<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>The genomic architecture of nucleolar organiser regions on the short arms of human acrocentric chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Barreira, Sofia</td>
</tr>
<tr>
<td><strong>Publication Date</strong></td>
<td>2015-09-29</td>
</tr>
<tr>
<td><strong>Item record</strong></td>
<td><a href="http://hdl.handle.net/10379/5376">http://hdl.handle.net/10379/5376</a></td>
</tr>
</tbody>
</table>

Some rights reserved. For more information, please see the item record link above.
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......................................................................................................................v
List of tables.......................................................................................................................xi
Abstract................................................................................................................................xii
Acknowledgements...........................................................................................................xiv

1 Introduction.....................................................................................................................1
  1.1 Human Reference Genome......................................................................................1
    1.1.1 Creation of the human reference genome......................................................1
  1.2 Functional Organisation of the Human Genome...................................................6
    1.2.1 Chromatin structure.......................................................................................6
    1.2.2 Histone modifications and chromatin modulation.........................................7
  1.3 Genome packaging..................................................................................................10
    1.3.1 Genome packaging in disease .....................................................................15
  1.4 Missing regions in the Human Reference Genome..............................................16
  1.5 Nucleolus................................................................................................................19
  1.6 Nucleolar Organiser Regions ................................................................................21
  1.7 Technologies for functional and spatial organisation analysis ............................26
    1.7.1 Chromatin Immunoprecipitation sequencing, ChIP-seq ..............................27
    1.7.2 Whole transcriptome shotgun sequencing, RNA-seq ................................28
    1.7.3 Chromosome conformation capture, 3C, 4C and 5C ................................31
    1.7.4 High-throughput conformation capture, Hi-C ...........................................32
  1.8 Cell lines used in this project................................................................................35
  1.9 Aims of this thesis.................................................................................................36

2 Molecular Biology and Bioinformatics Methods ........................................................37
  2.1 Tissue Culture .........................................................................................................37
  2.2 Isolation of nucleoli..............................................................................................38
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3</td>
<td>DNA extraction from purified nucleoli</td>
<td>39</td>
</tr>
<tr>
<td>2.4</td>
<td>Measurement of nucleic acids concentration and purity</td>
<td>41</td>
</tr>
<tr>
<td>2.5</td>
<td>Gel electrophoresis</td>
<td>41</td>
</tr>
<tr>
<td>2.6</td>
<td>Fluorescence in Situ Hybridisation, FISH</td>
<td>42</td>
</tr>
<tr>
<td>2.7</td>
<td>Mosaic alignment of Roche 454 reads</td>
<td>43</td>
</tr>
<tr>
<td>2.8</td>
<td>Quality control of sequencing reads</td>
<td>43</td>
</tr>
<tr>
<td>2.9</td>
<td>Alignment of PacBio reads</td>
<td>44</td>
</tr>
<tr>
<td>2.10</td>
<td>Generation of a consensus sequence</td>
<td>44</td>
</tr>
<tr>
<td>2.11</td>
<td>Generation of sub-sequences from PacBio reads</td>
<td>44</td>
</tr>
<tr>
<td>2.12</td>
<td>Culture and storage of BAC/plasmid clones</td>
<td>45</td>
</tr>
<tr>
<td>2.13</td>
<td>Plasmid purification from small cultures</td>
<td>46</td>
</tr>
<tr>
<td>2.14</td>
<td>Plasmid purification from large cultures</td>
<td>46</td>
</tr>
<tr>
<td>2.15</td>
<td>Analysis of transcriptome profile</td>
<td>47</td>
</tr>
<tr>
<td>2.16</td>
<td>BLAST search</td>
<td>48</td>
</tr>
<tr>
<td>2.17</td>
<td>PCR/RT-PCR</td>
<td>48</td>
</tr>
<tr>
<td>2.18</td>
<td>Purification of PCR products</td>
<td>49</td>
</tr>
<tr>
<td>2.19</td>
<td>cDNA cloning and sequencing</td>
<td>50</td>
</tr>
<tr>
<td>2.20</td>
<td>Bowtie alignment of Illumina reads</td>
<td>50</td>
</tr>
</tbody>
</table>

