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

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 thanks 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

Table of Figures

Table of Tables

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
et. al.	and others
ETS	External transcribed spacer
FIGE	Field-inversion gel electrophoresis
g	Gram
H1-hESC	Human embryonic stem cells line H1
HCl	Hydro choleric 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
^{o}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]

Chapter 1

Introduction

[Blank page]

1.1. Nucleolar organizer region

The nucleolar organizer region (NOR) is the chromosomal location in eukaryotes around which the nucleolus is formed. The nucleolus is a sub-nuclear, non-membranous structure which is defined by a heterochromatic shell (Mosgoeller 2004). It is the site of ribosome biogenesis and therefore essential for the cell survival. In the late 18th century Fontana described an oval shaped body in eel saliva cells (adapted from Mosgoeller 2004) that was later (in 1836) termed the nucleolus by Gabriel Gustav (as cited in Mosgoeller 2004). Emil Heitz was the first to report the chromosomal context of the nucleolus using *Zea mays* (as cited in Mosgoeller 2004). His work was further elaborated on by Barbara McClintock and she termed the genomic region around which the nucleolus is formed as the "nucleolar organizer" (McClintock 1934), which was later modified to "nucleolar organizer region". Although the location of the nucleolar organizer region was demarcated on the genome, the underlying sequence and function was still not known. In late 1950s densely stained particles were observed in the nucleoli that were speculated to be ribosomes (as cited in Birnstiel and Hyde 1963). In the early 1960s Edström *et al.* (1960) and Birnstiel *et al.* (1963) demonstrated that the nucleolar RNA composition is similar to that of cytoplasmic RNA, and differs to that of the nucleus. The discovery that the sizes of the RNA molecules in the nucleolus are the same as cytoplasmic RNA lead to the idea that ribosomes are stored in the nucleolus in addition to their known abundant presence in the cytoplasm (Birnstiel *et al.* 1963). However, it was not clear if ribosomes are synthesized in the nucleolus or are just stored there (McConkey and Hopkins 1964). Ritossa and Spiegelman (1965) had first reported that multiple copies of the ribosomal DNA (rDNA) units were arranged in clusters in the nucleolus of *Drosophila melanogaster* using an RNA-DNA hybridization technique, thus demonstrating that the NOR contains rDNA. This finding also established that ribosomes are not stored but are synthesized in the nucleolus. The presence of rDNA in the nucleoli of other organisms (Phillips *et al.* 1971) verified that the rDNA is a universal building block of the NORs.

The primary function of the nucleolus is ribosome biogenesis, which involves transcription of rDNA to produce pre-ribosomal RNA (pre-rRNA) therefore the general interpretation has been that NORs only consist of rDNA arrays. However, there is direct and indirect evidence that not only the rDNA arrays but also the regions surrounding them are part of the nucleolus. In human, NORs are present on the short arm of five acrocentric chromosomes. Quantification of the number of human acrocentric chromosome short arms in the nucleolus shows that other regions of these arms are also part of the nucleolus, in addition to the rDNA

arrays (Kalmárová *et al.* 2007). In human, satellite III repeats and KpnI elements (a LINE element with a K*pn*I restriction site) present on the short arm of acrocentric chr 15 and chr 21, respectively, were found to be associated with the nucleolus but are not part of the rDNA unit (Kaplan *et al.* 1993). Additional evidence for rDNA flanking regions being involved in the nucleolus comes from genetic variants of *Saccharomyces cerevisiae* (yeast) and *Zea mays* (maize) where the rDNA arrays were translocated to other chromosomal locations (McClintock 1934; Oakes *et al.* 1998; Oakes *et al.* 2006). In both organisms, in addition to a nucleolus at the new rDNA location, a small, diffuse nucleolus is formed around the original rDNA array position (McClintock 1934; Oakes *et al.* 1998; Oakes *et al.* 2006), suggesting that the flanking regions play some role in nucleolar formation. Further, the centromeres of the acrocentric chromosomes are found to be colocalized with nucleoli that were isolated by nucleolar purification in human (Stahl *et al.* 1976). Together these studies suggest that the term NOR should be redefined as *"the region around which the nucleolus is formed that consists of two components: a tandem array of rDNA units; and the regions surrounding this array (the flanking regions)"* (Figure 1.1). I will be using this as the definition of NORs throughout this thesis.

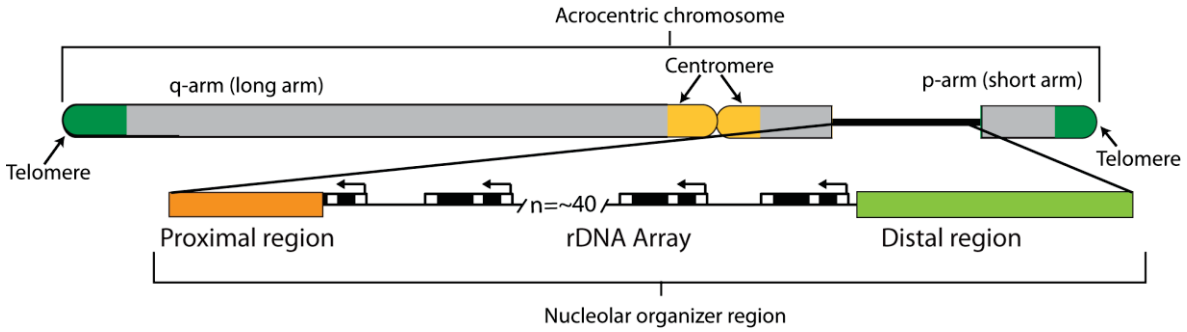


Figure 1.1: Schematic diagram of a human nucleolar organizer region (NOR) in an acrocentric chromosome.

The NOR is consists of an rDNA array (tandem black-white boxes) and the region surrounding it i.e. the rDNA flanking regions. The proximal flanking region (orange box) is towards the centromeric end and the distal region (green box) is towards the telomeric end of the rDNA array.

The ribosomal DNA: The rDNA is a multi-copy gene that encodes ribosomal RNA (rRNA), the core RNA component of the ribosome. In almost all eukaryotes characterised to date, a number of rDNA units are clustered to form tandem arrays that are arranged in a head-to-tail manner. However, in human a fraction of rDNA units have been reported to be arranged as inverted repeats (Caburet *et al.* 2005). The rDNA arrays are distributed on one or more chromosomes depending on the species. In primates, the number of NORs and rDNA copy

number vary depending on the individual (Table 1.1). A single rDNA unit consists of an rRNA precursor coding region and an intergenic spacer (Figure 1.2).

Table 1.1: Chromosome number and Copy number of rDNA in different primate species.

Organism	NOR bearing chromosomes	Estimated rDNA copy number in a diploid genome	Reference
Colobus polykomos (colobus monkey)	Marker Chromosome	68	(Henderson *et al.* 1977)
Tupaia glis (tree shrew)	27,28,29,30	739	(Henderson *et al.* 1977)
Lemur fuluvis (lemur)	group 20-30[a]	230	(Henderson *et al.* 1977)
Pithecia pithecia (saki)	21,22,23	546	(Henderson *et al.* 1977)
Saguinus nigricollis (marmoset)	14,17,18,19,20	744	(Henderson *et al.* 1977)
Ateles geoffroyi (spider monkey)	Marker Chromosome	Not determined	(Henderson *et al.* 1977)
Macaca mulatta (rhesus monkey)	20	140	(Henderson *et al.* 1974b)
Symphalangus syndactylus (siamang gibbon)	Single acrocentric chromosome	Not determined	(Henderson *et al.* 1976)
Hylobates lar (gibbon)	15	180	(Warburton *et al.* 1975)
Pongo pygmaeus albei (Sumatran orangutan)	11,12,13,14,15,16,17,22,23	422	(Henderson *et al.* 1979)
Gorilla gorilla beringei (mountain gorilla)	22,23	270	(Henderson *et al.* 1976)
Gorilla gorilla gorilla (gorilla)	22,23	Not determined	(Tantravahi *et al.* 1976)
Pan paniscus (pygmy chimpanzee)	14,15,17,22,23	414-488	(Henderson *et al.* 1976)
Pan troglodytes (chimpanzee)	14,15,17,22,23	488	(Henderson *et al.* 1974a; Tantravahi *et al.* 1976)
Homo sapiens (human)	13,14,15,21,22	320-460	(Henderson *et al.* 1972; Schmickel 1973; Tantravahi *et al.* 1976; Krystal and Arnheim 1978)

[a] Chromosomes were too small to identify.

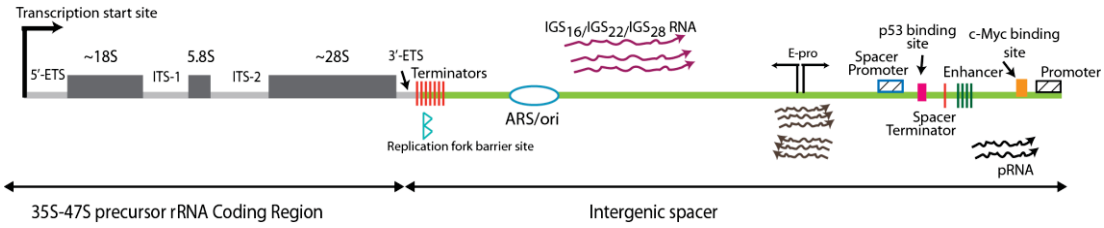


Figure 1.2: A eukaryotic rDNA unit.

A single rDNA unit consists of a precursor rRNA coding region (grey line) and an intergenic spacer (green line). The IGS harbours several functional elements (shown on the green line) and transcribes several functional noncoding RNAs (wiggly lines) that are described in detail in Section 1.4.3. The IGS functional elements i.e. rDNA promoter (shaded box with black outline) and terminators (vertical red lines), spacer promoter(shaded box with blue border) and terminator(vertical red line) , enhancer (vertical green lines), replication fork barrier (blue triangles), origin of replication (ARS/ori; blue bubble), protein binding sites (pink [p53] and orange[c-Myc] boxes), E-pro (vertical lines with both directions) shown in the diagram are a composite from different organisms.

The coding region consists of regions encoding the 18S, 5.8S and 25-28S rRNA species, and these are transcribed together to form a single pre-rRNA transcript. The coding region is transcribed by RNA polymerase I (Pol I), and the pre-rRNA transcript is post transcriptionally spliced and modified to form mature 18S, 5.8S and ~28S rRNA. The size of coding region varies from ~8 kb in yeast to ~13 kb in human (Table 1.2). The intergenic spacer (IGS) or ribosomal spacer is the region between coding regions of two adjacent rDNA units. The size of the IGS differs widely between species, ranging from ~2.5 kb in yeast to ~30 kb in human (Table 1.2). The IGS harbours several functional elements that include rRNA transcriptional regulators (a promoter and terminator), together with gene independent noncoding functional elements (NOCs), such as an origin of replication, a replication fork barrier, a transcription enhancer, and a bidirectional noncoding promoter (Figure 1.2). Several regulatory long noncoding RNAs are also transcribed from the IGS (Figure 1.2).

Although multiple copies of the rDNA are present in the genome, not all of them are transcriptionally active (Conconi *et al.* 1989; Lucchini and Sogo 1992; Dammann *et al.* 1995). The rDNA units can be divided into active and silent units depending on their transcriptional status (Hamperl *et al.* 2013). The silent rDNA units can be further divided into two types: units that are poised to be transcribed but are not actually transcribed and fully epigenetically silent copies. Moreover, not all the rDNA arrays take part in

transcription, with some entire arrays containing only silent rDNA units (Roussel 1996; McStay and Grummt 2008). The number of active rDNA units is not constant in a cell and can change depending on its ribosome requirement.

The rDNA flanking regions: I have defined NORs as including the region flanking the rDNA. However, in most species (including human) the flanking regions are poorly characterized. The flanking region on the telomere side of the rDNA array is called the distal flanking region, while the region on the centromere side is called the proximal flanking region (Figure 1.1).

