

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Characterization of the human nucleolar organizer regions

*A dissertation presented in partial fulfilment of the
requirements for the degree of*

Doctor of Philosophy

in

Genetics

at Massey University, Albany, New Zealand.

Saumya Agrawal

2014

© 2014
Saumya Agrawal
All rights reserved

To Maa and Papa

[Blank page]

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us

Charles Dickens

A Tale of Two Cities, Chapter 1, 1859

[Blank Page]

Abstract

The short arms of all the five human acrocentric chromosomes contain genomic region known as nucleolar organizer regions (NORs). The NOR is the site of nucleolus formation and therefore play critical role in cell survival. It has two components: a tandem array of ribosomal DNA (rDNA) units and regions surrounding the rDNA tandem array, the rDNA flanking regions. In this work, I have explored both components of the NOR to unravel their genomic and functional features.

Aside from their rRNA coding function, the rDNA intergenic spacer (IGS) are thought to contain many non-coding functional elements that are involved in a variety of cellular processes. The repetitive nature of the IGS has made these non-coding elements difficult to identify, therefore I employed phylogenetic footprinting to identify putative functional elements in the human rDNA. To implement phylogenetic footprinting, I performed whole genome assemblies to determine the rDNA sequences of six primate species. These primate rDNA sequences were compared with human rDNA to identify fifty-three conserved regions in the human IGS that correspond to known rDNA functional elements, as well as novel conserved regions with unknown function. The human IGS is known to transcribe noncoding RNAs and therefore, to identify transcripts from the novel conserved regions I performed RNA-seq analysis. Integration of phylogenetic footprinting and RNA-seq analysis results revealed that several conserved regions potentially actively transcribe a number of long poly(A)- transcripts that include a cancerous tissues specific transcript, which is antisense to the pRNA and another transcript from *cdc27* pseudogene present in different cell types and a small poly(A)- transcript specific to embryonic cells. The integration of phylogenetic footprinting and IGS chromatin profiling revealed enrichment of active histone modifications and transcription factors in the IGS conserved regions demonstrating that these regions potential act as transcription regulators. Three conserved potential origin of replication sites in the IGS were also identified. Further evidence of Pol II and Pol III association with the human IGS were provided that strongly demonstrate that aside of Pol I other two RNA polymerase machineries potentially transcribe the human IGS. Overall, this work provides an extensive dataset of potential functional conserved regions in the human IGS, and evidence for different functions associated with them.

The rDNA flanking regions thought to have role in the nucleolus formation/fusion. However, the genomic characteristic of the regions is unknown as they are missing from the current human genome assembly. Therefore, I characterized the rDNA flanking regions the distal rDNA flanking region (telomere side) and the proximal rDNA flanking region (centromere side) using the sequences of the regions that were identified by our collaborators. The

sequences of the flanking regions are highly conserved among the acrocentric chromosomes suggesting that they frequently exchange sequences. The proximal region similar to the pericentromeric regions is highly segmentally duplicated. On the other side, the distal region is merely segmentally duplicated but has two unique features a large inverted repeat region (~227 kb) and a long stretch of CER satellite repeats, potential binding site of a protein of unknown function. These parts of the genome are thought to be heterochromatic, however I employed a gene prediction pipeline that provide evidence for coding potential in both the flanking regions. Finally, it has been reported that the proximal junction point may be variable therefore, I designed a novel bioinformatics mapping technique, which suggests there are at least two distinct proximal junction points. Overall, this work demonstrates that the rDNA flanking regions are not merely heterochromatic wastelands but instead are highly complex and have its own genomic characteristics.

Taken together, the results from my work provide a platform for a more comprehensive characterization of the functional elements in the IGS and the rDNA flanking regions. This will lead to a better understanding of the biological processes that are related to NORs and will ultimately help to explore the mechanisms that underlie these processes, which are still far from being completely understood.

Acknowledgements

After finishing my Master's degree, I was sure that I would never enter in a University only very soon to realize that I could not survive without doing science. This return journey was not possible without the support, encouragement, advice and guidance of several amazing people whom I would like to thanks here.

First and foremost, I would like to thank to my supervisor Dr. Austen Ganley for providing me an excellent environment to develop my scientific thinking, giving me freedom to do experiments, being so patience and investing years of time to guide and shape me as a researcher. I would like to thank you for all the constructive feedbacks and guidance. I cannot think a better teacher than you whom I came across in my academic life. I hope my gratitude reflect from my work.