3 Rearranged rDNA repeats

3.1 Background | 51

3.2 Results

3.2.1 DNA preparation for 454 sequencing | 57
3.2.2 Roche 454 sequences and quality control | 61
3.2.3 Alignment of 454 sequences to rDNA repeat | 66
3.2.4 Aligning 454 reads to rDNA with 10% mismatches | 68
3.2.5 Paired-end alignments against rDNA repeat with 10% mismatches | 70
3.2.6 Generation of a new consensus rDNA sequence | 73
3.2.7 DNA sample preparation for SMRT sequencing ........................................... 76
3.2.8 PacBio sequencing and sequence quality control .................................... 79
3.2.9 PacBio alignments against rDNA using the BLASR aligner ....................... 81
3.2.10 Generation of new rDNA consensus from nucleolar PacBio .................. 84
3.2.11 Search for rDNA rearrangements with RPE-1 and HeLa PacBio reads .. 87
3.2.12 Analysis of CHM1 PacBio reads ............................................................ 88
3.2.13 Search for rearrangements with CHM1 Pacbio reads .......................... 89
3.2.14 Generation of a new consensus sequence from CHM1 reads ............... 91
3.2.15 Improvement of CHM1 alignments against the rDNA repeat ............. 94
3.3 Discussion ..................................................................................................... 98

4 Spatial Organisation of the Distal Junction .................................................. 102

4.1 Background .................................................................................................. 102
  4.1.1 Functional relevance of genome spatial organisation ............................ 102
  4.1.2 Techniques to observe genome folding .................................................. 104
  4.1.3 Hi-C data sets ......................................................................................... 107
  4.1.4 Distal Junction ....................................................................................... 109

4.2 Results ......................................................................................................... 111
  4.2.1 Hi-C data quality control ....................................................................... 111
  4.2.2 Hi-C data analysis for DJ ....................................................................... 111
  4.2.3 DJ interaction maps ............................................................................... 113
  4.2.4 GSE43070 ......................................................................................... 113
  4.2.5 GSE63525 ......................................................................................... 118
  4.2.6 GSE56869 ......................................................................................... 121
  4.2.7 Analysis of other large inverted repeats in the human genome .......... 123

4.3 Discussion ..................................................................................................... 124

5 Extension and characterisation of sequences along the distal side of
acrocentric short arms ...................................................................................... 130
5.1 Background ................................................................. 130
  5.1.1 Nucleolar Organiser Regions ................................. 130
  5.1.2 Monochromosomal hybrids for human chromosomes 13, 14, 15, 21 and 22 ................................................................. 132

5.2 Results ........................................................................ 134
  5.2.1 Search for BACs from the short arms of acrocentric chromosomes ...... 134
  5.2.2 Primer design and PCR on monochromosomal hybrids .................. 137
  5.2.3 Confirmation of placement of BACs with FISH ............................... 142
  5.2.4 Sequence composition of AL591856 ........................................... 146
  5.2.5 Analysis of the chromatin and gene expression profile of AL591856 .... 149
  5.2.6 Confirmation of transcripts from AL591856 through Reverse Transcriptase PCR .............................. 154

5.3 Discussion ....................................................................... 156

6 Conclusions and Future Work .............................................. 162

Appendix A – Hi-C figures....................................................... 168
Appendix B – Sequenced clones from AL591856............................. 174
Appendix C – AL59856 chromatin profile figures............................ 175
Bibliography............................................................................ 187
List of Figures

Figure 1.1 – Workflow of Clone contig sequencing method and whole genome shotgun sequencing method for *de novo* assembly of large genomes. ..........3

Figure 1.2 – Colour 3D FISH representation and classification of chromosomes in a human G0 fibroblast nucleus. ................................................. 11

Figure 1.3 - ANC-INC network model of nuclear organization based on spatially co-aligned active and inactive nuclear compartments. .......................14

Figure 1.4 - Examples of repeats found in the human genome. .................... 17

Figure 1.5 – Internal structure of the nucleolus. ........................................ 20

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. ........................................ 22

