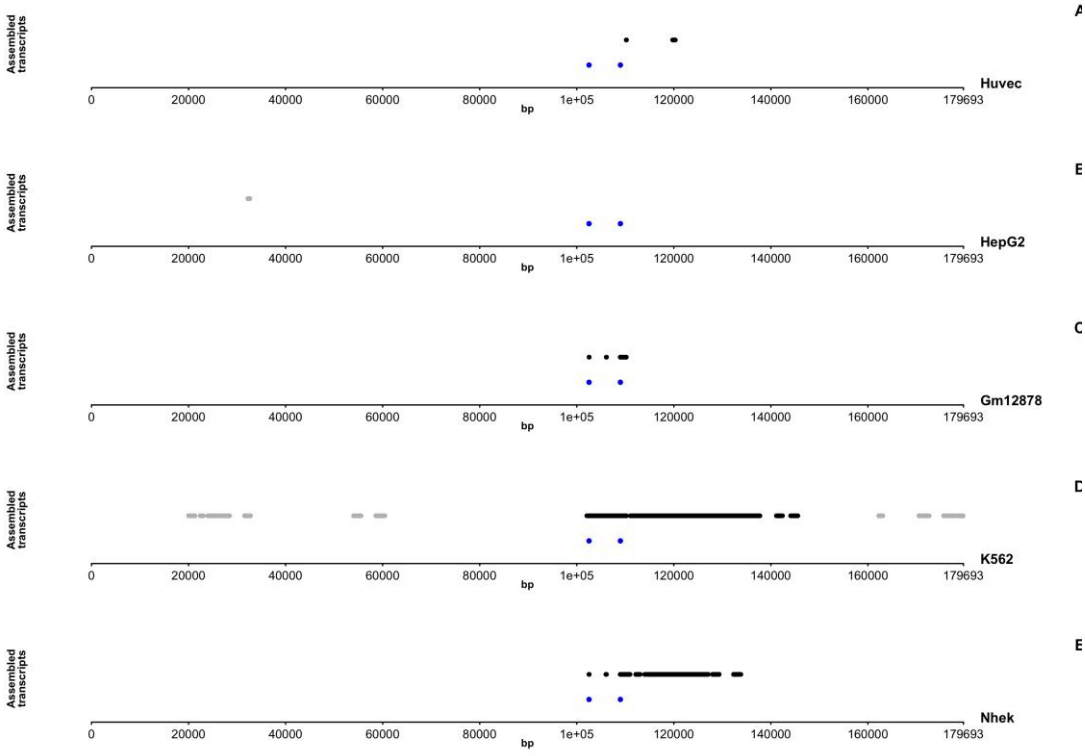

Figure C 12 - RNA-seq assembled transcripts from AL591856.

Bibliography

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F.*, et al.* (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science (New York, NY) *252*, 1651-1656.

Adekoya, E., Ait-Zahra, M., Allen, N., Anderson, M., Anderson, S., Anufriev, F., Ambruster, J., Ayele, K., Baker, J., Baldwin, J.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome biology *12*, R18.

Albiez, H., Cremer, M., Tiberi, C., Vecchio, L., Schermelleh, L., Dittrich, S., Kupper, K., Joffe, B., Thormeyer, T., von Hase, J.*, et al.* (2006). Chromatin domains and the interchromatin compartment form structurally defined and functionally interacting nuclear networks. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology *14*, 707-733.

Allfrey, V.G., Faulkner, R., and Mirsky, A.E. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. Proceedings of the National Academy of Sciences of the United States of America *51*, 786-794.

Altemose, N., Miga, K.H., Maggioni, M., and Willard, H.F. (2014). Genomic characterization of large heterochromatic gaps in the human genome assembly. PLoS computational biology *10*, e1003628.

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic acids research *9*, 3015-3027.

Athwal, R.S., Searle, B.M., and Jansons, V.K. (1985). Diphtheria toxin sensitivity in a monochromosomal hybrid containing human chromosome 5. The Journal of heredity *76*, 329-334.

Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., and Eichler, E.E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. Genome research *11*, 1005-1017.

Bannister, A.J., Zegerman, P., Partridge, J.F., Miska, E.A., Thomas, J.O., Allshire, R.C., and Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature *410*, 120-124.

Barnett, D.W., Garrison, E.K., Quinlan, A.R., Stromberg, M.P., and Marth, G.T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics (Oxford, England) *27*, 1691-1692.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M.*, et al.* (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic acids research *41*, D991-995.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. Methods (San Diego, Calif) *58*, 268-276.

Benevolenskaya, E.V. (2007). Histone H3K4 demethylases are essential in development and differentiation. Biochemistry and cell biology = Biochimie et biologie cellulaire *85*, 435-443.

Bennett, S. (2004). Solexa Ltd. Pharmacogenomics *5*, 433-438.

Bensimon, A., Simon, A., Chiffaudel, A., Croquette, V., Heslot, F., and Bensimon, D. (1994). Alignment and sensitive detection of DNA by a moving interface. Science (New York, NY) *265*, 2096-2098.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R.*, et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53-59.

Bernhard, W., and Granboulan, N. (1963). The fine structure of the cancer cell nucleus. Experimental cell research *24*, SUPPL9:19-53.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K.*, et al.* (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Bickmore, W.A., and van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. Cell *152*, 1270-1284.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E.*, et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

Blobel, G. (1985). Gene gating: a hypothesis. Proceedings of the National Academy of Sciences of the United States of America *82*, 8527-8529.

Bloom, S.E., and Goodpasture, C. (1976). An improved technique for selective silver staining of nucleolar organizer regions in human chromosomes. Human genetics *34*, 199-206.