I would like to thank my co-supervisor Assoc. Prof. Murray Cox for constant encouragement and support. I would also like to extend my gratitude to my project collaborator Prof. Brian McStay (NUI, Galway, Ireland). I also like to thank my co-supervisor Prof. Peter Lockhart for being on my advisory committee.

I would like to thank Massey University and Institute of Natural and Mathematical Sciences for Institute of Natural Sciences Scholarship that financially supported me during my PhD. Special thanks Muharram Khoussainova and Colleen van Es for always being so supportive regarding the administrative and financial issues.

I would like to thank Matt, Daniela and Val for their moral support and for sparing time from their extremely busy schedule to proof read my thesis. Val and Daniela, what can I say other than I am so lucky and blessed to have you guys as friends. Daniela a special thanks to you for all the care and keeping me in touch with the life outside the research world. Val thanks for mature discussions and for helping me to develop a better understanding about different aspects of the scientific world. Matt I would like to thank you for your guidance, support, friendship, listening all of mine crazy scientific theories, coffee and being so patience with thermostat setting of the office. A humble gratitude (I know it is not enough) to you Matt and Daniela especially for your help in the final stage of my thesis submission.

During my PhD, I become friends with some wonderful people whom I would like to thank. Ralph thanks for all the encouragement and moral support, Martina for all the wise advice, Laura for showing me different sides of the world and Jarod for being a constant source of moral support. I also would like to thank Inswasti (Ninin), Rashmi and Tatyana for their

friendship, Jyothsana for all the help especially during the early phase of my New Zealand life, Eli for all the moral support and John for stopping me driving on the highway.

I would like to thank several past and present members of Ganley group. First and foremost, Elizabeth I would like to thank you for being so caring, supportive and especially for the lunch on my first day in the lab. I would like to thank “the walking encyclopaedia” of our group Mark Walker for constant encouragement, support and sharing long thesis writing evenings. Special thanks to Nazanin, Aida, Naomi, Illog, Geng Geng and Ting for all their support. I also like to thank Yogesh Dalvi for teaching me molecular biology technique “PCR”.

I would like to thank my Mum, Dad and my sister Sonam for believing in me and in my dreams, for their unconditional love and encouragement in every moment of my life. I would also like to thank my uncle Arvind chacha and aunt Suman chachi for their constant support. It was impossible for me to finish this myriad task without the care and endless support from you all.

Table of Contents

Abstract	v
Acknowledgements	vii
Table of Contents	ix
Table of Figures	xv
Table of Tables	xviii
Abbreviations	xx
1. Chapter 1: Introduction	1
1.1. Nucleolar organizer region	3
1.2. Status of the NOR in human genome assembly	7
1.3. rDNA homogeneity in the primates	8
1.4. Human ribosomal DNA	9
1.4.1. Human 47S pre-rRNA coding region	11
1.4.2. Human rDNA IGS	11
1.4.3. Functional elements in the human IGS	12
1.4.3.1. Transcription regulators	14
1.4.3.2. Origin of replication	15
1.4.3.3. Replication fork barrier (RFB) site	15
1.4.3.4. rDNA transcription enhancer	16
1.4.3.5. Putative protein binding sites	16
1.4.3.6. Noncoding transcripts from the IGS	17
1.5. The chromatin state of the rDNA in human	18
1.6. Human rDNA array flanking regions	18
1.7. Role of rDNA in fundamental biological processes	20
1.8. Aims of the study	21
2. Chapter 2: Identification of potential functional elements in the intergenic spacer of the human ribosomal DNA	23
2.1. Introduction	25
2.1.1. Strategy to characterized potential functional elements in the IGS	25
2.2. Material and methods	31
2.2.1. Bioinformatics Techniques	31