Figure 1.7 - Human rDNA repeat extracted from BAC AL592188 [105424 - 149395] bp. .......................................................... 23

Figure 1.8 - Location of sequences identified adjacent to the rDNA repeats that also comprise NORs......................................................... 24

Figure 1.9 – Sequence characterisation of the DJ and PJ. ............................ 25

Figure 1.10 - Localisation of the DJ to the nucleolar periphery during interphase. .......................................................... 26

Figure 1.11 - Workflow of ChIP-seq method and analysis.......................... 27

Figure 1.12 - Workflow of RNA-seq technology and analysis..................... 30

Figure 1.13 - Methodology for chromosome conformation capture (3C). ........ 31

Figure 1.14 - The Hi-C method.................................................. 33

Figure 1.15 - Workflow of Hi-C reads analysis........................................ 34
Figure 3.1 - The five human acrocentric chromosomes, 13, 14, 15, 21 and 22....
.................................................................................................................................. 51
Figure 3.2 - Nucleolar organiser regions, NORs, are located in the short arms of
the acrocentric chromosomes.......................................................................................... 52
Figure 3.3 - Human rDNA repeat extracted from BAC AL592188. ....................... 53
Figure 3.4 - Combing of rDNA reveals canonical organisation for 18S (green)
and 28S (red) regions in 1Mb DNA fibres...................................................................... 54
Figure 3.5 - Gel electrophoresis of purified nucleolar DNA. ................................. 60
Figure 3.6 - FISH of nucleolar DNA. ......................................................................... 61
Figure 3.7 - Quality scores of 454 shotgun sequencing data across all bases and
sequence length distribution........................................................................................... 63
Figure 3.8 - Quality scores across all bases for the 454 paired-end file and
distribution of sequence lengths...................................................................................... 64
Figure 3.9 - Difference in the lengths of left and right paired-end reads.............. 65
Figure 3.10 - All 454 reads mapped against the rDNA repeat extracted from
AL592188. ..................................................................................................................... 67
Figure 3.11 - Remapping of all 454 reads with 10% mismatches. ......................... 69
Figure 3.12 – Strategy to look for rearrangements using the 454 paired-end reads
from nucleolar DNA. ..................................................................................................... 70
Figure 3.13 - Alignment of 454 paired-end reads to the rDNA repeat, allowing
10% mismatches per read............................................................................................... 72
Figure 3.14 - Comparison between the rDNA repeat extracted from AL592188
and the consensus generated from RPE-1 454 reads. ............................................. 74
Figure 3.15 - Gel electrophoresis of purified nucleolar DNA from HeLa............. 77
Figure 3.16 - FISH of RPE-1 nucleolar DNA (green) and rDNA (red)............. 78
Figure 3.17 - FISH of HeLa nucleolar DNA (green) and rDNA (red) ..................79
Figure 3.18 - PacBio quality report for RPE-1 nucleolar sample ..................80
Figure 3.19 - PacBio quality report for the HeLa nucleolar sample .................80
Figure 3.20 - Alignment of RPE-1 PacBio reads to the rDNA repeat extracted from AL592188 ..................................................82
Figure 3.21 - Alignment of nucleolar HeLa PacBio reads to the rDNA repeat... ........................................................................................................83
Figure 3.22 - Comparison between the rDNA repeat from AL592188 and the consensus generated from RPE-1 PacBio reads ..................................86
Figure 3.23 - Comparison between the rDNA repeat from AL592188 and the consensus generated from HeLa PacBio reads ..................................86
Figure 3.24 - BLAST alignment of generated consensus sequence from RPE-1 PacBio reads against the U13369 rDNA repeat ..................................86
Figure 3.25 - BLAST alignment of generated consensus sequence from HeLa PacBio reads against the U13369 rDNA repeat ..................................86
Figure 3.26 – Sequences from either side of the PacBio reads were used to look for rearrangements .................................................................87
Figure 3.27 - Alignment of CHM1 PacBio reads against the rDNA repeat extracted from AL592188 .................................................................90
Figure 3.28 - Comparison between the rDNA repeat from AL592188 and the consensus generated from CHM1 PacBio reads ..................................92
Figure 3.29 – Blast alignment of CHM1 consensus against the rDNA repeat U13369 ...............................................................................................92
Figure 3.30 - Alignment of CHM1 PacBio reads to the rDNA repeat extracted from AL592188 ..............................................................................96
Figure 3.31 - Comparison between the rDNA repeat from AL592188 and the consensus generated from CHM1 PacBio reads. .................................................................97