Table 1.2: Length of the rDNA coding region and intergenic spacer in different organism.

Organism	Approximate length of coding region (kb)	Approximate length of intergenic spacer (kb)	Reference
Saccharomyces cerevisiae (yeast)	8	2.5	(Georgiev *et al.* 1981)
Drosophila melanogaster (fruit fly)	8	4.1 to 5.1	(Tautz *et al.* 1988)
Xenopus laevis (African clawed frog)	8	3.4 or 5.7[a]	(Boseley *et al.* 1979)
Mus musculus (mouse)	13.5	31	(Grozdanov *et al.* 2003)
Homo sapiens (human)	13 kb	30 kb	(Gonzalez and Sylvester 1995)

[a] intra-individual spacer length variation

1.2. Status of the NOR in human genome assembly

Despite the much publicized fact that the sequence of almost the entire human genome is known (99.9%), several large gaps (~240 Mb) are present in the assembled genome sequence (Eichler *et al.* 2004). One of the major gaps is the absence of the sequences from the short arms of the five acrocentric chromosomes (Lander *et al.* 2001; Eichler *et al.* 2004). Consequently, the NORs are also absent from the current human genome assembly. It was thought that the short arms of the acrocentric chromosomes are mainly heterochromatic, highly duplicated, and devoid of genes (Lander *et al.* 2001; Venter *et al.* 2001), and the only assigned function they currently have is ribosome biogenesis. Since the focus of the human

genome sequencing effort was on identifying the euchromatic gene rich regions, there has not been a concerted attempt to complete the short arms of acrocentric chromosomes (Eichler *et al.* 2004). In later chromosome-specific projects, the focus remained on closing the gaps in the draft assembly that are in gene rich euchromatic regions. Since the long arms of the acrocentric chromosomes were thought to contain all the genes, the short arms remained unattended in these projects as well (Dunham *et al.* 1999; Hattori *et al.* 2000; Heilig *et al.* 2003; Dunham *et al.* 2004; Zody *et al.* 2006). The current knowledge of human NOR sequence is based on the sequence of a complete rDNA unit and small fragments of sequences from the rDNA flanking regions. The sequences of the flanking region away from the junction positions are still not known. Similar to human, the assemblies for other primates also do not include the NORs (Carbone *et al.* 2006; Gibbs *et al.* 2007; Scally *et al.* 2012). Further, the complete rDNA sequence for other primate is not known. In the absence of sequences for the rDNA and flanking regions, the nucleolar organizer regions from human and primates remain uncharacterized.

1.3. rDNA homogeneity in the primates

The multiple copies of the rDNA units are almost identical to each other in an individual. This phenomenon of homogeneity of the rDNA units is known as concerted evolution and is the result of a process called homogenization occurring in the rDNA (Elder and Turner 1995). In human and other apes, the rDNA arrays are distributed among acrocentric chromosomes (Arnheim *et al.* 1980). The rDNA units in different chromosomes are almost identical to each other, demonstrating that the rDNA arrays are also not evolving independently but in concerted way. Studies targeting the restriction site polymorphisms in the rDNA have shown that these rDNA variants are not restricted to a single rDNA array but are distributed among the acrocentric chromosomes (Krystal *et al.* 1981). The presence of the same variant in multiple rDNA copies that are distributed on different acrocentric chromosomes suggests that recombination between nonhomologous chromosomes in the rDNA array regions must be occurring. However, the majority of the rDNA units in a cell contain the same restriction mapping profile, further supporting the idea that the rDNA units are almost identical across the short arms of the acrocentric chromosomes (Arnheim *et al.* 1980; Krystal *et al.* 1981).

1.4. Human ribosomal DNA

In a human diploid genome, approximately 400-600 copies of the rDNA repeat unit are present (Tantravahi *et al.* 1976; Stults *et al.* 2008). The rDNA units are distributed as tandem arrays on the short arms of the five acrocentric chromosomes (chromosomes with arms of unequal length) i.e. chr 13, chr 14, chr 15, chr 21 and chr 22 (Henderson *et al.* 1972). In a single array the number of rDNA units varies from 2 to >140 (Stults *et al.* 2008). The sequences of all the rDNA units are almost identical to each other (Krystal *et al.* 1981). As in other species, the rDNA is transcribed in human by Pol I, and this generates a 47S pre-rRNA transcript that is post-transcriptionally processed to generate the mature 18S, 5.8S, and 28S rRNA species. Pol I is recruited to the promoter by a pre-initiation complex (PIC) (Budde and Grummt 1999). The PIC in human consists of two factors: an upstream binding factor (UBF) and a promoter selective factor (SL1). SL1 itself consists of a TATA-binding protein (TBP) and four TBP-Associated Factors: TAF_I41, TAF_I48, TAF_I63 and TAF_I110 (Gorski *et al.* 2007). UBF first binds to the promoter to facilitate the binding of SL1. SL1 in turn recruits Pol I to the promoter to initiate transcription.

A single rDNA unit is ~43 kb in length and is divided into a ~13 kb 47S pre-rRNA coding region (referred as the coding region from here onwards) and a 30 kb IGS (Figure 1.3). According to convention, position 1 of the human rDNA is the start of the coding region.

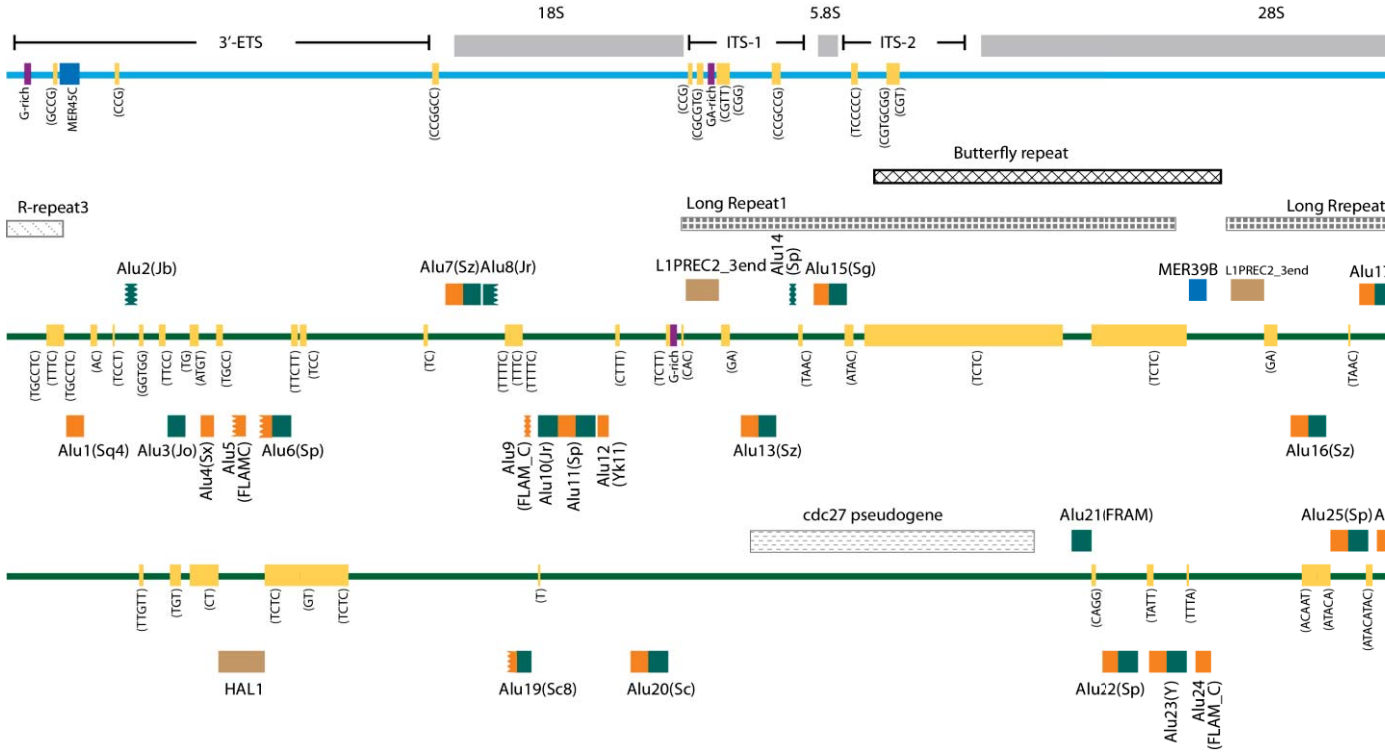


Figure 1.3: A complete human rDNA repeat unit.

Each human rDNA unit has a 13,357 bp coding region (blue line) and a 30,615 bp IGS (green line).The coding region encodes contains several repeat elements including Alus (monomers as orange and green boxes), microsatellites (yellow boxes), LIN complexity regions (purple boxes). The names of repeats element are indicated next to them. The IGS also has three other sequ boxes), butterfly repeats (crisscross boxes) and long repeats (checked boxes). It also has a cdc27 pseudogene (dashed box). represented as serrate margins. The elements above the rDNA are on the forward strand and elements below the rDNA are on nomenclature is based on the sequence of the forward strand and for clarity of the figure, microsatellites are placed on the rDN

1.4.1. Human 47S pre-rRNA coding region

The coding region consists of 18S, 5.8S and 28S rRNA encoding sequences that are separated from each other by internal transcribed spacer regions (ITS-1 and ITS-2) and are flanked by external transcribed spacer regions (5′-ETS and 3′-ETS) (Figure 1.3). The 18S and 5.8S rDNA sequences are highly conserved among vertebrates, and beyond vertebrates some regions are comparatively more conserved than others (Hillis and Dixon 1991). The 28S rDNA is composed of conserved regions that are interrupted by variable regions (also called divergent, D-domains or expansion segments) (Clark *et al.* 1984; Hassouna *et al.* 1984; Gonzalez *et al.* 1985). The conserved regions are highly conserved among the vertebrates, but the variable regions show length polymorphisms (Gorski *et al.* 1987). These variable regions even show sequence polymorphism between different human cell types (Leffers and Andersen 1993).

1.4.2. Human rDNA IGS

Adjacent rDNA coding regions are separated by an IGS. Like the rest of the human genome, the IGS also contains a number of repeat elements, such that 43.4% of the IGS is repetitive. The most abundant repeat element in the IGS are Alu elements that alone make up 13.3% of the IGS (Gonzalez and Sylvester 1995). Alu are from a class of retrotransposons known as SINE elements and are specific to primates. They are named because of the presence of an A*lu*I restriction site. A typical Alu element is about 280 bp in length and has a dimeric structure. Each monomer is made of a derivative of an ancestral 7SL RNA gene. The Alu element monomers are diverged from each other and are connected by a poly-A linker (Deininger 2011). Alu elements also have a 3′ poly-A tail of variable length. The IGS contains both complete (dimeric) and fragmented Alu elements (Figure 1.3). The fragmented Alu elements are mostly present near the rRNA terminator end of the IGS while complete Alus are dispersed throughout the IGS (Figure 1.3). Studies based on sequence hybridization have shown that several Alus in the human IGS are conserved across the IGS of apes and old world monkeys (Gonzalez *et al.* 1989; Gonzalez *et al.* 1993). In addition to Alu elements, other retrotransposon elements i.e. long interspersed elements (LINEs), and long terminal repeats (LTR) together with microsatellites (3-6 bp monomers that are tandemly repeated several times) (Figure 1.3) are also present in the human IGS. The LINE and LTRs elements present in the IGS are fragmented and together make up ~5% of the IGS. The human IGS also contains a 2,332 bp cdc27 pseudogene (Figure 1.3) that is conserved among the IGS of apes but is not found in old world monkeys (Gonzalez *et al.* 1989). Finally the IGS also has three other sequence elements *viz.* an R-repeat, a long repeat and a butterfly repeat (Figure

1.3), all of which were described by Gonzalez *et al.* (Gonzalez and Sylvester 1995). Two to three copies of the R-repeat are present in the human IGS. Each repeat is divided into a highly conserved ~400 bp GC rich block and a variable 150-322 bp TC rich block. They are present downstream of the coding region, at the start of the IGS. Two copies of the long repeats are present in the IGS. They are consist of different repeat elements i.e. a LTR element, two Alu elements and different microsatellites, and are present further into the IGS than the R-repeats. Butterfly repeats are 320 bp repeat element that are named because of the pattern with which they appear in Sanger sequencing gels. They are present in the same region of the IGS as the long repeats.