2.2.1.1. Comparative analysis of Sanger read assemblers to determine the efficiency of assembling rDNA repeat unit sequences.....	31
2.2.1.1.1. Extraction of potential rDNA reads from Sanger whole genome sequencing data:	31
2.2.1.1.2. Test assemblies.....	31
2.2.1.2. Whole genome assemblies to obtain the primate rDNA sequences	32
2.2.1.2.1. Datasets.....	32
2.2.1.2.2. Whole Genome Assembly.....	32
2.2.1.2.3. rDNA sequence construction.....	33
2.2.1.3. Primate rDNA BAC sequencing.....	34
2.2.1.3.1. NGS sequencing.....	34
2.2.1.3.2. Read preparation.....	34
2.2.1.3.3. Assembly.....	34
2.2.1.3.4. Mapping.....	34
2.2.1.4. rDNA Sequence analysis.....	34
2.2.1.5. Multiple sequence alignment and Similarity plot.....	36
2.2.1.6. ChIP-seq and RNA-seq analysis of the human rDNA sequence	36
2.2.1.6.1. Modified human genome assembly.....	36
2.2.1.6.2. Data set for ChIP-seq and RNA-seq analysis	37
2.2.1.6.3. ChIP-seq analysis	37
2.2.1.6.4. RNA-seq assembly:.....	39
2.2.2. Molecular Techniques	40
2.2.2.1. BAC filters and BAC clones:	40
2.2.2.2. Probe preparation:.....	40
2.2.2.2.1. Probe for screening BAC filters	40
2.2.2.2.2. Probe for identifying rDNA units in I-PpoI digested Southern blots	40
2.2.2.3. Southern Hybridization:	40
2.2.2.4. Verification of the presence of the rDNA unit in the <i>E. coli</i> containing BACs:	41
2.2.2.5. BAC extraction:.....	42
2.2.2.6. I-PpoI Digestion:	42
2.2.2.7. Field inversion gel electrophoresis	42
2.2.2.8. Southern blotting	43
2.3. Results	44
2.3.1. Whole genome assembly strategy to obtain the primate rDNA sequences	44
2.3.2. Selection of the primate species for the phylogenetic footprinting of human rDNA	46

2.3.3.	Comparison of sequence assemblers to determine the ability to assemble the rDNA sequence.....	48
2.3.3.1.	Dataset to assess the efficiency of the Sanger assemblers	48
2.3.3.2.	<i>De novo</i> assembly comparison of sequence assemblers using lib_12500 dataset:	49
2.3.4.	Reference human rDNA unit sequence.....	51
2.3.5.	Construction and verification of primate rDNA unit sequences	51
2.3.5.1.	Construction of primate rDNA sequences using whole genome assembly strategy.....	51
2.3.5.1.1.	Chimpanzee reference rDNA unit sequence	51
2.3.5.1.2.	Gorilla reference rDNA unit sequence.....	53
2.3.5.1.3.	Orangutan reference rDNA unit sequence	55
2.3.5.1.4.	Gibbon reference rDNA unit sequence	56
2.3.5.1.5.	Macaque reference rDNA unit sequence	58
2.3.5.1.6.	Marmoset reference rDNA unit sequence.....	59
2.3.5.2.	Verification of primate rDNA sequences obtained from WGA strategy using BAC clones	61
2.3.5.2.1.	Identification of BAC clones by screening BAC libraries.....	62
2.3.5.2.2.	Verification of the gorilla rDNA using BAC clones.....	67
2.3.5.2.3.	Verification of the orangutan rDNA using BAC clones	69
2.3.5.2.4.	Verification of the gibbon rDNA using BAC clones	71
2.3.5.2.5.	Verification of the macaque rDNA using BAC clones.....	73
2.3.5.2.6.	Verification of the marmoset rDNA using BAC clones	75
2.3.5.3.	The primate reference rDNA sequences	77
2.3.6.	Characterization of the human and six primate rDNA sequences	79
2.3.6.1.	Coding region	86
2.3.6.2.	Microsatellites:.....	86
2.3.6.3.	Satellites:.....	87
2.3.6.4.	Alu elements:	88
2.3.6.5.	Additional repeat elements:	90
2.3.7.	Phylogenetic footprinting to identify potential noncoding functional elements in the IGS	90
2.3.7.1.	Conservation of previously known features in the human IGS:	93
2.3.7.1.1.	rRNA coding regions	93
2.3.7.1.2.	c-Myc and p53 binding sites	93
2.3.7.1.3.	Noncoding transcripts	94
2.3.7.1.4.	Alu elements conservation.....	94