Figure 3.32 – Scheme of PacBio reads, containing the adapter, that reported rearrangements. .........................................................................................................99

Figure 4.1 - Genome-wide contact matrices for chromosome 14 using HindIII and NcoI as restriction enzymes. .................................................................105

Figure 4.2 – Hi-C data shows the human nucleus is segregated into open and closed chromatin compartments.................................................................106

Figure 4.3 - Location and arrangement of the large inverted repeat (white arrows) in the DJ contig (in green).................................................................110

Figure 4.4 - Strategy to analyse the spatial conformation of long-range intramolecular interactions of the DJ using Hi-C sequencing reads........112

Figure 4.5 - Intrachromosomal interactions in the DJ captured by Hi-C data from IMR90 cells in normal conditions..................................................114

Figure 4.6 - Intramolecular interactions in the DJ using Hi-C reads from IMR90 cells in normal conditions. .................................................................115

Figure 4.7 - Intramolecular interactions in the DJ using Hi-C reads from IMR90 cells in normal conditions (replicate sample).................................116

Figure 4.8 - DJ structural domain. .................................................................................117

Figure 4.9 - Intramolecular interactions in the DJ after treatment of IMR90 cells upon flavopiridol treatment.............................................................118

Figure 4.10 - Intrachromosomal interactions in the DJ obtained through analysis of Hi-C reads from the GSE63525 study. ........................................119

Figure 4.11 - Intrachromosomal interactions in the DJ obtained through analysis of Hi-C reads from the GSE63525 study. ........................................120
List of Figures

Figure 4.12 - Intrachromosomal interactions in the DJ with Hi-C reads from analysis of Hi-C reads from the GSE63525 study. .............................................121

Figure 4.13 - Observation of intramolecular interactions between the DJ large inverted repeats in DNase Hi-C reads from K562 cells.................................122

Figure 4.14 - Analysis of intrachromosomal interactions in two large inverted repeats present in the human genome..............................................................124

Figure 4.15 – The DJ chromatin intramolecular contacts form a loop structure centred at the large inverted repeat. .................................................................125

Figure 4.16 - ChIP-seq peaks or CTCF, H3K4me3, H3K36me3 and Pol II and DNase-seq in the DJ........................................................................................................127

Figure 5.1 - Schematic of an acrocentric chromosome.....................................131

Figure 5.2 - PCR with monochromosomal hybrids searching for regions of the DJ. .........................................................................................................................133

Figure 5.3 - Strategy to identify novel sequences in the short arms of acrocentric chromosomes using Hi-C sequencing data.................................136

Figure 5.4 - Gel electrophoresis of PCR product from AC013640 using DNA from monochromosomal somatic cell hybrids (mouse/human) as template. ......................................................................................................................139

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.........................................................140

Figure 5.6 - Gel electrophoresis of PCR product from AL591856 using DNA from monochromosomal somatic cell hybrids (mouse/human) as template. ..................................................................................................................141