1.4.3. Functional elements in the human IGS

It is thought that the human IGS contains a number of functional elements, and these include transcription regulators (promoter and terminators) together with gene independent functional elements (NOCs), such as an origin of replication, a replication fork barrier site, noncoding transcripts and putative protein binding sites (Learned 1983; Learned *et al.* 1986; Sáfrány *et al.* 1989; Sylvester *et al.* 1989; Kern *et al.* 1991; Little *et al.* 1993; Yoon *et al.* 1995; Gencheva and Russev 1996; Grandori *et al.* 2005). Compared to other model systems, namely yeast and *Xenopus,* in which the IGS functional elements have been well characterized, far less is known about these functional elements in the human IGS. In the following subsections, the functional elements thought to be present in the human IGS are described and are compared with yeast, *Xenopus* and mouse (Figure 1.4).

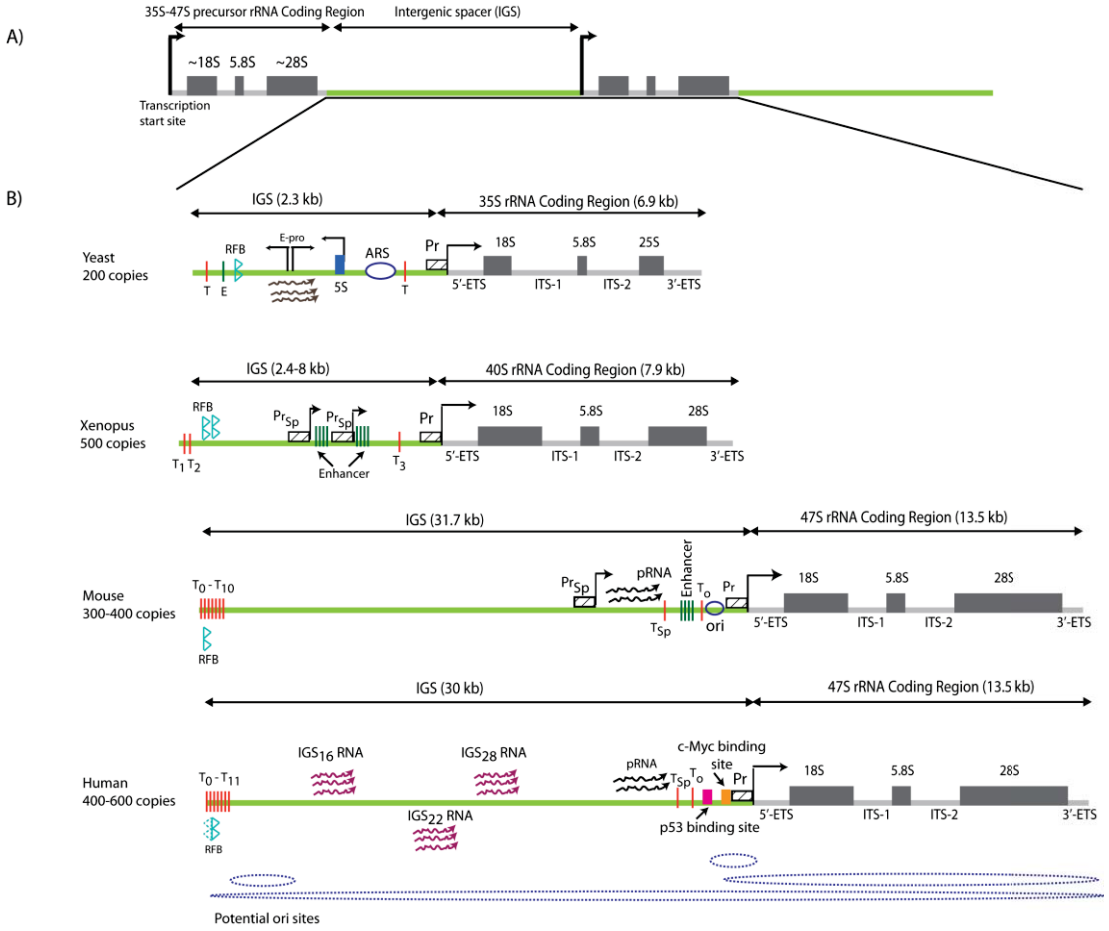


Figure 1.4: Functional elements in the IGS of yeast, *Xenopus*, mouse, and human.

A) Tandem arrangement of two rDNA units. An rDNA unit has a precursor rRNA coding region (grey line) and intergenic spacer (green line). The coding region encodes 18S, 5.8S and 28S rRNA (grey box), and these coding units are separated by internal spacer regions (ITS-1 and 2). B) The rDNA units of yeast, Xenopus, mouse and human. The known functional elements in the IGS viz. promoter (Pr), spacer promoter (Pr_{sp}), terminator(T_n), transcription enhancer, replication fork block site (RFB), origin of replication (ori/ARS), noncoding RNAs (wiggly arrows), protein binding sites (red and orange boxes), and the 5S rRNA gene (blue box) are shown around the IGS of each species. The sizes of the coding region and the IGS are shown above them. The name and rDNA copy number of each species are indicated to the left of each rDNA unit (adapted from Hamperl et al. 2013).

1.4.3.1. Transcription regulators

The IGS contains the promoter for rDNA transcription upstream of the coding region. In human, this promoter is made up of two domains: a ~45 bp core promoter element (CPE) that includes the transcription start site; and an upstream promoter element (UPE) also known as the upstream control element (UCE) (Haltiner *et al.* 1986). The CPE is sufficient for *in vitro* transcription while both domains are required *in vivo*. In human, the position of CPE is -45 to +18 while the UCE is located -156 to -107 (+1 is the initiation start site) (Learned 1983; Haltiner *et al.* 1986). These two elements are functionally nonequivalent. The CPE is essential for transcription and most mutations in it cause a decrease in transcription. The UPE includes a 21-bp stretch of DNA that is highly conserved between human, mouse, and rat. It plays an important role in modulating the efficiency of transcription but is not required for transcription *in vitro* from the human ribosomal promoter (Haltiner *et al.* 1986). Thus, the CPE differs from the UPE in its qualitative as well as its quantitative effects on transcription, since it determines the transcription start site. The position of the promoter and presence of a CPE and a UPE is conserved from yeast to human (Moss *et al.* 2007). The IGS in both mouse and *Xenopus* have a spacer promoter in addition to the rDNA promoter. In mouse, the spacer promoter is associated with transcription of a noncoding RNA known as the promoter RNA (pRNA). The pRNA is described in detail in section 1.4.3.6.

The human IGS contains an RNA Pol I terminator that consists of a 10-bp conserved sequence motif flanked with pyrimidine rich sequences (Bartsch *et al.* 1987; Pfleiderer *et al.* 1990). The terminator is also termed the "*Sal* box" due to the presence of an internal *Sal*I restriction site. Multiple copies of this terminator are arranged in a cluster downstream from the 28S rDNA (T_1-T_{10}) and one terminator is present proximal to the promoter (T_0). Termination occurs in human 360 bp downstream from the 3′ end of the 28S rDNA. Most transcribing Pol I molecules terminate at the first terminator, however the remaining terminators can serve to terminate polymerases that have managed to read through previous terminators. Similar to the human IGS, mouse IGS also has multiple terminators downstream of the rDNA coding region and a terminator upstream next to the promoter region (Németh *et al.* 2012). In mouse, the IGS also has a terminator that is known as the spacer terminator (T_{sp}) (Grummt *et al.* 1986). This is independent of rDNA transcription and terminates the transcription of the pRNA. In human, an uncharacterized spacer terminator 646 bp upstream of the transcription initiation site has also been reported (Németh *et al.* 2012).

1.4.3.2. Origin of replication

The presence of an origin of replication in the human rDNA is still not well established. The rDNA array covers a long stretch of the chromosomes (up to ~5 Mb) and therefore for faithful replication it is assumed that an origin of replication within the rDNA array is necessary. Supporting this conjecture, the inter origin distance in human is ~40 kb (Guilbaud *et al.* 2011), and since the rDNA array varies from ~80 kb to ~5 Mb, presence of an origin of replication inside the rDNA array would be necessary to maintain this level of origin spacing. Further, the rDNA units of both yeast and mouse have been found to contain an origin of replication in the IGS (Gögel *et al.* 1996; Muller *et al.* 2000). Unlike yeast, where the origin of replication [also known as the autonomously replicating sequence (ARS)] has a 13 bp consensus sequence, the origins of replication in mammals are not thought to be sequence specific (Muller *et al.* 2000) and therefore it is difficult to demarcate their positions accurately. Studies have reported various positions for the human IGS origin of replication using several different techniques (including DNA fibre analysis, and 2D gel electrophoresis). Specifically, Lebofsky & Bensimon (2005) have reported there is no preferential site and that replication can initiate anywhere in the rDNA. Little *et al.* (1993) concluded that there is no specific site that acts as an origin of replication, but that replication can start anywhere within the IGS. Gencheva *et al.* (1996) reported that two origins of replication are located in the IGS, one near the promoter region and the second next to the terminators. Coffman *et al.* (2006) have reported a preferential (at position around 37,371-37,728) and two additional origins of replication (at positions around 35,398-35,572 and 39,122-39,445) in the human IGS. Finally, Dimitrova (2011) has recently shown that the position of origin of replication depends on the cell cycle phase: in early S-phase replication can start at any position of the rDNA, while in the late S-phase it starts in the IGS at different locations. In mouse the origin of replication has been reported to be located upstream of the coding region, next to the promoter (Gögel *et al.* 1996), while in yeast a single origin is found near to the 5S rRNA gene in the IGS (Muller *et al.* 2000). Contrary to yeast and mouse, no fixed position has been reported for the origin of replication in *Xenopus* (Hyrien and Mechali 1993). In the *Xenopus* embryo replication reported to start randomly in the IGS (Hyrien and Mechali 1993)

1.4.3.3. Replication fork barrier (RFB) site

The RFB is a sequence in the IGS that arrests the movement of replication fork in the direction opposite to that of rDNA transcription to prevent the collision of the replication and Pol I transcription machineries. In human, the RFB site in the rDNA is thought to coincide with the terminators and is bipolar i.e. stops the replication fork in both directions (Little *et*

al. 1993). Although the RFB site in human is bipolar it stops the replication fork moving in the direction opposite to transcription more efficiently than the fork moving in the same direction (Little *et al.* 1993). Similar to human the RFB in mouse coincides with the terminator (López-Estraño *et al.* 1998), while in yeast and *Xenopus* the RFB is close to but independent of the terminator (Brewer *et al.* 1992; Wiesendanger *et al.* 1994). The IGS RFBs in mouse, *Xenopus* and yeast are all polar, unlike in human (Brewer *et al.* 1992; Wiesendanger *et al.* 1994; López-Estraño *et al.* 1998).

1.4.3.4. rDNA transcription enhancer

A region in the IGS has been reported to function as an enhancer in several species, although its presence is still to be established in human. In yeast, a 190 bp region 2 kb upstream of the transcription initiation site was reported to be an enhancer (Elion and Warner 1984), but later it was found that deletion of entire enhancer region does not affect growth or rRNA synthesis (Wai *et al.* 2001). In *Xenopus,* 6-12 copies of a 60/81 bp element (the 81 bp element is the same as the 60 bp element, but with an additional 21 bp added on) present between the two copies of the rDNA promoter act as enhancer (Pape *et al.* 1989). This 60/81 bp block enhances the rDNA transcription. These contradictory results for yeast and *Xenopus* provide no firm support either way for whether there is an enhancer in the human IGS.