2.3.7.1.5. Conservation of cdc27 pseudogene in apes	94
2.3.8. Search for potential functionality of conserved regions of unknown function:	95
2.3.8.1. Conservation of transcriptionally active regions	95
2.3.8.2. Conserved regions as potential transcriptional regulators	98
2.3.8.3. Origin of replication	102
2.3.9. Transcription machinery associated with the rDNA.....	105
2.4. Discussion.....	107
2.4.1. Potential transcripts and transcription regulatory elements in the human IGS	107
2.4.2. Cdc27 pseudogene as potential regulator	109
2.4.3. RNA Polymerase II and III machineries are associated with the human IGS	110
2.4.4. Potential origin of replication in IGS.....	111
2.4.5. CTCF association is not restricted near to the rDNA promoter but also present in the other regions of the IGS.....	111
2.4.6. Conservation of Alu elements in the primate rDNA	112
2.4.7. The limitations of ENCODE data to predict the function of the rDNA IGS	113
2.4.8. Comparison between the human and the yeast IGS phylogenetic footprinting analysis	114
2.4.9. Correlation between the increase size of the IGS and evolution of amniotes	114

3. Chapter 3: Characterization of the regions surrounding the human rDNA array:

The human rDNA flanking regions117

3.1. Introduction	119
3.1.1. The rDNA flanking regions sequences.....	119
3.1.2. Experimental strategy to characterize the rDNA flanking regions.....	120
3.2. Material and Methods.....	123
3.2.1. Sequencing and assembly of distal-rDNA and proximal-rDNA junction cosmids	123
3.2.2. Sequence mapping based screen for proximal-rDNA junctions.....	123
3.2.2.1. Data Acquisition.....	125
3.2.2.2. Reference sequence preparation	125
3.2.2.3. Pipeline	125
3.2.3. PCR amplification of the proximal-rDNA and distal-rDNA junction positions	126

3.2.3.1. Junction region amplification	126
3.2.3.2. Cloning and Transformation	127
3.2.4. Intra- and inter-chromosomal identity of the rDNA distal and proximal regions	129
3.2.5. Repeat content of the rDNA distal and proximal region contigs	129
3.2.6. Segmental duplication analysis of the rDNA distal and proximal contigs ..	129
3.2.7. Gene prediction pipeline for the rDNA distal and proximal contigs	130
3.3. Results.....	131
3.3.1. Verification of the proximal-rDNA junction	131
3.3.1.1. Extending linkage into the rDNA from the proximal region junction and searching various junction positions using cosmid sequences.....	131
3.3.1.2. Bioinformatics screen for proximal rDNA junctions.....	135
3.3.1.3. PCR amplification of the junction regions.....	136
3.3.2. Inter- and intra-chromosomal sequence conservation of the rDNA flanking regions	138
3.3.2.1. Intra-chromosomal sequence conservation of the rDNA distal and proximal regions.....	139
3.3.2.2. Inter-chromosomal sequence conservation of the rDNA distal and proximal flanking regions is high.....	140
3.3.2.3. Overall conservation of the distal region and the proximal region.....	145
3.3.3. Distal and Proximal contig construction.....	146
3.3.4. Characterization of the distal and proximal contigs.....	149
3.3.4.1. Repeat content of the distal and proximal contigs	149
3.3.4.2. A large inverted repeat in the distal contig	152
3.3.4.3. The level of segmental duplication in the proximal and distal contigs	153
3.3.4.4. Putative gene models in the distal and proximal contigs	156
3.4. Discussion.....	158
3.4.1. High conservation of the flanking regions across the acrocentric chromosomes	159
3.4.2. The proximal region is a segmental duplication hub	159
3.4.3. Putative genes in the masked and unmasked distal and proximal regions...	160
3.4.4. Significance of inverted repeat in the distal region.....	161
3.4.5. The flanking regions have a repeat content similar to the rest of the genome	162
3.4.6. The flanking regions boundary	163
3.4.7. The sequence of the short arm of acrocentric chromosomes beyond the identified rDNA flanking region sequences.....	164