Figure 5.7 - FISH of AC03640 BAC on human male metaphase slides..........143
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.8</td>
<td>FISH of AC103988.7 BAC on human male metaphase chromosomes</td>
</tr>
<tr>
<td>5.9</td>
<td>FISH of AL591856 on male metaphase chromosomes</td>
</tr>
<tr>
<td>5.10</td>
<td>FISH of BAC AL591856 and alpha satellite probes specific for individual acrocentric chromosomes on male metaphase chromosomes</td>
</tr>
<tr>
<td>5.11</td>
<td>Schematic of sequence homology for AL591856</td>
</tr>
<tr>
<td>5.12</td>
<td>FISH of AL591856 shows cross-hybridisation with the proximal side of rDNA</td>
</tr>
<tr>
<td>5.13</td>
<td>ChIP-seq peaks for H3K4me3, Pol II H3K36me3 and CTCF and assembled RNA-seq transcripts for K562</td>
</tr>
<tr>
<td>5.14</td>
<td>ChIP-seq peaks for H3K4me3, Pol II H3K36me3 and CTCF and assembled RNA-seq transcripts for Nhek</td>
</tr>
<tr>
<td>5.15</td>
<td>ChIP-seq peaks for H3K4me3, Pol II H3K36me3 and CTCF and assembled RNA-seq transcripts for Huvec</td>
</tr>
<tr>
<td>5.16</td>
<td>Reverse transcriptase PCR to confirm the occurrence of transcription in AL591856</td>
</tr>
<tr>
<td>5.17</td>
<td>Schematic of the two spliced variants identified by RT-PCR</td>
</tr>
<tr>
<td>5.18</td>
<td>Positioning Positioning of AL591856 in the short arms of acrocentric chromosomes</td>
</tr>
</tbody>
</table>
List of Figures

Figure A 1 - Intramolecular interactions in the DJ using Hi-C reads from Gm12878 cells in normal conditions (replicate sample)........................................................................................................168

Figure A 2 - Intramolecular interactions in the DJ using Hi-C reads from RWPE1 cells in normal conditions.................................................................................................................................169

Figure A 3 - Intramolecular interactions in the DJ using Hi-C reads from Huntington-Guilford Progeria Syndrome (HGPS) fibroblasts in normal conditions.............................................170

Figure A 4 - Intramolecular interactions in the DJ using Hi-C reads from HEK293 cells in normal conditions.................................................................................................................................171

Figure A 5 - Intramolecular interactions in the DJ using Hi-C reads from MCF-7 cells in normal conditions.................................................................................................................................172

Figure A 6 - Intrachromosomal interactions in 4 inverted repeats found in chromosomes 4, 10, 11 and 12 ..............................................................................................................................................173

Figure C 1 ChIP-seq peaks for histone modification H3K4me1 for BAC AL591856 ..........175

Figure C 2 - ChIP-seq peaks for histone modification H3K4me2 for BAC AL591856 ..........176

Figure C 3 - ChIP-seq peaks for histone modification H3K4me3 for BAC AL591856 ..........177

Figure C 4 ChIP-seq peaks for histone modification H3K9ac for BAC AL591856 ..............178

Figure C 5 - ChIP-seq peaks for histone modification H3K9me3 for BAC AL591856 ..........179

Figure C 6 - ChIP-seq peaks for histone modification H3K27ac for BAC AL591856 ..........180

Figure C 7 - ChIP-seq peaks for histone modification H3K27me3 for BAC AL591856 ..........181

Figure C 8 - ChIP-seq peaks for histone modification H3K36me3 for BAC AL591856 ..........182

Figure C 9 - ChIP-seq peaks for histone modification H4K20me1 for BAC AL591856 ..........183

Figure C 10 - ChIP-seq peaks for histone modification CTCF for BAC AL591856 ..........184

Figure C 11 - ChIP-seq peaks for histone modification Pol II for BAC AL591856 ..........185

Figure C 12 - RNA-seq assembled transcripts from AL591856.................................186
List of Tables

Table 1.1 - List of common histone modifications and associations ........................................8
Table 3.1 - Summary of statistics for shotgun (single-end) and paired-end 454 libraries.........62
Table 3.2 - Summary of statistics of the paired-end reads............................................................62
Table 3.4 - Nucleotide mismatches between AL592188 rDNA repeat and the new consensus generated from alignment of 454 reads ..................................................................................75
Table 3.5 - Reported mismatches for the RPE-1 and HeLa consensuses relative to the rDNA repeat from AL592188 ......................................................................................................................85
Table 3.6 - Mismatches between rDNA repeat from AL592188 and CHM1 consensus ..........93
Table 3.7 - Mismatches between rDNA repeat from AL592188 and CHM1 consensus (85% identity and alignment length at least 90% read length) .........................................................95
Table 4.1 - List of Hi-C data sets employed to study the spatial organisation of the DJ ..........107
Table 5.1 - Primer pairs for BAC AC013640 and expected product lengths .......................137
Table 5.2 - Primer pairs for BAC AC1039887.7 and expected product lengths .................138
Table 5.3 - Primer pairs for BAC AL591856 and expected product lengths .......................138
Table 5.4 - Primer pair sequence and expected product length for transcript from AL591856. 154