1.4.3.5. Putative protein binding sites

The human IGS has binding sites for proteins that have roles in rDNA transcriptional regulation. The oncogene protein c-Myc is a transcription binding factor that plays roles in a variety of physiological processes including cellular growth, proliferation, loss of differentiation, and cell death (Dang 1999; Grandori *et al.* 2000). c-Myc is upregulated in many cancerous cells and promotes cell proliferation (Lutz *et al.* 2002). In the rDNA, c-Myc binds next to the rDNA promoter where it facilitates the binding of Pol I to the rDNA promoter (Grandori *et al.* 2005). Increased c-Myc binding to the rDNA is correlated with increased rDNA transcription.

Another important protein factor that has been associated with the rDNA is p53. p53 is tumour suppressor protein and plays roles in genome stability by promoting DNA repair and apoptosis (Oren 2003). It has a highly sequence specific binding site, and a putative p53 binding site has been found in the IGS (Kern *et al.* 1991).

1.4.3.6. Noncoding transcripts from the IGS

Several noncoding RNAs that act both as rDNA regulators and as NOCs have been reported to be transcribed from the IGS in mammals (Jacob *et al.* 2012). In human and mouse a transcript that is post-transcriptionally modified to form a 150 bp noncoding RNA known as the pRNA is transcribed from a region 2 kb upstream of the rDNA transcription start site (Mayer *et al.* 2006; Mayer *et al.* 2008). The pRNA plays an important role in rDNA silencing by facilitating the binding of a complex called the nucleolar remodelling complex (NoRC) near to the rDNA promoter (Santoro and Grummt 2001; Mayer *et al.* 2008). NoRC induces transcriptional silencing of rDNA units by promoting the association of heterochromatin histone modifications on the rDNA (Mayer *et al.* 2008). In addition, a noncoding RNA antisense to the entire human rDNA 48S rRNA coding region has been also reported to down-regulate rDNA transcription (Bierhoff *et al.* 2010). This transcript antisense to the coding region is reported to increase the heterochromatic histone modification H4K20me3 at the promoter region and therefore decrease the rDNA transcription (Bierhoff *et al.* 2010). Noncoding RNA from the IGS has also been shown to play a role in the sequestration of proteins in the nucleolus that has a specific amino acid motif known as nucleolar detention sequence (Audas *et al.* 2012).

Recently it has been shown that the human IGS produces a ~400 bp transcript known as IGS_{28}RNA from position ~28 kb of the rDNA during acidosis together with anaerobic condition. Anaerobic is marked by reduction in energy supply and therefore can be deleterious for cell. However, cells have pH sensitive mechanism that reduces the rDNA transcription and therefore the energy requirement of cell (Mekhail *et al.* 2004; Mekhail *et al.* 2006). IGS_{28} RNA immobilizes VHL, a tumour suppressor protein that is over expressed during acidosis in anaerobic conditions (Audas *et al.* 2012). VHL is thought to play role in the reduction rRNA transcription and therefore reduces the energy requirement of the cell (Mekhail *et al.* 2004; Mekhail *et al.* 2006). The IGS also produces additional noncoding RNAs, IGS_{16}RNA and IGS_{22}RNA, in response to heat shock (Audas *et al.* 2012). These transcripts bind to another nucleolar protein, Hsp70 (Audas *et al.* 2012), a heat shock protein that prevents aggregation of pre-existing proteins and facilitates folding of newly synthesised peptides (Mayer and Bukau 2005). During heat shock condition, a translocation of the Hsp70 to nucleolus has been observed. It has been thought that Hsp70 may have a role in protecting nucleolar protein (Pelham 1984; Welch and Feramisco 1984; Welch and Suhan 1986).

1.5. The chromatin state of the rDNA in human

The chromatin markers associated with the rDNA units are thought to govern the accessibility of the rDNA to transcription regulatory factors and therefore are thought to play a major role in defining whether the rDNA units are active or inactive (McStay and Grummt 2008). Active rDNA units have been shown to be associated with the histone modifications H3K4me3, H2A.Z and H3K27ac (Grummt 2007; McStay and Grummt 2008) that are enriched in the actively transcribed regions of the genome (Barski *et al.* 2007). In contrast, the inactive rDNA units are associated with the repressive histone modifications H3K27me3 and H4K20me3 (Grummt 2007; McStay and Grummt 2008). The histone modifications are interchangeable depending on the cellular conditions and demand for rRNA in the cell (Tanaka *et al.* 2010). A variety of transcription factors also modulates the activity of the rDNA, and these include UBF and the CCCTC-binding factor (CTCF) (van de Nobelen *et al.* 2010; Hernández-Hernández *et al.* 2012). UBF is essential for the recruitment of Pol I machinery to the rDNA promoter, hence is a key player in rDNA transcription (Grummt 1999). CTCF, a zinc finger protein, acts as an epigenetic regulator in the rDNA. Initially, CTCF was reported to act as an insulator in the rDNA that binds near to the rDNA promoter and inhibits rDNA transcription (Torrano *et al.* 2006). However, a recent study has shown that CTCF promotes rDNA transcription by recruiting H2.AZ and H3K4me2 to the rDNA (van de Nobelen *et al.* 2010). Thus, it may function as an insulator or a promoter depending on the conditions. Similar to the rest of the human genome, there are a number of different histone modifications and transcription factors that combine to have an integrated effect on the transcriptional state of the rDNA.

1.6. Human rDNA array flanking regions

In human, much less is known about the regions flanking the rDNA than is known about the rDNA, for which a number of characterizations have been performed and a complete sequence has been reported. The human rDNA flanking regions consist of a distal and a proximal flanking region on each acrocentric chromosome (Figure 1.1). The distal junction position was first identified by Worton *et al.* (1988) using restriction mapping of an X-derived translocation chromosome that has 3-5 rDNA units and the distal end of the short arm of chr 21 on the chr X. The restriction mapping results shows that the junction is approximately 3.7 kb upstream of the transcription initiation site. The remaining acrocentric chromosomes have the same restriction map for the distal region, suggesting that the

junction point and the distal flanking regions are conserved among these chromosome arms. Sylvester *et al.* (1989) obtained a ~4.5 kb sequence of a clone that has ~2 kb of the distal region and remaining rDNA sequence. The sequence shows that the position of the distal junction point is at position 39,029 in the IGS. This junction is 3,970 bp upstream of the transcription initiation site, similar to the distal-rDNA junction position found by Worton *et al* (1988). Later Gonzalez and Sylvester (1997) sequenced a clone that was previously identified to be from the distal flanking region to obtain ~8.3 kb of the distal region sequence. This sequence consists of a variety of sequence elements, including Alu elements, rRNA pseudogene fragments, ESTs and CpG rich regions (Figure 1.5). A 390 bp fragment from the distal flanking region sequence was amplifiable from all five acrocentric chromosomes, further demonstrating the conservation of the distal flanking region between all the acrocentric chromosomes (Gonzalez and Sylvester 2001).

Figure 1.5: Schematic diagram showing the distal rDNA junction position.

The distal junction position in the IGS of the rDNA unit next to the junction is represented as vertical line. The distal flanking region (green line) is consists of sequence element that includes rDNA pseudogenes (ψIGS and ψ28S), ETSs (blue boxes), Alu elements (orange boxes) and CpG rich region (brown box). The rDNA unit next to the junction and the rDNA unit next to it are shown. Figure not to scale.

The other side of the rDNA array i.e. the proximal flanking region has been characterized by Sakai *et al.* (1995). They identified a cosmid that has the proximal junction point by screening a chr 21 specific cosmid library using probes from the distal region and rDNA. The rDNA positive clones that are negative to the distal region probe were subjected to E*co*RV restriction digestion (rDNA does not have E*co*RV restriction site) followed by restriction mapping using E*co*RI and B*am*HI to identify a clone with the proximal junction region. The proximal junction position in the clone is at position 6,229 bp in ITS-1 of the rDNA unit (Figure 1.6). The identified proximal junction position in ITS-1 was shown to be present in chr 15, chr 21 and chr 22 by using mouse somatic cell hybrids that have only one human acrocentric chromosome. However, this junction was not amplifiable for chr 13, chr

14 and for a different hybrid cell for chr 21, suggesting that the proximal junction may be variable. The proximal region adjoining the rDNA array has a 68 bp unique sequence followed by tandem repeats of a ~147 bp satellite repeat known as ACRO1 (Figure 1.6). The ACRO1 satellite has been found on the short arm of all five acrocentric chromosomes together as well as in the pericentromeric region of chr 3 (Sakai *et al.* 1995).

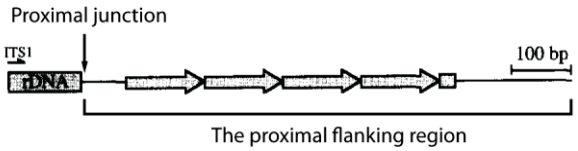


Figure 1.6: Schematic diagram showing the proximal rDNA junction position.

The position of the proximal junction point in the rDNA (grey box) is shown as a vertical arrow. The unique 64 bp region next to the junction is represented as thin black line next to the arrow and the ACRO 1 satellite array is represented as a tandem array of solid arrows (adapted from Sakai et al. 1995).

1.7. Role of rDNA in fundamental biological processes

The primary biological function of the rDNA is transcription of rRNA. rRNA contributes 80% of the total cellular RNA (Moss *et al.* 2007) and is a major building block of the ribosome. Since ribosomes are the molecular machinery responsible for protein synthesis, the rDNA is essential for cell survival. The rDNA also play role in cell proliferation and cellular stress response.

The rDNA and cell proliferation: The demand of protein increases during cellular proliferation, and this in turn increases the demand for ribosomes (Holland 2004; Dai and Lu 2008; Grzmil and Hemmings 2012). Increase in the utilization of ribosomes increases the demand for rRNA, which is met by an increased number active of rDNA copies (Drygin *et al.* 2010). Since the IGS contains transcriptional regulators, it is possible that these IGS elements may play roles in enhancing rRNA transcription during cancerous cell proliferation.

rDNA and the stress response: Protein synthesis and therefore the demand for ribosomes decreases in response to environment stresses such as nutritional starvation, temperature stress, and oxidative stress, which consequently decreases the number of active rDNA copies. The transcription factor JNK2, a common environmental stress response protein, inactivates the Pol I machinery, resulting in silencing of rDNA units during oxidative stress

(Mayer *et al.* 2005). Further, several noncoding RNAs are transcribed from the human IGS in response to acidosis, anaerobic conditions and heat shock. These are known to facilitate the localization of the proteins VHL and Hsp70 to the nucleolus. VHL helps to maintain energy equilibrium during acidosis by reducing rRNA transcription (Mekhail *et al.* 2006), while Hsp70 thought to play role in the prediction of nucleolar proteins during heat shock (Pelham 1984; Welch and Feramisco 1984; Welch and Suhan 1986). Both these proteins neutralize the deleterious effects related to rDNA transcription during stress conditions and are thought to prevent apoptosis (Pelham 1984; Welch and Feramisco 1984; Welch and Suhan 1986; Mekhail *et al.* 2006).

1.8. Aims of the study

There are two major aims for this PhD thesis.

1) The primary biological function of the rDNA is the transcription of rRNA for ribosome biogenesis. Besides ribosome biogenesis, the rDNA also plays role in other cellular processes. The IGS contains functional elements that play roles in the regulation of rDNA-dependent biological processes in addition to rRNA transcription. Although these elements are well characterized in yeast, comparatively less is known about them in human, and the positions of several previously identified functional elements in the human rDNA are not established. As the length of the IGS in human (~30 kb) is ~10 times greater than that in yeast (~3 kb), it is likely that the human IGS contains a number of functional elements. Because of highly repetitive nature of the human IGS, it is challenging to design wet lab experiments to characterize it. Hence, I decided to employ computational methods to search for potential functional elements in the IGS. Therefore the first aim of this study is:

Aim 1: To use a bioinformatics approach to identify and characterize the functional elements present in the intergenic spacer of the human rDNA.