3.4.8.	Computational challenges to study the rDNA flanking regions.....	164
4.	Chapter 4: Conclusions and Future Directions	167
4.1.	Conclusions	169
4.1.1.	The functional regions in the human IGS.....	169
4.1.2.	Characterization of the rDNA flanking regions.....	171
4.1.3.	The broader significance of the study.....	172
4.2.	Future Directions	172
4.2.1.	Verification of the IGS transcripts identified using publically available datasets	172
4.2.2.	Exploring the role of the transcripts from the IGS	173
4.2.3.	Verification of the identified origin of replication.....	174
4.2.4.	Role of the identified conserved regions	175
4.2.5.	Comparative analysis of human IGS transcripts.	175
4.2.6.	Role of Pol II in IGS transcription.....	176
4.2.7.	Phylogenetic footprinting of the rDNA flanking regions	176
4.2.8.	Role of the rDNA flanking regions in nucleolar formation/fusion.....	177
5.	Appendix I: Tables and Figures	179
6.	Appendix II: Statement of contributions.....	201
7.	Appendix III: Publication arising from this work.....	205
	References	217

Table of Figures

Figure 1.1: Schematic diagram of a human nucleolar organizer region (NOR) in an acrocentric chromosome.	4
Figure 1.2: A eukaryotic rDNA unit.	6
Figure 1.3: A complete human rDNA repeat unit.	10
Figure 1.4: Functional elements in the IGS of yeast, <i>Xenopus</i> , mouse, and human.	13
Figure 1.5: Schematic diagram showing the distal rDNA junction position.	19
Figure 1.6: Schematic diagram showing the proximal rDNA junction position.	20
Figure 2.1: Schematic overview of the identification of potential functional elements in the human IGS study	26
Figure 2.2: Schematic diagram of modified chromosome 21 reference sequence used for RNA-seq and CHIP-seq analysis of the human rDNA.	37
Figure 2.3: Microsatellite assembly using different size paired-end reads.	45
Figure 2.4: Primate phylogenetic tree showing the genera selected for human rDNA phylogenetic footprinting.	47
Figure 2.5: WGA contigs containing chimpanzee rDNA sequence.	53
Figure 2.6: WGA contigs containing gorilla rDNA sequence.	54
Figure 2.7: WGA contigs containing orangutan rDNA sequence.	56
Figure 2.8: WGA contigs containing gibbon rDNA sequence.	57
Figure 2.9: WGA contigs containing Macaque rDNA sequence.	59
Figure 2.10: WGA contigs containing marmoset rDNA sequence.	60
Figure 2.11: Gorilla BAC library filter CHORI-255 1A with signals for the rDNA positive BAC clones.	63
Figure 2.12: Orangutan BAC library filter CHORI-276 3F with signals for the rDNA positive BAC clones.	64
Figure 2.13: Gibbon BAC library filter CHORI-271 10A with signals for the rDNA positive BAC clones.	65
Figure 2.14: Verification of the presence of the rDNA unit in BAC clones.	66
Figure 2.15: Estimating the length of rDNA units in the gorilla BAC clones.	68
Figure 2.16: Variation between gorilla WGA rDNA and BAC clones gorilla rDNA.	69
Figure 2.17: Estimating the length of rDNA units in orangutan BAC clones.	70
Figure 2.18: Variation between the orangutan WGA rDNA and BAC clones orangutan rDNA.	71
Figure 2.19: Estimating the length of rDNA units in gibbon BAC clones.	72
Figure 2.20: Variation between gibbon WGA rDNA and BAC clones gibbon rDNA.	73
Figure 2.21: Estimating the length of the rDNA units in the macaque BAC clones.	74