Table B 1 – Sequenced cDNA clones from AL591856 ................................................................174
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.
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 template (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 chromosomes (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 (ETs) 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

<table>
<thead>
<tr>
<th>Histone Modification</th>
<th>Transcription regulation</th>
<th>Association</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me1</td>
<td>Activation</td>
<td>Enhancers</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>Activation</td>
<td>Promoters and enhancers</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>Activation</td>
<td>Promoters</td>
</tr>
<tr>
<td>H3K9me2</td>
<td>Repression</td>
<td>Heterochromatin</td>
</tr>
<tr>
<td>H3K9me3</td>
<td>Repression</td>
<td>Heterochromatin</td>
</tr>
<tr>
<td>H3K9ac</td>
<td>Activation</td>
<td>Active regulatory regions</td>
</tr>
<tr>
<td>H3K27me1</td>
<td>Activation</td>
<td>Euchromatin</td>
</tr>
<tr>
<td>H3K27me2</td>
<td>Repression</td>
<td>Polycomb-repressed regions</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>Repression</td>
<td>Polycomb-repressed regions</td>
</tr>
<tr>
<td>H3K36me3</td>
<td>Activation</td>
<td>Transcribed regions</td>
</tr>
<tr>
<td>H4K20me1</td>
<td>Activation</td>
<td>Transcribed regions</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>Activation</td>
<td>Active regulatory regions</td>
</tr>
</tbody>
</table>

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).

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).
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.
Chapter 1

(Nebran-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).

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.](image)

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).
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).
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 37°C with 5% CO₂. 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 37°C 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 4°C 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 4°C 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 4°C for 5 minutes. Liquid was discarded and 5 mL of Buffer A was added (10 mM Heps, pH 7.9, 10 mM KCl, 1.5 mM MgCl₂, 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 4°C for 5 minutes. Pellet was resuspended in 3 mL S1 solution (0.25 M Sucrose, 10 mM MgCl₂ + 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 MgCl₂ + 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 4°C. 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 MgCl₂ + one tablet Protease inhibitor, complete mini EDTA-free) and centrifuged at 3000xg for 10 min at 4°C. Supernatant was discarded and nucleolar pellet resuspended in 0.5 mL of S2 solution and further sonicated at 1430xg for 5 min at 4°C. Resulting pellet contained highly purified nucleoli (checked under microscope) and stored at -80°C 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 50°C, 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 -80°C 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 37°C 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 27°C 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 3-5 min 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 (37°C) 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 37°C incubator.

A glycerol stock was prepared for long-term storage at -80°C. 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 -80°C.
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 37°C overnight. Afterwards, cells were centrifuged at 4000xg for 15 min at 4°C. Low copy BAC DNA and high copy Plasmid 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
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 MgCl₂, 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 98°C followed
by 35 cycles of 10 sec at 98°C, 20 sec at 50-70°C, 4 min 72°C for 20-30 sec per kb and final extension for 2 min at 72°C.

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.](image)

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.
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).

![Diagram of canonical and non-canonical rDNA repeats](image)

**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.
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

<table>
<thead>
<tr>
<th>SID</th>
<th>Sample</th>
<th>HQ reads</th>
<th>HQ bases</th>
<th>Avg Read Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>17426</td>
<td>RPE-1</td>
<td>665,998</td>
<td>233,740,085</td>
<td>351</td>
</tr>
<tr>
<td>17463</td>
<td>RPE-1 paired-end</td>
<td>563,816</td>
<td>199,000,909</td>
<td>353</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>SID</th>
<th># Reads</th>
<th>%Reads</th>
<th>Avg left length</th>
<th>Avg right length</th>
</tr>
</thead>
<tbody>
<tr>
<td>17363</td>
<td>423,876</td>
<td>75.18%</td>
<td>165</td>
<td>166</td>
</tr>
</tbody>
</table>

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).