Bobrow, M., Joness, L.F., and Clarke, G. (1971). A complex chromosomal rearrangement with formation of a ring 4. Journal of medical genetics *8*, 235-239.

Bodnar, A.G., Ouellette, M., Frolkis, M., Holt, S.E., Chiu, C.P., Morin, G.B., Harley, C.B., Shay, J.W., Lichtsteiner, S., and Wright, W.E. (1998). Extension of life-span by introduction of telomerase into normal human cells. Science (New York, NY) *279*, 349-352.

Boisvert, F.M., van Koningsbruggen, S., Navascues, J., and Lamond, A.I. (2007). The multifunctional nucleolus. Nature reviews Molecular cell biology *8*, 574-585.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) *30*, 2114-2120.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Muller, S., Eils, R., Cremer, C., Speicher, M.R.*, et al.* (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS biology *3*, e157.

Boulon, S., Westman, B.J., Hutten, S., Boisvert, F.M., and Lamond, A.I. (2010). The nucleolus under stress. Molecular cell *40*, 216-227.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K.*, et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.

Branco, M.R., and Pombo, A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS biology *4*, e138.

Brown, D.D., and Gurdon, J.B. (1964). Asence of ribosomal RNA synthesis in the anucleolate mutant Xenopus Laevis. Proceedings of the National Academy of Sciences of the United States of America *51*, 139-146.

Burgess, R.C., Burman, B., Kruhlak, M.J., and Misteli, T. (2014). Activation of DNA damage response signaling by condensed chromatin. Cell reports *9*, 1703-1717.

Burke, D.T., Carle, G.F., and Olson, M.V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science (New York, NY) *236*, 806-812.

Caburet, S., Conti, C., Schurra, C., Lebofsky, R., Edelstein, S.J., and Bensimon, A. (2005). Human ribosomal RNA gene arrays display a broad range of palindromic structures. Genome research *15*, 1079-1085.

Carter, D., Chakalova, L., Osborne, C.S., Dai, Y.F., and Fraser, P. (2002). Long-range chromatin regulatory interactions in vivo. Nature genetics *32*, 623-626.

Catasti, P., Chen, X., Mariappan, S.V., Bradbury, E.M., and Gupta, G. (1999). DNA repeats in the human genome. Genetica *106*, 15-36.

Chaisson, M.J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC bioinformatics *13*, 238.

Chao, S.H., and Price, D.H. (2001). Flavopiridol inactivates P-TEFb and blocks most RNA polymerase II transcription in vivo. The Journal of biological chemistry *276*, 31793-31799.

Chen, H., Chen, J., Muir, L.A., Ronquist, S., Meixner, W., Ljungman, M., Ried, T., Smale, S., and Rajapakse, I. (2015). Functional organization of the human 4D Nucleome. Proceedings of the National Academy of Sciences of the United States of America *112*, 8002-8007.

Chen, Y.C., Liu, T., Yu, C.H., Chiang, T.Y., and Hwang, C.C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PloS one *8*, e62856.

Choo, K.H., Earle, E., Vissel, B., and Filby, R.G. (1990). Identification of two distinct subfamilies of alpha satellite DNA that are highly specific for human chromosome 15. Genomics *7*, 143-151.

Cohen, D., Chumakov, I., and Weissenbach, J. (1993). A first-generation physical map of the human genome. Nature *366*, 698-701.

Consortium, E.P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science (New York, NY) *306*, 636-640.

Cremer, T., and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature reviews Genetics *2*, 292-301.

Cremer, T., and Cremer, M. (2010). Chromosome territories. Cold Spring Harbor perspectives in biology *2*, a003889.

Cremer, T., Cremer, M., Dietzel, S., Muller, S., Solovei, I., and Fakan, S. (2006). Chromosome territories--a functional nuclear landscape. Current opinion in cell biology *18*, 307-316.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A.*, et al.* (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America *107*, 21931-21936.

Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K., and Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome research *19*, 24-32.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S.*, et al.* (2015). Ensembl 2015. Nucleic acids research *43*, D662-669.

Cuthbert, A.P., Trott, D.A., Ekong, R.M., Jezzard, S., England, N.L., Themis, M., Todd, C.M., and Newbold, R.F. (1995). Construction and characterization of a highly stable human: rodent monochromosomal hybrid panel for genetic complementation and genome mapping studies. Cytogenetics and cell genetics *71*, 68-76.

Dammann, R., Lucchini, R., Koller, T., and Sogo, J.M. (1995). Transcription in the yeast rRNA gene locus: distribution of the active gene copies and chromatin structure of their flanking regulatory sequences. Molecular and cellular biology *15*, 5294-5303.

Daujat, S., Zeissler, U., Waldmann, T., Happel, N., and Schneider, R. (2005). HP1 binds specifically to Lys26-methylated histone H1.4, whereas simultaneous Ser27 phosphorylation blocks HP1 binding. The Journal of biological chemistry *280*, 38090-38095.

Dawson, M.A., Bannister, A.J., Gottgens, B., Foster, S.D., Bartke, T., Green, A.R., and Kouzarides, T. (2009). JAK2 phosphorylates histone H3Y41 and excludes HP1alpha from chromatin. Nature *461*, 819-822.

de Koning, A.P., Gu, W., Castoe, T.A., Batzer, M.A., and Pollock, D.D. (2011). Repetitive elements may comprise over two-thirds of the human genome. PLoS genetics *7*, e1002384.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science (New York, NY) *295*, 1306-1311.