Figure 2.22: Variation between macaque WGA rDNA and BAC clones macaque rDNA.	75
Figure 2.23: Estimating the length of the rDNA units in marmoset BAC clones.	76
Figure 2.24: Variation between marmoset WGA rDNA and BAC clones marmoset rDNA.	77
Figure 2.25: The complete chimpanzee rDNA repeat unit.	80
Figure 2.26: The complete gorilla rDNA repeat unit.	81
Figure 2.27: The complete orangutan rDNA repeat unit.	82
Figure 2.28: The complete gibbon rDNA repeat unit.	83
Figure 2.29: The complete macaque rDNA repeat unit.	84
Figure 2.30: The complete marmoset rDNA repeat unit.	85
Figure 2.31: Sequence similarity plot for human rDNA with five different primate species <i>viz.</i> chimpanzee, gorilla, orangutan, gibbon and macaque.	92
Figure 2.32: Sequence conservation plot for the rRNA coding regions.	93
Figure 2.33: Sequence conservation plot for the <i>cdc27</i> pseudogene.	94
Figure 2.34: The long poly(A)- and small poly(A)+ transcripts in the human IGS from different cell types.	97
Figure 2.35: Chromatin, transcription factor and transcript landscape of the IGS in embryonic cells H1-hESC.	99
Figure 2.36: Chromatin marks and TFs associated with conR-53.	100
Figure 2.37: Chromatin marks associated with conR-23 to conR-31.	101
Figure 2.38: Chromatin marks associated and TFs with conR-16.	102
Figure 2.39: Origin replication complex (ORC) binding in the HeLa-S3 cell type.	104
Figure 2.40: Pol machineries and related transcription factors that associate with the human IGS.	106
Figure 3.1: Schematic overview of the characterization of the rDNA flanking regions	122
Figure 3.2: Workflow for junction verification mapping pipeline.	124
Figure 3.3: Structure of distal region and proximal region clones around the rDNA junction	134
Figure 3.4: The positions of the primer pairs for the amplification of the proximal-rDNA and distal-rDNA junctions	137
Figure 3.5: Amplification of the flanking region-rDNA junction.	138
Figure 3.6: Inter-chromosomal variation in the near proximal region due to Alu elements and 147 bp ACRO1 repeat.	144
Figure 3.7: Average intra- and inter-chromosomal identities between distal and proximal flanking region clones	145
Figure 3.8: Scheme to construct the distal and proximal contigs.	147
Figure 3.9: Locations of the distal region clones in the distal contig.	148
Figure 3.10: Locations of the proximal region clones in the proximal contig.	149

Figure 3.11: Repeat composition of the distal and proximal contigs.....	150
Figure 3.12: Locations of novel and satellite repeats in the distal and proximal contigs. ...	151
Figure 3.13: HMM logo for the 138 bp ACRO138 repeats.....	152
Figure 3.14: Sequence similarity between the arms of the large inverted repeat in the distal contig.	153
Figure 3.15: Segmental duplication in the proximal and distal contigs.....	155
Figure 3.16: Gene models in the distal and proximal contigs.....	157
Appendix Figure 1: Chromatin, transcription factor and transcript landscape of the IGS in lymphoblastoid cell GM12878.	183
Appendix Figure 2: Chromatin, transcription factor and transcript landscape of the IGS in umbilical vein endothelial cell HUVEC.	184
Appendix Figure 3: Chromatin, transcription factor and transcript landscape of the IGS in adenocarcinomic cell A549.	185
Appendix Figure 4: Chromatin, transcription factor and transcript landscape of the IGS in cervical carcinoma cell HeLa-S3.....	186
Appendix Figure 5: Chromatin, transcription factor and transcript landscape of the IGS in leukaemia cell K562.	187

Table of Tables

Table 1.1: Chromosome number and Copy number of rDNA in different primate species.....	5
Table 1.2: Length of the rDNA coding region and intergenic spacer in different organism. ...	7
Table 2.1: List of histone modifications mapped to the human rDNA sequences	29
Table 2.2: List of transcription factors mapped to the human rDNA sequence	29
Table 2.3: The cell types included in this study	30
Table 2.4: Details of WGS data for the primates.....	32
Table 2.5: Comparative analysis of assemblers to evaluate their efficiency to assemble the human rDNA sequence.....	50
Table 2.6: Statistics of the potential chimpanzee rDNA contigs.....	52
Table 2.7: Statistics of the potential gorilla rDNA contigs.	54
Table 2.8: Statistics of the potential orangutan rDNA contigs.....	55
Table 2.9: Statistics of the potential gibbon rDNA contigs.....	57
Table 2.10: Statistics of the potential macaque rDNA contigs.....	58
Table 2.11: Statistics of the potential marmoset rDNA contigs.	60
Table 2.12: BAC filters screened to identify the rDNA containing BAC clones.....	62
Table 2.13: The number of rDNA reads represented by the WGA rDNA sequence.....	78
Table 2.14: The length variation between the WGA and BAC rDNA sequences of the six primate species.	78
Table 2.15: rDNA sequence comparison between human and the six primate species.....	86
Table 2.16: Repeat composition of the primate rDNA sequences.....	87
Table 2.17: Pairwise sequence comparisons showing the level of sequence conservation between human and ape Alu elements.	89
Table 3.1: The part of the flanking region reference used for mapping the WGS reads.....	125
Table 3.2: Primer pairs used for rDNA junction verification.....	127
Table 3.3: PCR protocols used for different primer pairs to amplify the junction region. ...	127
Table 3.4: Sequence comparison between the human rDNA and proximal junction rDNA.	132
Table 3.5: Sequence comparison between the human rDNA and segmentally duplicated rDNA fragments.	132
Table 3.6: Results for different steps of the junction mapping pipeline.....	135
Table 3.7: Sequence similarity matrix for the near-distal region cosmid and BAC clones..	141
Table 3.8: Sequence similarity matrix for the far distal region BAC clones.....	141
Table 3.9: Sequence similarity matrix for the near-proximal region cosmids and BAC clones.	143
Table 3.10: Segmental duplication comparison between the distal and proximal contigs. ...	155