2) The heterochromatic DNA is known to surround nucleoli in human cells. The nature of the rDNA flanking regions isn't known but because of their adjoin position to the rDNA array the general assumption is that these peri-nucleolar regions include the flanking regions, and therefore they must be heterochromatic. The assumption that the rDNA flanking regions are heterochromatic has led to the idea that they are non-functional. Therefore, these regions remain unplaced in the human genome assembly and are largely uncharacterized because of the focus of the human genome assembly on the euchromatic regions. However, a number of studies have suggested that the rDNA flanking regions play roles in nucleolar architecture

and/or function. Supporting this, the distal region forms discrete foci at the boundary of the nucleoli (Floutsakou *et al.* 2013). This indicates that the flanking region is not inert and non-functional but likely to have a biological role. However, it is difficult to establish any role of the rDNA flanking regions, as properties of the underlying sequence are still unknown. Therefore the second aim of this study is:

Aim 2: To use bioinformatics approaches to characterize the sequences of the rDNA flanking regions and to search for elements that may play roles in nucleolar organization and function.

Chapter 2

Identification of potential functional elements in the intergenic spacer of the human ribosomal DNA

[Blank page]

2.1. Introduction

In a human diploid genome, approximately 400-600 copies of ribosomal DNA (rDNA) repeat units are present (Henderson *et al.* 1972; Schmickel 1973; Stults *et al.* 2008). The rDNA units are distributed as tandem arrays on the short arm of the five acrocentric chromosomes (chromosomes with arms of unequal length) i.e. chr 13, chr 14, chr 15, chr 21 and chr 22 (Henderson *et al.* 1972). In a single array the number of rDNA units varies from 2 to >140 (Stults *et al.* 2008). Despite the high copy number, only a fraction of the rDNA units is transcriptionally active. The number of active and inactive rDNA units is not constant and varies between the tissues of an individual and between the same tissues of different individuals depending on the physiological conditions. A single rDNA unit is consists of a ~13 kb coding region followed by a ~30 kb intergenic spacer (Gonzalez and Sylvester 1995). The coding region encodes for the ribosomal RNA that is a key building block of ribosomes. In addition to ribosome biogenesis, the rDNA also regulates other cellular process that includes cell proliferation, cellular stress response, genomic stability and aging (Montanaro *et al.* 2008; Drygin *et al.* 2010; Ide *et al.* 2010; Audas *et al.* 2012; Saka *et al.* 2013). In the model organism, yeast functional elements that regulate these processes are known to be present in the IGS, which include a transcription promoter, a terminator, replication fork barrier, origin of replication, E-pro (Elion and Warner 1984; Brewer *et al.* 1992; Muller *et al.* 2000; Saka *et al.* 2013). However, in compare to yeast far less is known about their presence and position in the human IGS. Therefore, the focus of this study is to identify the potential functional element in the human IGS.

2.1.1. Strategy to characterized potential functional elements in the IGS

To identify the functional elements in the IGS it is difficult to use molecular techniques like deletion of the region of interest. The multiple copies and locations of the rDNA units means, it is not straightforward to make genetic changes in the regions of interest in the human IGS to determine their function. Furthermore, there is no certainty that the effect of any changes made in a single rDNA repeat unit will have a detectable effect. The protocols and molecular techniques required to alter the rDNA units at a global level and observe the consequent phenotypic changes are only well established in yeast, and are currently unavailable in human. Thus, it is challenging to characterise functional elements in the human rDNA using experimental approaches. Therefore, I have decided to take a comparative genomics approach, phylogenetic footprinting, to identify potential functional elements in the human IGS (Figure 2.1).

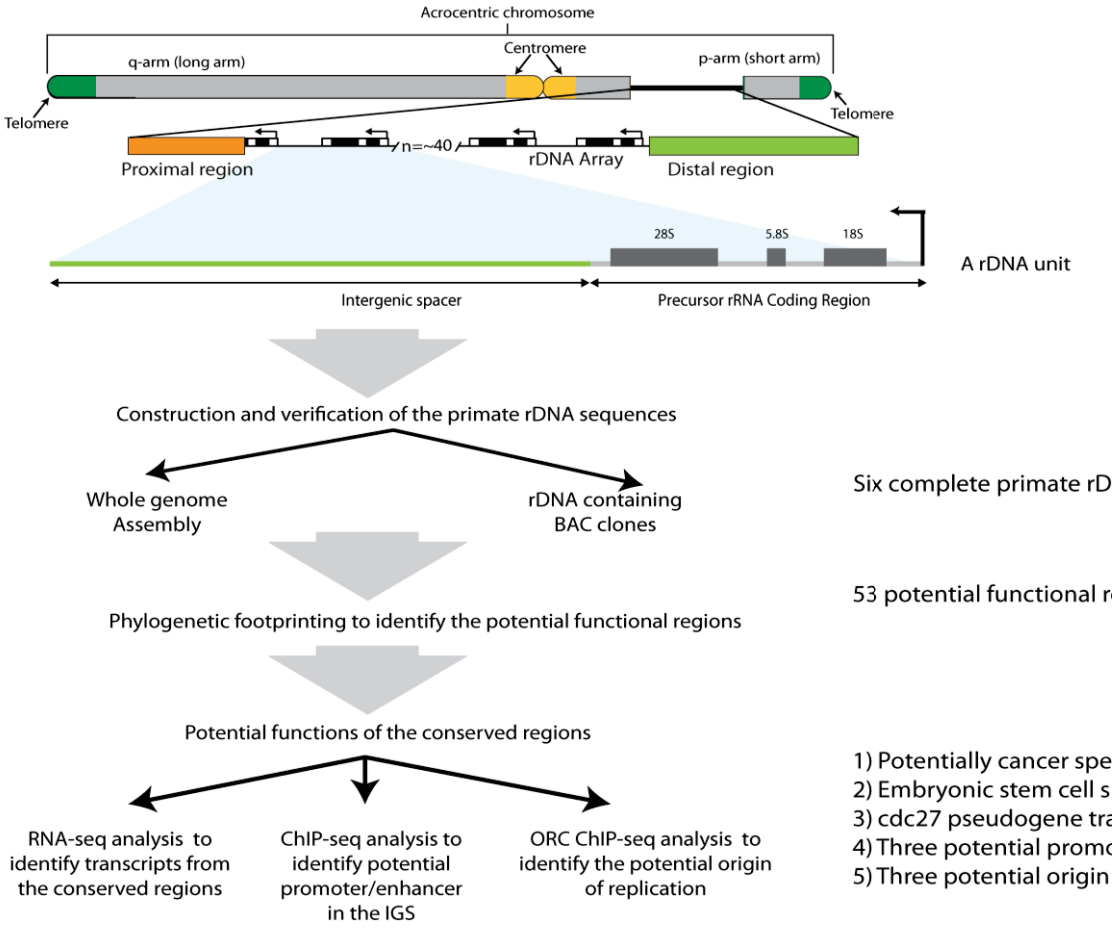


Figure 2.1: Schematic overview of the identification of potential functional elements in th

The flow diagram show the progression of the project and the different analysis performed regions. The major outcome of the study is shown on the right side of the figure.

The principle behind phylogenetic footprinting is that mutations in functional elements are deleterious, therefore changes in the sequences of functional elements are selected against and change at a slower rate than non-functional sequence in evolutionary time (Tagle *et al.* 1988). Comparison of the orthologous sequences from related species results in the functional elements appearing as "phylogenetic footprints" i.e. highly conserved regions in the multiple sequence alignment in a background of non-functional, poorly conserved sequences (Tagle *et al.* 1988). The success of the phylogenetic footprinting depends on the evolutionary relatedness of the species selected for comparison. Previous studies have shown that the inclusion of closely related species along with more distantly related species give the ability to identify conserved regions with high confidence (McCue *et al.* 2002; Stone *et al.* 2005).

Phylogenetic footprinting has been successfully applied to identify known and novel functional elements in the IGS of *Saccharomyces cerevisiae* (Ganley *et al.* 2005). The rDNA of five species that are related to the *Saccharomyces cerevisiae* were compared. Functional elements i.e. a bi-directional promoter E-pro, cohesion association region, replication fork barrier, and other potential gene independent functional elements (NOCs) correspond to the identified conserved regions in the IGS. Further, this technique has also been applied to other regions of the human genome to identify promoters and other gene regulators for variety of genes (Tagle *et al.* 1988; Bachman *et al.* 1996; Berezikov *et al.* 2005). To identify the potential functional elements in the human IGS, I searched for the phylogenetic footprints by comparing the human IGS with different primate IGS sequences. However, there were no primate rDNA IGS sequences available except for human. Therefore, to perform phylogenetic footprinting I first needed to obtain the rDNA sequences for the primate species I selected for this study. The whole genome shotgun sequencing (WGS) data of an organism contains the nucleotide information of its entire genome. However, current sequencing technologies can only determine the sequences of relatively small DNA fragments. Thus, to obtain the nucleotide information of the entire genome the genomic sample is sheared into small fragments and sequenced, and these sequences reads are merged together to construct the sequence of the genome. The process of merging the reads is known as *de novo* whole genome sequence assembly (WGA). To construct the rDNA sequence of the selected primates for the human rDNA phylogenetic analysis study, I used publically available whole genome sequencing data from different primate genome projects to perform WGA. Further, to verify that the obtained WGA rDNA sequences are not misassembled, I decided to identify and sequence BAC (bacterial artificial chromosome) clones containing the rDNA. A BAC is an engineered DNA molecule constructed by inserting a large DNA fragment (usually 100-200 kb) into a bacterial plasmid. Since BACs can contain large fragments of the

genome, they are used to construct the whole genome libraries. Children's Hospital Oakland Research Institute, USA (CHORI) provides BAC genome libraries for the primates in the form of BAC filters (BAC clones that have been gridded onto membranes). I decided to screen the filters from CHORI to identify the rDNA-containing BAC clones and then compare them to the rDNA sequences obtained by WGA.

The IGS are known to be transcribed to produce different regulatory long noncoding RNAs. Therefore, one potential function of the conserved regions identified using phylogenetic footprinting can be transcribing RNA transcripts that may have regulatory roles. To identify the full range of potential transcripts from the IGS, I decided to map RNA-seq data for different cell types. The data used for analysis was publically available from ENCODE project and sequencing was performed at Cold Spring Harbour Laboratory (CSHL). To identify potential IGS transcripts I selected long (>200 bp in length) polyadenylated denoted as poly(A)+ and long non-polyadenylated denoted as poly(A)- RNA-seq data for the analysis. Further, to identify potential micro RNA, small nucleolar RNA, tRNA and small nuclear RNA from the IGS, poly(A)+ RNA-seq data were selected to be mapped to the IGS. In the ENCODE project RNAs were fractionated according to their location in the cell i.e. cytosol and nucleus before sequencing (Djebali *et al.* 2012). The noncoding transcripts from the IGS are known to be located in the nucleolus (Audas *et al.* 2012; Jacob *et al.* 2012). Moreover, it has been recently shown that the larger fraction of the noncoding and intergenic RNAs are located in the nucleus compared to the cytosol (Djebali *et al.* 2012). Therefore, I decided to search for the transcripts from the IGS using RNA-seq data from the nucleus.

Previous studies have shown that several transcripts originate from the human IGS. This implies that transcriptional regulators (promoters, enhancers and insulators) of these IGS transcripts are present in the IGS and these may lie within the conserved regions. The histone modifications and transcription factors (TFs) associated with a genomic region determines and regulates its transcriptional activity (Caparros *et al.* 2009). Therefore, to identify potential transcription regulators in human IGS, I decided to map ChIP-seq data from ENCODE project for various histone modifications and TFs. The list of factors I mapped is given in Table 2.1 and Table 2.2.