Derenzini, M., Trere, D., Pession, A., Montanaro, L., Sirri, V., and Ochs, R.L. (1998). Nucleolar function and size in cancer cells. The American journal of pathology *152*, 1291-1297.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R., Lander, E.S.*, et al.* (1987). A genetic linkage map of the human genome. Cell *51*, 319-337.

Dostie, J., and Dekker, J. (2007). Mapping networks of physical interactions between genomic elements using 5C technology. Nature protocols *2*, 988-1002.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., *et al.* (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome research *16*, 1299-1309.

Duitama, J., Zablotskaya, A., Gemayel, R., Jansen, A., Belet, S., Vermeesch, J.R., Verstrepen, K.J., and Froyen, G. (2014). Large-scale analysis of tandem repeat variability in the human genome. Nucleic acids research *42*, 5728-5741.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research *30*, 207-210.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., *et al.* (2009). Real-time DNA sequencing from single polymerase molecules. Science (New York, NY) *323*, 133-138.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., *et al.* (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43-49.

Finch, J.T., Lutter, L.C., Rhodes, D., Brown, R.S., Rushton, B., Levitt, M., and Klug, A. (1977). Structure of nucleosome core particles of chromatin. Nature *269*, 29-36.

Finlan, L.E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J.R., and Bickmore, W.A. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS genetics *4*, e1000039.

Floutsakou, I., Agrawal, S., Nguyen, T.T., Seoighe, C., Ganley, A.R., and McStay, B. (2013). The shared genomic architecture of human nucleolar organizer regions. Genome research *23*, 2003-2012.

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., *et al.* (2006). Copy number variation: new insights in genome diversity. Genome research *16*, 949-961.

Gebrane-Younes, J., Fomproix, N., and Hernandez-Verdun, D. (1997). When rDNA transcription is arrested during mitosis, UBF is still associated with non-condensed rDNA. Journal of cell science *110 (Pt 19)*, 2429-2440.

Gheldof, N., Smith, E.M., Tabuchi, T.M., Koch, C.M., Dunham, I., Stamatoyannopoulos, J.A., and Dekker, J. (2010). Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. Nucleic acids research *38*, 4325-4336.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., *et al.* (2005). Galaxy: a

platform for interactive large-scale genome analysis. Genome research *15*, 1451-1455.

Gilbert, N., Gilchrist, S., and Bickmore, W.A. (2005). Chromatin organization in the mammalian nucleus. International review of cytology *242*, 283-336.

Giorgetti, L., Galupa, R., Nora, E.P., Piolot, T., Lam, F., Dekker, J., Tiana, G., and Heard, E. (2014). Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell *157*, 950-963.

Gonzalez, I.L., and Sylvester, J.E. (1995). Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. Genomics *27*, 320-328.

Gonzalez, I.L., Wu, S., Li, W.M., Kuo, B.A., and Sylvester, J.E. (1992). Human ribosomal RNA intergenic spacer sequence. Nucleic acids research *20*, 5846.

Gosalia, N., Neems, D., Kerschner, J.L., Kosak, S.T., and Harris, A. (2014). Architectural proteins CTCF and cohesin have distinct roles in modulating the higher order structure and expression of the CFTR locus. Nucleic acids research *42*, 9612-9622.

Green, E.D., and Olson, M.V. (1990). Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. Proceedings of the National Academy of Sciences of the United States of America *87*, 1213-1217.

Grob, A., Colleran, C., and McStay, B. (2014). Construction of synthetic nucleoli in human cells reveals how a major functional nuclear domain is formed and propagated through cell division. Genes & development *28*, 220-230.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W.*, et al.* (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948-951.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y.*, et al.* (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900-910.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P.*, et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.

Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., Marc, S., Bernardi, G., Lathrop, M., and Weissenbach, J. (1994). The 1993-94 Genethon human genetic linkage map. Nature genetics *7*, 246-339.

Hadjur, S., Williams, L.M., Ryan, N.K., Cobb, B.S., Sexton, T., Fraser, P., Fisher, A.G., and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature *460*, 410-413.

Han, H.J., Russo, J., Kohwi, Y., and Kohwi-Shigematsu, T. (2008). SATB1 reprogrammes gene expression to promote breast tumour growth and metastasis. Nature *452*, 187-193.

Han, M., and Grunstein, M. (1988). Nucleosome loss activates yeast downstream promoters in vivo. Cell *55*, 1137-1145.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674.

Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H., Yada, T., Park, H.S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D.K.*, et al.* (2000). The DNA sequence of human chromosome 21. Nature *405*, 311-319.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A.*, et al.* (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics *39*, 311-318.

Henderson, A.S., Warburton, D., and Atwood, K.C. (1972). Location of ribosomal DNA in the human chromosome complement. Proceedings of the National Academy of Sciences of the United States of America *69*, 3394-3398.

Hernandez-Verdun, D. (2011). Assembly and disassembly of the nucleolus during the cell cycle. Nucleus (Austin, Tex) *2*, 189-194.

Hilliker, A.J., and Appels, R. (1989). The arrangement of interphase chromosomes: structural and functional aspects. Experimental cell research *185*, 267-318.

Hirose, Y., and Manley, J.L. (1998). RNA polymerase II is an essential mRNA polyadenylation factor. Nature *395*, 93-96.

Hommelsheim, C.M., Frantzeskakis, L., Huang, M., and Ulker, B. (2014). PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications. Scientific reports *4*, 5052.

Hubner, B., Cremer, T., and Neumann, J. (2013). Correlative microscopy of individual cells: sequential application of microscopic systems with increasing resolution to study the nuclear landscape. Methods in molecular biology (Clifton, NJ) *1042*, 299-336.