Table 3.11: Putative gene models from the distal contig.....	157
Table 3.12: Putative gene models from the proximal contig.....	158
Appendix Table 1: Assembly statistics for the primate whole genome assemblies.	181
Appendix Table 2: Coordinates for the conserved regions in the human IGS.....	181
Appendix Table 3: Sequencing statics of the distal and proximal cosmids.....	189
Appendix Table 4: Assembly statistics for the distal and proximal cosmid assemblies.....	189
Appendix Table 5: Sequence similarity matrix for the distal region BAC clones.....	190
Appendix Table 6: Repeat statistics of the distal contig.....	192
Appendix Table 7: Repeat statistics of the proximal contig.....	193
Appendix Table 8: Segmentally duplicated regions from the proximal contig.....	194
Appendix Table 9: Segmentally duplicated regions from the proximal contig.....	197
Appendix Table 10: Multiple sequence alignment for the ACRO138 repeat.....	198
Appendix Table 11: Consensus sequence of ACRO138 repeats.....	199

Abbreviations

APC	Anaphase-promoting complex
BAC	Bacterial artificial chromosome
BCM	Baylor College of Medicine
Bdp1	Transcription factor TFIIIB component B homolog
BI	Broad Institute
BLAST	Basic Local Alignment Search Tool
bp	Base pair
BRF1	Transcription factor IIIB 90 kDa subunit
Brf2	b-related factor 2
cdc27	cell division cycle 27
ChIP-seq	Chromatin immunoprecipitation-sequencing
CHORI	Children's Hospital Oakland Research Institute, USA
chr	Chromosome
CRA	Celera Genomics
CTCF	CCCTC-Binding factor
EDTA	Ethylenediaminetetraacetic acid
<i>et. al.</i>	and others
ETS	External transcribed spacer
FIGE	Field-inversion gel electrophoresis
g	Gram
H1-hESC	Human embryonic stem cells line H1
HCl	Hydro chloric acid
HeLa	Henrietta Lacks
hrs	Hours
HUVEC	Human umbilical vein endothelial cells
IGS	Intergenic spacer
ITS	Internal transcribed spacer
JCVI	J. Craig Venter Institute
kb	Kilo base pair = 1000bp
L	Liter
LINE	Long Interspersed Nuclear Element
LTR	Long terminal repeat
min	Minute
ml	Mili liter
MSA	Multiple sequence alignment
NaCl	Sodium chloride
NAHR	Non-allelic homologous recombination
NaOH	Sodium hydroxide
NOR	Nucleolar organizer region
NoRC	Nucleolar remodelling complex
°C	degree Celsius
ORC	Origin replication complex
ORI	Origin of replication
POL I	RNA polymerase I
POL II	RNA polymerase II
POL III	RNA polymerase III
pRNA	Promoter RNA
rDNA	Ribosomal DNA
RFB	Replication fork block
RNA-seq	RNA Sequencing
SC	Sanger center
sec	Second

SINE	Short Interspersed Nuclear Elements
SL1	Selectivity factor 1
SSC	Saline-sodium citrate
TBE	Tris/Borate/EDTA
TBP	TATA Binding Protein
TBS	Tris-buffered saline
UBF	Upstream binding factor
V/cm	voltage/centimetre
WGA	Whole genome assembly
WGS	Whole genome sequencing

[Blank Page]