Table 2.1: List of histone modifications mapped to the human rDNA sequences

Histone Modification	Description
H2A variant Z	Play role in several functions including Polycomb silencing, transcription activation and nucleosome assembly.
H3K9ac	Associated with regions that have open chromatin structure (less nucleosome occupancy)
H3K9me1	Associated with regions that have open chromatin structure (less nucleosome occupancy)
H3K27ac	Associated with transcriptional initiation and open chromatin structure.
H3K4me1	Enriched at enhancers and downstream of transcription start sites
H3K4me2	Enriched at enhancers and downstream of transcription start sites
H3K4me3	Enriched at promoters
H3K79me2	Marks the transcriptional transition region - the region between the initiation marks and the elongation marks
H4K20me1	Associated with active promoters and/or transcribing regions
H3K09me3	Promotes a repressive heterochromatic state
H3K27me3	Promotes a repressive heterochromatic state

Table 2.2: List of transcription factors mapped to the human rDNA sequence

Transcription factors	Description
Bdp1	cofactor of RNA Pol III
Brf1	cofactor of RNA Pol III
Brf2	cofactor of RNA Pol III
c-Myc	Associated with activation of rDNA (Pol I) and Pol II transcribed genes.
CTCF	Zinc finger protein enriched at insulators/promoters
Pol-II	POLR 2A subunit of RNA Pol II
Pol-III	POLR 3G subunit of RNA Pol III
TBP	Associated with all three polymerases (Pol I, Pol II and Pol III)
UBF	Involved in the recruiting Pol I to the rDNA
ZNF143	Transcription activator associated with Pol II and Pol III

The ENCODE project provides RNA-seq and ChIP-seq data for 147 different cell types out of which 18 were given higher priority by the ENCODE project based on the physiological conditions represented by them (http://www.genome.gov/26524238; Dunham *et al.* 2012). These 18 cell types were further divided into two tiers: tier-1 (includes three cell types) and tier-2 (includes 15 cell types) based on the level of priority in the ENCODE project. As described previously, the rDNA is thought to play a key role in several processes including cellular proliferation (Section 1.7). Cancerous cells are well-established examples of rapidly proliferating cells. Therefore, to identify the transcripts and transcriptional regulators associated with the rDNA that may play roles in cellular proliferation, I decided to select both noncancerous and cancerous cell types for comparative ChIP-seq and RNA-seq analysis. For this study, I have selected all cell types from tier-1 *viz.* non-cancerous cell types GM12878 and H1-hESC, and cancerous cell type K562. Since tier-1 contains only one cancerous cell type, I selected A549 and HeLa-S3 from tier-2 to include other cancerous cell types. Further, to have equal representation of the cancerous and noncancerous cell types in the analysis I have included one additional noncancerous cell type, HUVEC, from tier-2 for noncancerous cell types. The cell types A549, HeLa-S3 and HUVEC were selected over other 12 cell types in tier-2 because ChIP-seq data were present for most of the epigenetic factors included in the study. The details of selected cell types are given in Table 2.3

Table 2.3: The cell types included in this study

Cell type	Description
GM12878	Lymphoblastoid
HUVEC	Human umbilical vein endothelial cells
H1-hESC	Human embryonic stem cells line H1
K562	Leukaemia
HeLa-S3	Cervical carcinoma
A549	Adenocarcinomic alveolar basal epithelial cells

2.2. Material and methods

2.2.1. Bioinformatics Techniques

The bioinformatics methods used for this project are described below:

2.2.1.1. Comparative analysis of Sanger read assemblers to determine the efficiency of assembling rDNA repeat unit sequences

2.2.1.1.1. Extraction of potential rDNA reads from Sanger whole genome sequencing data:

Human whole genome sequencing data (12,146,378 reads) from the J. Craig Venter Institute (JCVI) sequencing center were obtained from the Ensemble database (currently available through the NCBI trace archive using the query: SPECIES_CODE = "HOMO SAPIENS"AND CENTER_NAME = "JCVI" AND STRATEGY = "WGA"). The reads obtained were grouped according to their insert size. Eleven different groups representing insert sizes of 2,250 bp, 9,500 bp, 11,000 bp, 12,000 bp, 12,500 bp, 13,000 bp, 40,000 bp, 42,000 bp, 43,000 bp, 44,000 bp and 45,000 bp were obtained. The paired end reads for each group were mapped to the human rDNA sequence (GenBank accession no. U13369) using gsMapper (ver. 2.3) (Roche/454)**.** The following parameters were used for the mapping: minimum 100 bp read overlap, 95% identity, and 200 expected depth. Expected depth represents the number of times a nucleotide position is expected to be sequenced in the data from a sequencing run. Reads mapped to the rDNA together with their mate pair were extracted from the original data to use as a test dataset for the assemblers.

2.2.1.1.2. Test assemblies

Five Sanger read assemblers *viz.* gsAssembler (ver. 2.3), Celera Assembler (ver. 6.1), MIRA (ver. 3.2.1), Phrap (ver. 1.090518) and Arachne (ver. r37405) were tested for the efficiency of rDNA repeat unit assembly. All assemblers were tested using default parameters. All assemblies were run on a 64-bit server with six-core Intel Xeon @ 2.67GHz processor and 512 GB RAM. The obtained assemblies were screened for human rDNA containing contigs using the GenBank human complete rDNA unit accession number U13369, as the query sequence.

2.2.1.2. Whole genome assemblies to obtain the primate rDNA sequences

2.2.1.2.1. Datasets

Whole genome sequencing data for the six other primates *viz.* chimpanzee (*Pan troglodytes)*, gorilla (*Gorilla gorilla)*, orangutan (*Pongo abelii),* gibbon (*Nomascus leucogenys)*, rhesus macaque (*Macaca mulatta)* and common marmoset (*Callithrix jacchus)* were obtained from the Ensemble database (the details of the data are described in Table 2.4).

Table 2.4: Details of WGS data for the primates

Organism	Sequencing Center	Sequencing coverage	Total number of reads	Number of reads used for assembly
Pan troglodytes	BI	3.6X	13,288,075	9,598,997
Gorilla gorilla	SC	2.1X	8,171,818	8,099,999
Pongo abelii	BCM	6X	11,882,059	11,863,254
Nomascus leucogenys	BCM	6X	12,747,839	8,399,955
Macaca mulatta	BCM	6X	7,512,808	7,512,808
Callithrix jacchus	BCM	6X	11,416,033	8,626,709

Currently data can be obtained from NCBI using the following queries:

a) SPECIES_CODE = "Pan troglodytes " AND STRATEGY = "WGS" AND CENTER_NAME = "BI"
b) SPECIES_CODE = "Gorilla gorilla " AND STRATEGY = "WGA" AND CENTER_NAME = "SC"
c) SPECIES_CODE = "Pongo abelii" AND STRATEGY = "WGA" AND CENTER_NAME = "BCM"
d) SPECIES_CODE = "Nomascus leucogenys" AND STRATEGY = "WGA" AND CENTER_NAME = "BCM"
e) SPECIES_CODE = "Macaca mulatta" AND STRATEGY = "WGA" AND CENTER_NAME = "BCM"
f) SPECIES_CODE = "Callithrix jacchus" AND STRATEGY = "WGA" AND CENTER_NAME = "BCM"

2.2.1.2.2. Whole Genome Assembly

Whole genome assemblies (WGA) for chimpanzee, gorilla, gibbon, macaque and callithrix were performed using Arachne ver. r37405 on a 64-bit server with an Intel quad core Xeon @ 3.2GHz processor and 75 GB RAM. Whole genome assembly for orangutan was performed using Arachne ver. r37578 on a 64-bit server with six-core an Intel Xeon @ 2.67GHz processor and 512 GB RAM**.** Default parameters were used for all the assemblies.

2.2.1.2.3. rDNA sequence construction

The following steps were repeated for each of the primate species to obtain the rDNA repeat unit sequence:

a) Step-1: Potential rDNA containing contigs were identified by screening the obtained WGA with human rDNA unit sequence extracted from BAC AL353644 (see Section 2.3.4 for the details of the extracted human rDNA sequence) using BLAST.

b) Step-2: Contigs with an average read coverage <10 were removed. The read coverage for rDNA containing contigs was obtained using CLC genomic workbench (CLC bio, Inc.). To remove the segmentally duplicated rDNA regions in the other part of the genome, contigs smaller than 1kb were also discarded. The low coverage ends (reads coverage <5) of the remaining contigs were trimmed before proceeding for further analysis.

c) Step-3: Depending on the number of rDNA units present in the contigs obtained from Step-2 two different strategies were employed to construct the primate rDNA repeat unit sequence. The strategies are described as below:

i. The first strategy was employed if step-2 yielded a contig longer than the human rDNA unit sequence (assuming that all the primate rDNA units are ~40 kb) and thus has more than one rDNA unit. This contig was selected for the construction of rDNA sequence. The partial unit in the contig was removed to obtain the complete rDNA repeat unit sequence. The presence of more than one rDNA unit in the contig was established by self-comparison using BLAST**.**

ii. The second strategy was employed if the contigs obtained from Step-2 were smaller than the human rDNA sequence and more than one overlapping contig was required to completely cover the human rDNA sequence. The overlapping contigs were merged using Consed (ver. 19) (Gordon *et al.* 1998). Consed first determines the overlapping regions between the contigs using cross_match (http://www.phrap.org/phredphrap/general.html), next it merges the reads in the overlapping regions and creates a new contig by generating a consensus sequence using the merged reads. The files containing contig information in .ACE format that were generated by Arachne were used as input for Consed.

d) Step-4: The contig obtained from Step-3 was rearranged such that the base 1 of the sequenced is the start of the 45S rRNA coding region, which is followed by the IGS. The 45S rRNA coding region in the primate rDNA sequence was demarcated by comparing it with the human 45S rRNA coding sequence using BLAST.

2.2.1.3. Primate rDNA BAC sequencing

2.2.1.3.1. NGS sequencing

Identified BACs (Section 2.3.5.2.1) were sequenced on the Illumina HiSeq 2000 platform to obtain 2x100 bp paired end reads with a 250 bp insert size. The sequencing was done by New Zealand Genomics Limited (NZGL), Otago University, Dunedin, New Zealand**.**

2.2.1.3.2. Read preparation

The quality of the paired end reads was checked using FastQC (Cox *et al.* 2010). Low quality ends of reads were trimmed using DynamicTrim (Cox *et al.* 2010) with a quality score cutoff of 13, and short reads were removed with a length cutoff of 25 bp using LengthSort. Both programs are part of the SolexaQA package (Cox *et al.* 2010). Contaminating reads from *E. coli* were removed by mapping reads to the *E. coli* genome (GenBank accession no. CP000948.1).

2.2.1.3.3. Assembly

Processed reads were assembled using Abyss (ver. 3.81) (Simpson *et al.* 2009). K-mer values 26-45 were used to obtain the optimum assembly for each dataset. The optimum assemblies were determined by comparing the contigs for each NGS assembly with the WGA rDNA sequence for the corresponding primate. The assembly with contigs that represent most of the corresponding WGA primate rDNA sequence was selected as the optimum assembly.

2.2.1.3.4. Mapping

Processed reads were mapped to the corresponding reference primate rDNA sequence using bowtie2 (ver. 2.0.4) (Langmead and Salzberg 2012) with parameters –N 1 –L 30. The consensus sequence from the mapped reads was generated using a minimum coverage cutoff of 5. The CLC Genomic workbench was used to obtain the consensus sequence. The mapped BAC rDNA sequence was aligned to the corresponding WGA rDNA sequence using MAFFT server (http://mafft.cbrc.jp/alignment/server; Katoh *et al.* 2009). Strategy E-INS-I and scoring matrix 1 PAM were used for the alignment.