Hudson, T.J., Stein, L.D., Gerety, S.S., Ma, J., Castle, A.B., Silva, J., Slonim, D.K., Baptista, R., Kruglyak, L., Xu, S.H.*, et al.* (1995). An STS-based map of the human genome. Science (New York, NY) *270*, 1945-1954.

Inoue, J., Mitsuya, K., Maegawa, S., Kugoh, H., Kadota, M., Okamura, D., Shinohara, T., Nishihara, S., Takehara, S., Yamauchi, K.*, et al.* (2001).

Construction of 700 human/mouse A9 monochromosomal hybrids and analysis of imprinted genes on human chromosome 6. Journal of human genetics *46*, 137-145.

Jacobs, P.A., Hunt, P.A., Mayer, M., and Bart, R.D. (1981). Duchenne muscular dystrophy (DMD) in a female with an X/autosome translocation: further evidence that the DMD locus is at Xp21. American journal of human genetics *33*, 513-518.

Jager, R., Migliorini, G., Henrion, M., Kandaswamy, R., Speedy, H.E., Heindl, A., Whiffin, N., Carnicer, M.J., Broome, L., Dryden, N.*, et al.* (2015). Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nature communications *6*, 6178.

Jantzen, H.M., Admon, A., Bell, S.P., and Tjian, R. (1990). Nucleolar transcription factor hUBF contains a DNA-binding motif with homology to HMG proteins. Nature *344*, 830-836.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290-294.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science (New York, NY) *316*, 1497-1502.

Jones, K.W., and Corneo, G. (1971). Location of satellite and homogeneous DNA sequences on human chromosomes. Nature: New biology *233*, 268-271.

Joshi, A.A., and Struhl, K. (2005). Eaf3 chromodomain interaction with methylated H3-K36 links histone deacetylation to Pol II elongation. Molecular cell *20*, 971-978.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research *110*, 462-467.

Kedinger, C., Gniazdowski, M., Mandel, J.L., Jr., Gissinger, F., and Chambon, P. (1970). Alpha-amanitin: a specific inhibitor of one of two DNA-pendent RNA polymerase activities from calf thymus. Biochemical and biophysical research communications *38*, 165-171.

Kim, K., Kim, J.M., Kim, J.S., Choi, J., Lee, Y.S., Neamati, N., Song, J.S., Heo, K., and An, W. (2013). VprBP has intrinsic kinase activity targeting histone H2A and represses gene transcription. Molecular cell *52*, 459-467.

Kim, S.S., Jung, S.C., Kim, H.J., Moon, H.R., and Lee, J.S. (1999). Chromosome abnormalities in a referred population for suspected chromosomal aberrations: a report of 4117 cases. Journal of Korean medical science *14*, 373-376.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the

vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231-1245.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. Nature *436*, 876-880.

Kind, J., Pagie, L., Ortabozkoyun, H., Boyle, S., de Vries, S.S., Janssen, H., Amendola, M., Nolen, L.D., Bickmore, W.A., and van Steensel, B. (2013). Single-cell dynamics of genome-nuclear lamina interactions. Cell *153*, 178-192.

Kleinjan, D.A., and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. American journal of human genetics *76*, 8-32.

Koberna, K., Malinsky, J., Pliss, A., Masata, M., Vecerova, J., Fialova, M., Bednar, J., and Raska, I. (2002). Ribosomal genes in focus: new transcripts label the dense fibrillar components and form clusters indicative of "Christmas trees" in situ. The Journal of cell biology *157*, 743-748.

Koohy, H., Down, T.A., and Hubbard, T.J. (2013). Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme. PloS one *8*, e69853.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L.*, et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science (New York, NY) *318*, 420-426.

Kornberg, R.D. (1974). Chromatin structure: a repeating unit of histones and DNA. Science (New York, NY) *184*, 868-871.

Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., and Turner, D.J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nature methods *6*, 291-295.

Kurz, A., Lampel, S., Nickolenko, J.E., Bradl, J., Benner, A., Zirbel, R.M., Cremer, T., and Lichter, P. (1996). Active and inactive genes localize preferentially in the periphery of chromosome territories. The Journal of cell biology *135*, 1195-1205.

Kustatscher, G., Hegarat, N., Wills, K.L., Furlan, C., Bukowski-Wills, J.C., Hochegger, H., and Rappsilber, J. (2014). Proteomics of a fuzzy organelle: interphase chromatin. The EMBO journal *33*, 648-664.

Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M.*, et al.* (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nature biotechnology *30*, 771-776.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Landry, J.J., Pyl, P.T., Rausch, T., Zichner, T., Tekkedil, M.M., Stutz, A.M., Jauch, A., Aiyar, R.S., Pau, G., Delhomme, N.*, et al.* (2013). The genomic and transcriptomic landscape of a HeLa cell line. G3 (Bethesda, Md) *3*, 1213-1224.

Lawrence, R.J., and Pikaard, C.S. (2004). Chromatin turn ons and turn offs of ribosomal RNA genes. Cell cycle (Georgetown, Tex) *3*, 880-883.

Lee, W.P., Stromberg, M.P., Ward, A., Stewart, C., Garrison, E.P., and Marth, G.T. (2014). MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PloS one *9*, e90581.

Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G., and Webb, W.W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. Science (New York, NY) *299*, 682-686.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707-719.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) *25*, 2078-2079.

Libby, R.T., Hagerman, K.A., Pineda, V.V., Lau, R., Cho, D.H., Baccam, S.L., Axford, M.M., Cleary, J.D., Moore, J.M., Sopher, B.L.*, et al.* (2008). CTCF cis-regulates trinucleotide repeat instability in an epigenetic manner: a novel basis for mutational hot spot determination. PLoS genetics *4*, e1000257.