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.
rDNA repeat extracted from AL592188 (43,972 bp)

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

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

<table>
<thead>
<tr>
<th>Nucleotide Position</th>
<th>rDNA AL592188</th>
<th>RPE-1</th>
<th>HeLa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1756</td>
<td>C</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>2602</td>
<td>C</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>3714</td>
<td>G</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>7794</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

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.
rDNA repeat extracted from AL592188

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.

rDNA repeat extracted from AL592188

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](image)

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).
rDNA repeat extracted from AL592188

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

<table>
<thead>
<tr>
<th>Nucleotide position</th>
<th>AL592188 nucleotide</th>
<th>CHM1 consensus nucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>139</td>
<td>T</td>
<td>Y (C or T)</td>
</tr>
<tr>
<td>1756</td>
<td>C</td>
<td>Y (C or T)</td>
</tr>
<tr>
<td>2338</td>
<td>A</td>
<td>R (A or G)</td>
</tr>
<tr>
<td>2512</td>
<td>C</td>
<td>S (C or G)</td>
</tr>
<tr>
<td>2602</td>
<td>C</td>
<td>Y (C or T)</td>
</tr>
<tr>
<td>2620</td>
<td>A</td>
<td>M (A or C)</td>
</tr>
<tr>
<td>2891</td>
<td>G</td>
<td>S (C or G)</td>
</tr>
<tr>
<td>2955</td>
<td>T</td>
<td>Y (C or T)</td>
</tr>
<tr>
<td>3207</td>
<td>C</td>
<td>M (A or C)</td>
</tr>
<tr>
<td>3714</td>
<td>G</td>
<td>R (A or G)</td>
</tr>
<tr>
<td>5734</td>
<td>A</td>
<td>R (A or G)</td>
</tr>
<tr>
<td>5819</td>
<td>A</td>
<td>R (A or G)</td>
</tr>
<tr>
<td>6523</td>
<td>G</td>
<td>K (G or T)</td>
</tr>
<tr>
<td>6524</td>
<td>G</td>
<td>K (G or T)</td>
</tr>
<tr>
<td>6525</td>
<td>G</td>
<td>S (C or G)</td>
</tr>
<tr>
<td>7794</td>
<td>C</td>
<td>Y (C or T)</td>
</tr>
<tr>
<td>12834</td>
<td>G</td>
<td>R (A or G)</td>
</tr>
<tr>
<td>13521</td>
<td>G</td>
<td>R (A or G)</td>
</tr>
<tr>
<td>13556</td>
<td>A</td>
<td>M (A or C)</td>
</tr>
<tr>
<td>14399</td>
<td>T</td>
<td>Y (C or T)</td>
</tr>
</tbody>
</table>
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.
rDNA repeat extracted from AL592188

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.

![Diagram of PacBio reads with an adapter](image)

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-I 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

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

<table>
<thead>
<tr>
<th>Primer</th>
<th>Primer sequence</th>
<th>Product size</th>
</tr>
</thead>
<tbody>
<tr>
<td>ac013640_1f</td>
<td>TGGGCCGCTGCTATTTTTATG</td>
<td>-</td>
</tr>
<tr>
<td>ac013640_1r</td>
<td>TAAGCACTTAGCTGCCTACTGAA</td>
<td>224 bp</td>
</tr>
<tr>
<td>ac013640_2f</td>
<td>TTGGGGTCACATTACTGCCC</td>
<td>-</td>
</tr>
<tr>
<td>ac013640_2r</td>
<td>TGTCGCACAAATCTCAAA</td>
<td>323 bp</td>
</tr>
<tr>
<td>ac013640_3f</td>
<td>CCAGGTTTCTGGCCTTCCTTT</td>
<td>-</td>
</tr>
<tr>
<td>ac013640_3r</td>
<td>TTTTCGCTTCCTCAAGC</td>
<td>479 bp</td>
</tr>
</tbody>
</table>
Table 5.2 - Primer pairs for BAC AC1039887.7 and expected product lengths