2.2.1.4. rDNA Sequence analysis

The repeat regions in the rDNA sequences were identified using RepeatMasker (http://www.repeatmasker.org) with parameter “DNA source” set as “human”. Further, Alu elements in the IGS were also identified using DFAM database (ver. 1.1)

(http://dfam.janelia.org; Wheeler *et al.* 2013). Other sequence elements in the IGS were identified using the sequence aligner YASS (Yet Another Similarity Searcher) (Noe and Kucherov 2005) and BLAST (blastn) (Altschul *et al.* 1990).

The YASS was selected over the more commonly used BLAST tool for the pairwise sequence comparisons because of its specificity. YASS performs better than BLAST for the high repeat content sequences (Noe and Kucherov 2005). The reason behind the better performance is difference in two parameters that are common in their search algorithm i.e. 1) number of seeds and 2) difference in the scoring of transition/transversions in the seeds (Noe and Kucherov 2005). A seed is a small piece of the query sequence of predefined length that is used to search for significant match with other sequences to start the alignment. YASS utilizes groups of overlapping seeds while BLAST uses only one seed to report a hit. Searching for multiple seeds help YASS to span a larger region for the mach compared to the BLAST, which increase the probability to find a match for the query sequence. Further, YASS employs transition-constrained seeds to search the match. Transition-constrained seeds treat transition mutations (purine ↔ purine, pyrimidine ↔ pyrimidine) differently than transversions (purine ↔ pyrimidine) during the scoring seeds. Transitions are more common than transversions particularly in microsatellites (Ebersberger *et al.* 2002; Vowles and Amos 2004). The rDNA contains a number of microsatellites, therefore there are likely to be many transitions between the species. Therefore, during the pairwise comparison it is essential to treat both processes differently.

For the identification of the conserved region in the IGS, I used MAFFT rather than ClustalW to perform the multiple sequence alignments. ClustalW is widely used for multiple sequence alignment, but it has limitations when performing multiple sequence alignments of sequences that have multiple highly conserved sequences interrupted by less conserved regions, such as is found in the rDNA. ClustalW employs a progressive method for alignment where sequences are aligned depending on the phylogenetic tree generated based on their pairwise comparisons (Katoh *et al.* 2005). Multiple trees are generated and the alignment corresponding to highest scoring tree is reported as the optimal alignment. This alignment represents the maximum global sequence similarity among the sequences. In contrast to ClustalW, MAFFT uses an iterative refined method to generate the alignments where the sequences are first aligned using highest scoring phylogenetic tree (Katoh *et al.* 2005; Katoh *et al.* 2009). The generated alignment is then refined by comparing the subgroups in the tree to improve the alignments of a subset of sequences within the entire multiple sequence alignments. Further, the E-INS-i module in MAFFT employs generalized affine gap cost to score the gaps. The generalized affine gap cost facilitates large gaps in the alignment that is useful when searching blocks of high conserved regions, as less conserved

regions can be spanned by large gaps without affecting the total score of the alignment. In the case of the rDNA alignment of the primate sequences where pairwise sequence identity drops very rapidly, E-INS-i performs more accurately than ClustalW.

2.2.1.5. Multiple sequence alignment and Similarity plot

Primate rDNA sequences were aligned to the human rDNA sequence to generate multiple sequence alignment (MSA) with MAFFT (ver. 6.935b) (Katoh *et al.* 2005; Katoh *et al.* 2009) using strategy E-INS-i (--genafpair), 1 PAM scoring matrix (--kimura 1) and gap penalty zero (--ep 0) (**command:** mafft --genafpair --maxiterate 6 --thread 6 --cluastalout --kimura 1 --ep 0 --reorder fasta_input_file > seq.aln). Where required the obtained alignments were adjusted by visual inspection. To obtain the similarity plot first all the columns in the MSA that had gaps in the human rDNA reference sequence were removed. The modified MSA was used as input for SynPlot (http://hscl.cimr.cam.ac.uk/syn_plot.html; Gottgens *et al.* 2001) using a sliding window of 75 with increments of 1 bp to generate the similarity plot. The human rDNA annotations were mapped onto the similarity plot using GFF files as input for human rDNA sequence features. The conserved regions in the MSA were extracted using SynPeak (http://hscl.cimr.cam.ac.uk/syn_plot_peaks.htm; Gottgens *et al.* 2001). The following parameters were used to extract the conserved regions: window size 10, increment 1 and minimum identity 0.8. Conserved regions less than 10 bp apart were merged together.

2.2.1.6. ChIP-seq and RNA-seq analysis of the human rDNA sequence

2.2.1.6.1. Modified human genome assembly

Human rDNA sequence is currently absent from the human genome assembly (hg19). In human, the rDNA is present on the five acrocentric chromosomes *viz.* chr 13, 14, 15, 21 and 22. For the ChIP-seq and RNA-seq rDNA analyses I decided to introduce the rDNA sequence extracted from AL353644 (Section 2.3.4) and the proximal (towards centromere end of rDNA) and distal (towards telomere end of rDNA) flanking region sequences (details in Section 3.1.1) into the hg19 chr 21 sequence. A concatenated sequence with proximal-rDNA-distal sequence was prepared (Figure 2.2). Since the proximal contig is already present on chr 21 of the assembly, I first removed this proximal sequence from chr 21 and then inserted the concatenated sequence at the proximal contig position. The modified human genome assembly with the rDNA and flanking region sequences on chr 21 was used for further analysis.

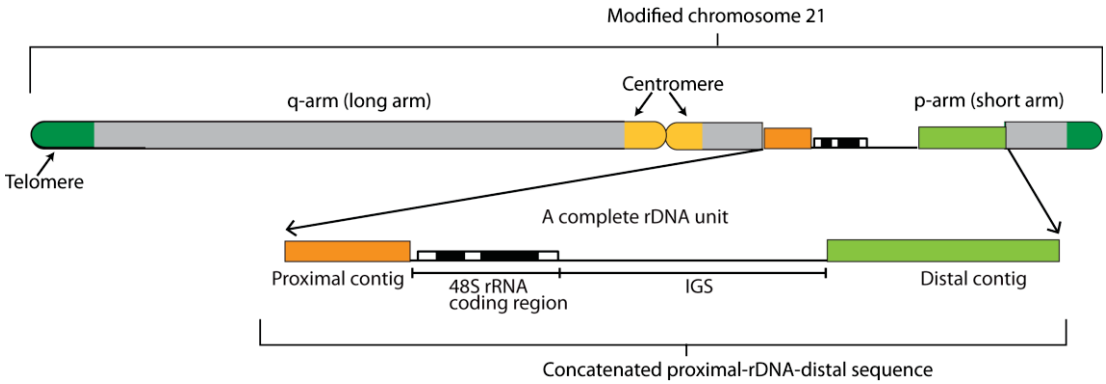


Figure 2.2: Schematic diagram of modified chromosome 21 reference sequence used for RNA-seq and ChIP-seq analysis of the human rDNA

A concatenated sequence consisting of the proximal contig (orange box), a single rDNA unit (white-black boxes) sequence and the distal contig (green box) was inserted to the short arm of the original chr 21 reference sequence.

2.2.1.6.2. Data set for ChIP-seq and RNA-seq analysis

Data for histone modifications, transcription factors (TBP, Pol II, Pol III, Brf3, Brf1, Brd1, ZNF143 and UBF), CTCF and an Input for cell types GM12878, K562, H-1hESC, HUVEC, HeLa-S3 and A549 were downloaded from ENCODE (https://genome.ucsc.edu/encode). Long poly(A)+, long poly(A)- and small RNA-seq data from the nucleus for GM12878, K562, H-1hesc, HUVEC, HeLa-S3, HepG2 and A549 were obtained from the CSHL long RNA-seq and CSHL short RNA-seq database (https://genome.ucsc.edu/encode**)**.

2.2.1.6.3. ChIP-seq analysis

The following steps were performed to analyze all the histone modification and TFs datasets for the six cell lines:

a) **Step-1:** Low quality ends of reads were trimmed using DynamicTrim with quality score cutoff of 13 and short reads were removed with a length cutoff of 25 bp using LengthSort. Both programs are part of the SolexaQA package (Cox *et al.* 2010).

b) **Step-2:** Processed reads were mapped to the modified human genome assembly using bowtie (ver. 0.12.8) (Langmead *et al.* 2009). The following parameters were used to map the reads: uniquely mapped reads to the genome (-m 1), maximum three mismatches in the seed (-n 3) and seed length (-l 30). The alignment was obtained in a SAM file format.

command: bowtie -l 30 -n 3 -p 7 --chunkmbs 1024 -a --best --strata -m 1 modified_human_genome_index ChIP_seq_trimmed.fastq -S ChIP_seq_mapped.sam

c) **Step-3:** Mapped reads were sorted according to the position mapped to the reference sequence using command SortSam.jar. The ChIP-seq data in this study were generated on the Illumina platform, which has a limitation in resolving GC rich regions. Therefore, contrary to the idealistic situation where each base has same coverage, the GC rich regions have lower sequence coverage than other regions. This variation in the sequence coverage among the regions is known as coverage bias. To remove the coverage bias and provide equal representation of all the regions in the analysis, multiple reads mapped to same location were removed using command MarkDuplicates.jar. All the replicates for each sample were merged using command MergeSamFiles.jar. All the three commands used in Step-3 are part of Picard tools (ver. 1.6.1).

   Command: java –jar SortSam.jar I=mapped_ChIPseq_chr21.sam O=mapped_ChIPseq _chr21_sorted.sam SO=coordinate MAX_RECORDS_IN_RAM=5000000

   Command: java -jar MarkDuplicates.jar INPUT=mapped_ChIPseq_chr21_sorted.sam O= mapped_ChIPseq _chr21_dr.sam M=stat_dr.txt REMOVE_DUPLICATES=true

   Command: java -jar MergeSamFiles.jar I=mapped_ChIPseq _chr21_dr1.sam I=mapped_ChIPseq_chr21_dr2.sam O=mapped_ChIPseq_chr21_dr_merge.sam SORT_ORDER=coordinate.

d) **Step-4:** The fragment size was calculated using the merged SAM file using run_spp.R ver. 1.11 (http://code.google.com/p/phantompeakqualtools/; Kharchenko *et al.* 2008).

   Command: Rscript run_spp.R -c=ChIP_seq_mapped_dr_merge.bam -savp -out=ChIP_seq_mapped_dr_merge.txt

e) **Step-5:** The merged and sorted SAM files were used to call the peaks using callpeak function. The noise was removed from the signal by subtracting the corresponding Input signal using bdgcmp function. Both call peak and bdgcmp are functions of MACS2 (ver. 2.0.10.20120913) (https://github.com/taoliu/MACS/; Zhang *et al.* 2008).

   Command: macs2 callpeak -c ChIP_seq_control_mapped_dr_merge.sam -t ChIP_seq_mapped_dr_merge.sam -g 'hs' --keep-dup all -n ChIP_seq --trackline -B -m 5 50 --nomodel --shiftsize <fragment_length/2>

   Command: macs2 bdgcmp -t ChIP_seq_treat_pileup.bdg -c ChIP_seq_control_lambda.bdg -o ChIP_seq_treat_minus_linear.bdg -m subtract

f) **Step-6:** Peaks corresponding to the rDNA sequence were extracted for further analysis.

g) **Step-7:** The peaks were visualized using Integrative Genomics Viewer (IGV) ver. 2.3 (Robinson *et al.* 2011; Thorvaldsdottir *et al.* 2013).

2.2.1.6.4. RNA-seq assembly:

The following steps were performed for all the selected long poly(A)+ and long poly(A)- RNA-seq datasets:

a) **Step-1:** Low quality ends of reads were trimmed using DynamicTrim with a quality score cutoff of 13 and short reads with a length cutoff of 25 bp were removed using LengthSort.

b) **Step-2:** Processed reads were mapped to the modified human genome assembly using STAR aligner (ver. 2.2.0) (Dobin *et al.* 2013). The following parameters were used to map the reads: mapped with the score ≤5 the maximum score (--outFilterMultimapScoreRange 5), mapped ≤ 15 times on the genome (--outFilterMultimapNmax 15), maximum 10 mismatches (--outFilterMismatchNmax 10) and with 10000 bp maximum distance between the paired read (--alignMatesGapMax 10000)**.** The alignments were obtained in a SAM file format.