Lichter, P., Cremer, T., Borden, J., Manuelidis, L., and Ward, D.C. (1988). Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. Human genetics *80*, 224-234.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O.*, et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, NY) *326*, 289-293.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E.*, et al.* (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476-482.

Long, E.O., and Dawid, I.B. (1980). Repeated genes in eukaryotes. Annual review of biochemistry *49*, 727-764.

Lorch, Y., LaPointe, J.W., and Kornberg, R.D. (1987). Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. Cell *49*, 203-210.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251-260.

Lupski, J.R., and Stankiewicz, P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS genetics *1*, e49.

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C.B., Krumm, A.*, et al.* (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. Nature methods *12*, 71-78.

Maegenis, R.E., Donlon, T.A., and Wyandt, H.E. (1978). Giemsa-11 staining of chromosome 1: a newly described heteromorphism. Science (New York, NY) *202*, 64-65.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z.*, et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376-380.

Markaki, Y., Gunkel, M., Schermelleh, L., Beichmanis, S., Neumann, J., Heidemann, M., Leonhardt, H., Eick, D., Cremer, C., and Cremer, T. (2010). Functional nuclear organization of transcription and DNA replication: a topographical marriage between chromatin domains and the interchromatin compartment. Cold Spring Harbor symposia on quantitative biology *75*, 475-492.

Markaki, Y., Smeets, D., Fiedler, S., Schmid, V.J., Schermelleh, L., Cremer, T., and Cremer, M. (2012). The potential of 3D-FISH and super-resolution structured illumination microscopy for studies of 3D nuclear architecture: 3D structured illumination microscopy of defined chromosomal structures visualized by 3D (immuno)-FISH opens new perspectives for studies of nuclear architecture. BioEssays : news and reviews in molecular, cellular and developmental biology *34*, 412-426.

Masson, C., Bouniol, C., Fomproix, N., Szollosi, M.S., Debey, P., and Hernandez-Verdun, D. (1996). Conditions favoring RNA polymerase I transcription in permeabilized cells. Experimental cell research *226*, 114-125.

McClintock, B. (1934). The relation of a particular chromosomal element to the development of the nucleoli in Zea mays. ZZellforsch *21*, 294-326.

McConkey, E.H., and Hopkins, J.W. (1964). The relationship of the nucleolus to the synthesis of ribosomal RNA in HeLa cells. Proceedings of the National Academy of Sciences of the United States of America *51*, 1197-1204.

McCord, R.P., Nazario-Toole, A., Zhang, H., Chines, P.S., Zhan, Y., Erdos, M.R., Collins, F.S., Dekker, J., and Cao, K. (2013). Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford progeria syndrome. Genome research *23*, 260-269.

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S.D., Wickens, M., and Bentley, D.L. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. Nature *385*, 357-361.

McStay, B., and Grummt, I. (2008). The epigenetics of rRNA genes: from molecular to chromosome biology. Annual review of cell and developmental biology *24*, 131-157.

Mehta, G.D., Kumar, R., Srivastava, S., and Ghosh, S.K. (2013). Cohesin: functions beyond sister chromatid cohesion. FEBS letters *587*, 2299-2312.

Mercer, T.R., Edwards, S.L., Clark, M.B., Neph, S.J., Wang, H., Stergachis, A.B., John, S., Sandstrom, R., Li, G., Sandhu, K.S., *et al.* (2013). DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. Nature genetics *45*, 852-859.

Miele, A., and Dekker, J. (2008). Long-range chromosomal interactions and gene regulation. Molecular bioSystems *4*, 1046-1057.

Miga, K.H., Newton, Y., Jain, M., Altemose, N., Willard, H.F., and Kent, W.J. (2014). Centromere reference models for human chromosomes X and Y satellite arrays. Genome research *24*, 697-707.

Miller, J.R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics *95*, 315-327.

Misteli, T. (2010). Higher-order genome organization in human disease. Cold Spring Harbor perspectives in biology *2*, a000794.

Monaco, A.P., and Larin, Z. (1994). YACs, BACs, PACs and MACs: artificial chromosomes as research tools. Trends in biotechnology *12*, 280-286.

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773-777.

Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., and Marra, M. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques *45*, 81-94.

Morrow, J.F., Cohen, S.N., Chang, A.C., Boyer, H.W., Goodman, H.M., and Helling, R.B. (1974). Replication and transcription of eukaryotic DNA in Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America *71*, 1743-1747.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods *5*, 621-628.

Mourad, R., Hsu, P.Y., Juan, L., Shen, C., Koneru, P., Lin, H., Liu, Y., Nephew, K., Huang, T.H., and Li, L. (2014). Estrogen induces global reorganization of chromatin structure in human breast cancer cells. PloS one *9*, e113354.

Muro, E., Gebrane-Younis, J., Jobart-Malfait, A., Louvet, E., Roussel, P., and Hernandez-Verdun, D. (2010). The traffic of proteins between nucleolar organizer regions and prenucleolar bodies governs the assembly of the nucleolus at exit of mitosis. Nucleus (Austin, Tex) *1*, 202-211.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science (New York, NY) *320*, 1344-1349.

Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature *502*, 59-64.

Nemeth, A., Conesa, A., Santoyo-Lopez, J., Medina, I., Montaner, D., Peterfia, B., Solovei, I., Cremer, T., Dopazo, J., and Langst, G. (2010). Initial genomics of the human nucleolus. PLoS genetics *6*, e1000889.

Nicholson, J.M., and Cimini, D. (2013). Cancer karyotypes: survival of the fittest. Frontiers in oncology *3*, 148.