<table>
<thead>
<tr>
<th>Primer</th>
<th>Primer sequence</th>
<th>Product size</th>
</tr>
</thead>
<tbody>
<tr>
<td>ac103988.7_1f</td>
<td>GGAGCTCTTTGCCTGCCTAAT</td>
<td>-</td>
</tr>
<tr>
<td>ac103988.7_1r</td>
<td>ATGTTTAGCGTTCCATAACACGA</td>
<td>439 bp</td>
</tr>
<tr>
<td>ac103988.7_2f</td>
<td>ATTTGGGAGGGGTGGGGAATTATTA</td>
<td>-</td>
</tr>
<tr>
<td>ac103988.7_2r</td>
<td>AGCCACACAGTTAGATGCTGTTA</td>
<td>227 bp</td>
</tr>
<tr>
<td>ac103988.7_3f</td>
<td>TTGCCACTTTTGTGAACGGCT</td>
<td>-</td>
</tr>
<tr>
<td>ac103988.7_3r</td>
<td>AGCTGGAGTTCGGTGAGAGA</td>
<td>377 bp</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Primer</th>
<th>Primer sequences</th>
<th>Product size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer_105_F</td>
<td>GTAGTGAAATAGTTAATGCAGCTGA</td>
<td>-</td>
</tr>
<tr>
<td>Primer_105_R</td>
<td>CCCACTGAAAGATGTACCCACAA</td>
<td>212 bp</td>
</tr>
<tr>
<td>Primer_102_F</td>
<td>TGGCCGGGCATTCAATTTTC</td>
<td>-</td>
</tr>
<tr>
<td>Primer_102_R</td>
<td>CCGTCCTGTGCTGGTGACGT</td>
<td>211 bp</td>
</tr>
<tr>
<td>Primer_84_F</td>
<td>CACTGCTGCTTGAAGGGCAACA</td>
<td>-</td>
</tr>
<tr>
<td>Primer_84_R</td>
<td>GCTCCCTCTGGCCGTTGACTG</td>
<td>174 bp</td>
</tr>
</tbody>
</table>

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, 22 with 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.
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.
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.

<table>
<thead>
<tr>
<th>Primer Pair</th>
<th>Sequence</th>
<th>Product length</th>
</tr>
</thead>
<tbody>
<tr>
<td>XR1109f</td>
<td>GAGAATCCCGGCTCTGCTGC</td>
<td>-</td>
</tr>
<tr>
<td>XR1109r</td>
<td>GCAGGAAGAAGTCTCTCATCAGG</td>
<td>198 bp</td>
</tr>
</tbody>
</table>

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’- TTCCAGAACACTGCTRCKRGGTCTGAATGTGTGGTCCCTCACATAGGA
-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
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 organising 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
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.
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.
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 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.
### Sequenced clones from AL591856

#### Table B 1 – Sequenced cDNA clones from AL591856

<table>
<thead>
<tr>
<th>AL591856 cDNA</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clone 2</td>
<td>GAATTCCGATTGCAGGAAGAAGTCTCTTCATCAGGTACAGATGACAGTGCCTTGATCCTGGGGATTCTCAATCACTAGTGAATTC</td>
</tr>
<tr>
<td>Clone 3</td>
<td>GAATTCCGATTGCAGGAAGAAGTCTCTTCATCAGGTACAGATGACAGTGCCTTGATCCTGGGGATTCTCAATCACTAGTGAATTC</td>
</tr>
<tr>
<td>Clone 4</td>
<td>GAATTCCGATTGCAGGAAGAAGTCTCTTCATCAGGTACAGATGACAGTGCCTTGATCCTGGGGATTCTCAATCACTAGTGAATTC</td>
</tr>
<tr>
<td>Clone 5</td>
<td>GAATTCCGATTGCAGGAAGAAGTCTCTTCATCAGGTACAGATGACAGTGCCTTGATCCTGGGGATTCTCAATCACTAGTGAATTC</td>
</tr>
</tbody>
</table>
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


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


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


Bibliography


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