Niedojadlo, J., Perret-Vivancos, C., Kalland, K.H., Cmarko, D., Cremer, T., van Driel, R., and Fakan, S. (2011). Transcribed DNA is preferentially located in the perichromatin region of mammalian cell nuclei. Experimental cell research *317*, 433-444.

Nishino, Y., Eltsov, M., Joti, Y., Ito, K., Takata, H., Takahashi, Y., Hihara, S., Frangakis, A.S., Imamoto, N., Ishikawa, T.*, et al.* (2012). Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. The EMBO journal *31*, 1644-1653.

O'Connor, M., Peifer, M., and Bender, W. (1989). Construction of large DNA segments in Escherichia coli. Science (New York, NY) *244*, 1307-1312.

Olins, A.L., and Olins, D.E. (1974). Spheroid chromatin units (v bodies). Science (New York, NY) *183*, 330-332.

Ong, C.T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature reviews Genetics *12*, 283-293.

Orlando, V. (2000). Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. Trends in biochemical sciences *25*, 99-104.

Osoegawa, K., Mammoser, A.G., Wu, C., Frengen, E., Zeng, C., Catanese, J.J., and de Jong, P.J. (2001). A bacterial artificial chromosome library for sequencing the complete human genome. Genome research *11*, 483-496.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nature reviews Genetics *12*, 87-98.

Parada, L.A., McQueen, P.G., and Misteli, T. (2004). Tissue-specific spatial organization of genomes. Genome biology *5*, R44.

Parada, L.A., McQueen, P.G., Munson, P.J., and Misteli, T. (2002). Conservation of relative chromosome positioning in normal and cancer cells. Current biology : CB *12*, 1692-1697.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T.*, et al.* (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell *132*, 422-433.

Patel, R.K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PloS one *7*, e30619.

Pederson, T. (2011). The nucleolus. Cold Spring Harbor perspectives in biology *3*.

Perry, R.P. (1962). The cellular sites of synthesis of ribosomal and 4S RNA. Proceedings of the National Academy of Sciences of the United States of America *48*, 2179-2186.

Pflueger, D., Rickman, D.S., Sboner, A., Perner, S., LaFargue, C.J., Svensson, M.A., Moss, B.J., Kitabayashi, N., Pan, Y., de la Taille, A.*, et al.* (2009). N-myc downstream regulated gene 1 (NDRG1) is fused to ERG in prostate cancer. Neoplasia (New York, NY) *11*, 804-811.

Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194-1211.

Pianese, G. (1896). Beitrag zur Histologie und Aetiologie der Carcinoma. Histologische und experimentelle Untersuchungen. Beitr Pathol Anat Allg Pathol *142*, 1-193.

Pinkel, D., Straume, T., and Gray, J.W. (1986). Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. Proceedings of the National Academy of Sciences of the United States of America *83*, 2934-2938.

Pomerantz, M.M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M.P., Doddapaneni, H., Beckwith, C.A., Chan, J.A., Hills, A., Davis, M.*, et al.* (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nature genetics *41*, 882-884.

Ponting, C.P., and Hardison, R.C. (2011). What fraction of the human genome is functional? Genome research *21*, 1769-1776.

Pope, B.D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D.L., Wang, Y., Hansen, R.S., Canfield, T.K.*, et al.* (2014). Topologically associating domains are stable units of replication-timing regulation. Nature *515*, 402-405.

Popken, J., Brero, A., Koehler, D., Schmid, V.J., Strauss, A., Wuensch, A., Guengoer, T., Graf, A., Krebs, S., Blum, H., *et al.* (2014). Reprogramming of fibroblast nuclei in cloned bovine embryos involves major structural remodeling with both striking similarities and differences to nuclear phenotypes of in vitro fertilized embryos. Nucleus (Austin, Tex) *5*, 555-589.

Prober, J.M., Trainor, G.L., Dam, R.J., Hobbs, F.W., Robertson, C.W., Zagursky, R.J., Cocuzza, A.J., Jensen, M.A., and Baumeister, K. (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science (New York, NY) *238*, 336-341.

Puvion-Dutilleul, F., Puvion, E., and Bachellerie, J.P. (1997). Early stages of pre-rRNA formation within the nucleolar ultrastructure of mouse cells studied by in situ hybridization with a 5'ETS leader probe. Chromosoma *105*, 496-505.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics *13*, 341.

Raisner, R.M., Hartley, P.D., Meneghini, M.D., Bao, M.Z., Liu, C.L., Schreiber, S.L., Rando, O.J., and Madhani, H.D. (2005). Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. Cell *123*, 233-248.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., *et al.* (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1665-1680.

Raska, I., Reimer, G., Jarnik, M., Kostrouch, Z., and Raska, K., Jr. (1989). Does the synthesis of ribosomal RNA take place within nucleolar fibrillar centers or dense fibrillar components? Biology of the cell / under the auspices of the European Cell Biology Organization *65*, 79-82.

Reich, E., Franklin, R.M., Shatkin, A.J., and Tatum, E.L. (1961). Effect of actinomycin D on cellular nucleic acid synthesis and virus production. Science (New York, NY) *134*, 556-557.

Rickman, D.S., Soong, T.D., Moss, B., Mosquera, J.M., Dlabal, J., Terry, S., MacDonald, T.Y., Tripodi, J., Bunting, K., Najfeld, V., *et al.* (2012). Oncogene-mediated alterations in chromatin conformation. Proceedings of the National Academy of Sciences of the United States of America *109*, 9083-9088.

Roach, J.C., Boysen, C., Wang, K., and Hood, L. (1995). Pairwise end sequencing: a unified approach to genomic mapping and sequencing. Genomics *26*, 345-353.

Roeder, R.G., and Rutter, W.J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. Nature *224*, 234-237.

Roh, T.Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. Proceedings of the National Academy of Sciences of the United States of America *103*, 15782-15787.

Roh, T.Y., Cuddapah, S., and Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes & development *19*, 542-552.

Ronaghi, M., Uhlen, M., and Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. Science (New York, NY) *281*, 363, 365.

Rouquette, J., Genoud, C., Vazquez-Nin, G.H., Kraus, B., Cremer, T., and Fakan, S. (2009). Revealing the high-resolution three-dimensional network of chromatin and interchromatin space: a novel electron-microscopic approach to reconstructing nuclear architecture. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology *17*, 801-810.

Roussel, P., Andre, C., Comai, L., and Hernandez-Verdun, D. (1996). The rDNA transcription machinery is assembled during mitosis in active NORs and absent in inactive NORs. The Journal of cell biology *133*, 235-246.

Russell, J., and Zomerdijk, J.C. (2006). The RNA polymerase I transcription machinery. Biochemical Society symposium, 203-216.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America *74*, 5463-5467.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109-113.

Savino, T.M., Gebrane-Younes, J., De Mey, J., Sibarita, J.B., and Hernandez-Verdun, D. (2001). Nucleolar assembly of the rRNA processing machinery in living cells. The Journal of cell biology *153*, 1097-1110.

Scharf, S.J., Horn, G.T., and Erlich, H.A. (1986). Direct cloning and sequence analysis of enzymatically amplified genomic sequences. Science (New York, NY) *233*, 1076-1078.

Schermelleh, L., Heintzmann, R., and Leonhardt, H. (2010). A guide to super-resolution fluorescence microscopy. The Journal of cell biology *190*, 165-175.

Schofer, C., Weipoltshammer, K., Almeder, M., Muller, M., and Wachtler, F. (1996). Redistribution of ribosomal DNA after blocking of transcription induced by actinomycin D. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology *4*, 384-391.

Schotta, G., Lachner, M., Sarma, K., Ebert, A., Sengupta, R., Reuter, G., Reinberg, D., and Jenuwein, T. (2004). A silencing pathway to induce H3-K9

and H4-K20 trimethylation at constitutive heterochromatin. Genes & development *18*, 1251-1262.

Schwartz, D.C., Li, X., Hernandez, L.I., Ramnarain, S.P., Huff, E.J., and Wang, Y.K. (1993). Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. Science (New York, NY) *262*, 110-114.

She, X., Horvath, J.E., Jiang, Z., Liu, G., Furey, T.S., Christ, L., Clark, R., Graves, T., Gulden, C.L., Alkan, C.*, et al.* (2004). The structure and evolution of centromeric transition regions within the human genome. Nature *430*, 857-864.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proceedings of the National Academy of Sciences of the United States of America *89*, 8794-8797.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nature genetics *38*, 1348-1354.

Sirri, V., Roussel, P., and Hernandez-Verdun, D. (2000). The AgNOR proteins: qualitative and quantitative changes during the cell cycle. Micron (Oxford, England : 1993) *31*, 121-126.

Smeets, D., Markaki, Y., Schmid, V.J., Kraus, F., Tattermusch, A., Cerase, A., Sterr, M., Fiedler, S., Demmerle, J., Popken, J.*, et al.* (2014). Three-dimensional super-resolution microscopy of the inactive X chromosome territory reveals a collapse of its active nuclear compartment harboring distinct Xist RNA foci. Epigenetics & chromatin *7*, 8.

Smit, A.F. (1996). The origin of interspersed repeats in the human genome. Current opinion in genetics & development *6*, 743-748.

Smit, A.H.R.G., P. (2013-2015). RepeatMasker Open-4.0.

Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N., and de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes & development *20*, 2349-2354.

Stone, N.E., Fan, J.B., Willour, V., Pennacchio, L.A., Warrington, J.A., Hu, A., de la Chapelle, A., Lehesjoki, A.E., Cox, D.R., and Myers, R.M. (1996). Construction of a 750-kb bacterial clone contig and restriction map in the region of human chromosome 21 containing the progressive myoclonus epilepsy gene. Genome research *6*, 218-225.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C.*, et al.* (2007). Relative impact of

nucleotide and copy number variation on gene expression phenotypes. Science (New York, NY) *315*, 848-853.

Stults, D.M., Killen, M.W., Pierce, H.H., and Pierce, A.J. (2008). Genomic architecture and inheritance of human ribosomal RNA gene clusters. Genome research *18*, 13-18.

Stults, D.M., Killen, M.W., Williamson, E.P., Hourigan, J.S., Vargas, H.D., Arnold, S.M., Moscow, J.A., and Pierce, A.J. (2009). Human rRNA gene clusters are recombinational hotspots in cancer. Cancer research *69*, 9096-9104.

Suka, N., Suka, Y., Carmen, A.A., Wu, J., and Grunstein, M. (2001). Highly specific antibodies determine histone acetylation site usage in yeast heterochromatin and euchromatin. Molecular cell *8*, 473-479.

Sullivan, B.A., and Karpen, G.H. (2004). Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. Nature structural & molecular biology *11*, 1076-1083.

Sylvester, J.E., Whiteman, D.A., Podolsky, R., Pozsgay, J.M., Respess, J., and Schmickel, R.D. (1986). The human ribosomal RNA genes: structure and organization of the complete repeating unit. Human genetics *73*, 193-198.

Talasz, H., Lindner, H.H., Sarg, B., and Helliger, W. (2005). Histone H4-lysine 20 monomethylation is increased in promoter and coding regions of active genes and correlates with hyperacetylation. The Journal of biological chemistry *280*, 38814-38822.

Tanabe, H., Muller, S., Neusser, M., von Hase, J., Calcagno, E., Cremer, M., Solovei, I., Cremer, C., and Cremer, T. (2002). Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. Proceedings of the National Academy of Sciences of the United States of America *99*, 4424-4429.

Tanabe, H., Nakagawa, Y., Minegishi, D., Hashimoto, K., Tanaka, N., Oshimura, M., Sofuni, T., and Mizusawa, H. (2000). Human monochromosome hybrid cell panel characterized by FISH in the JCRB/HSRRB. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology *8*, 319-334.

Tashiro, S., and Lanctot, C. (2015). The International Nucleome Consortium. Nucleus (Austin, Tex) *6*, 89-92.

Therizols, P., Illingworth, R.S., Courilleau, C., Boyle, S., Wood, A.J., and Bickmore, W.A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science (New York, NY) *346*, 1238-1242.

Thiry, M., and Lafontaine, D.L. (2005). Birth of a nucleolus: the evolution of nucleolar compartments. Trends in cell biology *15*, 194-199.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Molecular cell *10*, 1453-1465.

Tonkin, E.T., Wang, T.J., Lisgo, S., Bamshad, M.J., and Strachan, T. (2004). NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome. Nature genetics *36*, 636-641.

van Berkum, N.L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. Journal of visualized experiments : JoVE.

van Koningsbruggen, S., Gierlinski, M., Schofield, P., Martin, D., Barton, G.J., Ariyurek, Y., den Dunnen, J.T., and Lamond, A.I. (2010). High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. Molecular biology of the cell *21*, 3735-3748.

van Steensel, B., Delrow, J., and Henikoff, S. (2001). Chromatin profiling using targeted DNA adenine methyltransferase. Nature genetics *27*, 304-308.

Vega, H., Waisfisz, Q., Gordillo, M., Sakai, N., Yanagihara, I., Yamada, M., van Gosliga, D., Kayserili, H., Xu, C., Ozono, K.*, et al.* (2005). Roberts syndrome is caused by mutations in ESCO2, a human homolog of yeast ECO1 that is essential for the establishment of sister chromatid cohesion. Nature genetics *37*, 468-470.

Verma, R.S., Dosik, H., and Lubs, H.A. (1978). Size and pericentric inversion heteromorphisms of secondary constriction regions (h) of chromosomes 1, 9, and 16 as detected by CBG technique in Caucasians: classification, frequencies, and incidence. American journal of medical genetics *2*, 331-339.

Voelkerding, K.V., Dames, S.A., and Durtschi, J.D. (2009). Next-generation sequencing: from basic research to diagnostics. Clinical chemistry *55*, 641-658.

Vogel, F., and Schroeder, T.M. (1974). The internal order of the interphase nucleus. Humangenetik *25*, 265-297.

Wallace, J.A., and Felsenfeld, G. (2007). We gather together: insulators and genome organization. Current opinion in genetics & development *17*, 400-407.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A.*, et al.* (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120-124.

Wang, Y., Li, X., and Hu, H. (2014). H3K4me2 reliably defines transcription factor binding regions in different cells. Genomics *103*, 222-228.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q.*, et al.* (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics *40*, 897-903.

Warburton, D., Gersen, S., Yu, M.T., Jackson, C., Handelin, B., and Housman, D. (1990). Monochromosomal rodent-human hybrids from microcell fusion of human lymphoblastoid cells containing an inserted dominant selectable marker. Genomics *6*, 358-366.

Warburton, P.E., Giordano, J., Cheung, F., Gelfand, Y., and Benson, G. (2004). Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome research *14*, 1861-1869.

Weber, J.L., and Myers, E.W. (1997). Human whole-genome shotgun sequencing. Genome research *7*, 401-409.

Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T.*, et al.* (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature *451*, 796-801.

Weth, O., Paprotka, C., Gunther, K., Schulte, A., Baierl, M., Leers, J., Galjart, N., and Renkawitz, R. (2014). CTCF induces histone variant incorporation, erases the H3K27me3 histone mark and opens chromatin. Nucleic acids research *42*, 11941-11951.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bahler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature *453*, 1239-1243.

Witcher, M., and Emerson, B.M. (2009). Epigenetic silencing of the p16(INK4a) tumor suppressor is associated with loss of CTCF binding and a chromatin boundary. Molecular cell *34*, 271-284.

Wright, F.A., Lemon, W.J., Zhao, W.D., Sears, R., Zhuo, D., Wang, J.P., Yang, H.Y., Baer, T., Stredney, D., Spitzner, J.*, et al.* (2001). A draft annotation and overview of the human genome. Genome biology *2*, RESEARCH0025.

Zentner, G.E., Saiakhova, A., Manaenkov, P., Adams, M.D., and Scacheri, P.C. (2011). Integrative genomic analysis of human ribosomal DNA. Nucleic acids research *39*, 4949-4960.

Zhang, X., Cowper-Sal lari, R., Bailey, S.D., Moore, J.H., and Lupien, M. (2012a). Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. Genome research *22*, 1437-1446.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.

Zhang, Y., McCord, R.P., Ho, Y.J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012b). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell *148*, 908-921.

Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U.*, et al.* (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature genetics *38*, 1341-1347.

Zuin, J., Dixon, J.R., van der Reijden, M.I., Ye, Z., Kolovos, P., Brouwer, R.W., van de Corput, M.P., van de Werken, H.J., Knoch, T.A., van, I.W.F.*, et al.* (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proceedings of the National Academy of Sciences of the United States of America *111*, 996-1001.