   Command: STAR --genomeDir genome_STAR_index --readFilesIn rnaseq_1.fastq rnaseq_2.fastq --outSAMstrandField intronMotif --runThreadN 10 --outFilterMultimapScoreRange 5 --outFilterMultimapNmax 15 --outFilterMismatchNmax 10 --outFilterIntronMotifs RemoveNoncanonical --alignMatesGapMax 10000

c) **Step-3:** Since the modified chr 21 contains rDNA sequence, alignments for the reads mapped to modified chr 21 were extracted from the alignment file obtained from Step-2

d) **Step-4:** All the data were sorted according to the reference sequence coordinates using the command SortSamFiles.jar from Picard tools (ver. 1.6.1).

   Command: java –jar SortSam.jar I=mapped_rnaseq_chr21.sam O=mapped_rnaseq_chr21_sorted.sam SO=coordinate MAX_RECORDS_IN_RAM=5000000

e) **Step-5:** The sorted SAM files were used as input for Cufflinks (ver 2.0.2) (Roberts *et al.* 2011) to assemble the RNA-seq data. Default parameters were used for the assembly except the minimum isoform abundance required to be assembled was changed from 1% to 0.5% (-F 0.05).

   Command: cufflinks -N --total-hits-norm -p 8 --no-update-check -F 0.05 <mapped_rnaseq_chr21_sorted.sam>

2.2.2. Molecular Techniques

The molecular biology methods used for this project are described below:

2.2.2.1. BAC filters and BAC clones:

BAC filters for the five primates (orangutan, gibbon, gorilla, macaque and marmoset) were obtained from Children's Hospital Oakland Research Institute, USA (CHORI; http://www.chori.org).

2.2.2.2. Probe preparation:

The 18S rDNA region 4,328-4,922 (coordinate based on GenBank entry U13369) of the human rDNA was used as a probe. This region was selected for use because it is highly conserved among the primates. Two different methods were used to prepare the probes:

2.2.2.2.1. Probe for screening BAC filters

For screening BAC filters, the rDNA probe was prepared using DIG high prime DNA Labeling Kit II (Roche; Catalogue no. 11585614910). Standard PCR reactions were performed using male human genomic DNA as template (Promega; Catalogue no. G1471) and primers HS_18S_rDNA_F (5'-AGCTCGTAGTTGGATCTTGG-3') and HS_18S_rDNA_R (5'- GTGAGGTTTCCCGTGTTGAG -3') to amplify the 18S rDNA region. The obtained PCR product was denatured and mixed with random hexanucleotides, dNTP, digoxigenin-dUTP (alkali-labile), and Klenow enzyme, and incubated overnight at 37^{o}C. The reaction was stopped by heating the mixture to 65^{o}C for 10 mins.

2.2.2.2.2. Probe for identifying rDNA units in I-P*po*I digested Southern blots

For I-*Ppo*I digested Southern blots, the rDNA probe was prepared using PCR DIG labeling mix (Roche; Catalogue No. 11585550910). Standard PCR reactions were performed with dNTP replaced by PCR DIG labeling mix (dATP, dCTP, dGTP, dTTP and digoxigenin-11-dUTP). Male human genomic DNA used as template (Promega; Catalogue no. G1471) and primers HS_18S_rDNA_F and HS_18S_rDNA_R were used.

2.2.2.3. Southern Hybridization:

Blots were incubated in 1X PSE buffer (hybridization buffer) at 65^{o}C for 30 min. Denatured probe was added to the hybridization buffer and the membrane was hybridized overnight at 65^{o}C. The next day, blots were washed 3 times with 0.1X PSE for 5 mins, twice with 1X TBS for 5 mins followed by incubation in blocking solution for 1 hr. All these steps were

performed at 65°C. Blocked blots were incubated in antibody solution for 30 mins at 25°C. Blots were washed in blocking solution for 10 mins, followed by 1X TBS for 10mins and then equilibrated in alkaline phosphatase buffer for 15 mins. All these steps were performed at 25°C. Blots were placed in a development folder and the chemiluminescent substrate CDP-Star (Roche; Catalogue No. 11685627001) was applied to the membrane and incubated for 5 min at 37°C. Signals were detected using Luminescent Image Analyzer System (Fujifilm; LAS4000).

1X PSE (Hybridization buffer)

0.3 M anhydrous sodium phosphate (4.26 g/100 mls)

7% SDS (7 g/100 mls)

1 mM EDTA (200 µl of 0.5M EDTA)

Milli-Q water (make up to 100 mls)

10X TBS

1 M Tris (24.23 g/L)

1.5 M NaCl (80.06 g/L)

Milli-Q water (make up to 1000 mls)

Adjust pH to 7.6 using HCl

Blocking solution

1.5% skimmed milk powder (15g/L) was dissolved in 1X TBS (up to 1L) by stirring at 60°C. The solution was filtered using filter paper (Whatman; Catalogue no. 1001 090).

Alkaline phosphatase buffer

100 mM Tris (200 mls of 0.5M Tris-HCl)

100 mM NaCl (5.8 g/L)

5 mM $MgCl_2$ (1 g/L)

Milli-Q water (make up to 1000 mls)

pH adjusted to 9.5 using NaOH

2.2.2.4. Verification of the presence of the rDNA unit in the E. *coli* containing BACs:

The *E. coli* containing BACs were obtained from the CHORI as stab cultures and streaked on LB agar plates containing chloramphenicol. The isolates were tested for the presence of

rDNA units by colony PCR using primers 18S_F and 18S_R. The isolates with the expected product were used further for BAC extraction.

2.2.2.5. BAC extraction:

Overnight *E. coli* cultures containing the BAC of interest were grown by inoculating 1000mls of LB media containing chloramphenicol (30 µg/L) and incubated at 37°C and 300 rpm for 16 hrs. NucleoBond® Xtra Maxi Plus (Macherey-Nagel, Germany; Catalog No. 740414.10) plasmid purification kit was used to extract the BACs from the overnight culture. Genomic DNA contamination in the purified BACs was determined by digesting with *Eco*RI (Roche; Catalogue no. 11175084001) at 37°C for 1 hr and running the digested product on an agarose gel.

LB media
171 mM NaCl (10 g/L)
Bacto Tryptone or Bacto Peptone (10 g/L)
Bacto yeast extract (5 g/L)
Milli-Q water (make up to 1000 mls)

Chloramphenicol (Stock solution 30 mg/ml)
0.03 g Chloramphenicol
2.70 mls ethanol
Mixed and filtered using microfilter (Biofil syringe filter; Catalogue No. FPE-204-030)

2.2.2.6. I-P*po*I Digestion:

10 µl of purified BAC DNA was digested by incubating with 100U of I-P*po*I (Promega, USA; Catalogue No. R7031) overnight at 37°C in a 20 µl reaction as per the manufacturer's instructions.

2.2.2.7. Field inversion gel electrophoresis

To determine the size of the rDNA in the BACs, I-P*po*I digested products (1 µl of Sample + 1 µl of dye) were loaded on to 1% pulsed field certified agarose (Bio-Rad; Catalogue No. 162-0137) in 0.5X TBE gels. A 5 kb ladder (Bio-Rad; Catalogue No. 170-3624) was run next to the samples to determine fragment sizes. To aid resolution of multiple bands, the ladder was mixed with dye and water in the ratio of 1:1:2 and incubated for 2 hrs at 37°C. The incubated mixture was heated at 50°C for 15 min and placed in ice for 10 min. Gels

were electrophoresed in 0.5X TBE at 14^{o}C in a CHEF Mapper® XA pulsed field gel electrophoresis system (Bio-Rad; Catalogue no. 170-3670 to 170-3673) for 31 hrs using the following FIGE settings: 180 V forward voltage (5.4 V/cm), 120 V reverse voltage (3.6 V/cm), switch time 0.4 sec to 2 sec, linear ramp. Gels were stained in 20 ng/ml ethidium bromide solution for 1 hr and destained in water with gentle shaking for 2hrs. The bands were visualized under UV light using gel imager (Bio-Rad).

0.5X TBE

45 mM Tris (5.4 g/L)

45 mM Boric acid (2.75 g/L)

1mM EDTA (pH8) (2 mls of 500 mM EDTA)

Milli-Q water (make up to 1000 mls)

2.2.2.8. Southern blotting

The FIGE gel was transferred to Biobond Plum^{TM} SS nylon membranes (Sigma Aldrich; Catalogue no. N4781-1EA) using a standard Southern blot protocol (Sambrook and Russell 2001). Before the transfer, gels were incubated in 0.2M HCl for 20 mins, denaturation solution for 30 mins, and neutralization solution for 1hr. Following the transfer, the membranes were treated with 2X SSC for 5 min before the DNA was cross-linked to the membrane using a UV crosslinker (UVP, Model no. CX2000). Finally, the membrane was air-dried.

0.2M HCl

6ml of conc. HCl in 300mls H_2O

Denaturation Solution

0.5M NaOH (20 g/L)

1.5M NaCl (87.6 g/L)

Neutralization solution

2M NaCl (116.8 g/L)

0.5M Tris-HCl (60.5 g/L)

pH => 7.4 (95 mls conc. HCl)

20XSSC

3M NaCl (175.2 g/L)

0.3M Na-citrate (88.1 g/L)

2.3. Results

2.3.1. Whole genome assembly strategy to obtain the primate rDNA sequences

To identify the potential functional elements in the human IGS using the phylogenetic footprinting primate rDNA sequences are required to be compared with the human rDNA sequence. However, at present, the only available complete primate rDNA sequence is that of human. Despite the fact that the genomes of several primate species have been sequenced no complete rDNA sequences are present in GenBank for these primates. Therefore, the first step in order to perform phylogenetic footprinting was to acquire the rDNA sequences of the primate species. Since the whole genome sequencing (WGS) data of several primate species is publically available I decided to employed whole genome assembly (WGA) to construct the primate rDNA sequences. The WGS data of an organism contains the sequence information of its entire genome. Primates have 200-500 copies of the rDNA unit and therefore, compared to the unique regions such as typical protein-coding genes, the rDNA is expected to be represented 200-500 times more in WGS data. The rDNA units are almost identical to each other hence it is difficult to differentiate between the reads from one unit to another. Therefore, genome assemblers pile up all the reads from the different rDNA units together to generate a single rDNA sequence. Hence, the rDNA sequences that are produced by whole genome assemblies represent the consensus sequence of all the rDNA repeat units for that species.

The genome of the primates has been sequenced on either Sanger or NGS platforms. To perform the WGA for constructing the primate rDNA sequences I have selected WGS data from the Sanger platform because of the longer reads and larger insert size compare to the NGS data. It is known from the sequences of the complete rDNA units of human and mouse that the IGS of mammalian rDNA consists of a variety of different repeat elements, including Alu repeats and blocks (up to ~1.7 kb) of microsatellites (Gonzalez and Sylvester 1995; Grozdanov *et al.* 2003). Microsatellites are made of small identical sequences (tandem array of 2-6 bp monomers) that are tandemly repeated several times, so it is not possible to differentiate between the reads from different parts of a microsatellite region. The short reads with small insert sizes from microsatellite regions will pile up together during assembly to form a small cluster (Figure 2.3.a). To correctly assemble the long regions of microsatellites, long reads with large insert size are required (Figure 2.3.b). Further, Alu repeats present in the IGS are also present in the other parts of the genome (non-rDNA Alu repeats). Therefore, to avoid reads from non-rDNA Alus during the rDNA assembly, long reads with large insert

size are required so that the rDNA and non-rDNA regions surrounding the Alu elements can be differentiated. Additionally, the rDNA is segmentally duplicated in other parts of the genome. The sequence identity between the duplicated rDNA regions and the rDNA unit is lower than the sequence identity between different copies of the rDNA units. Therefore, to avoid reads from these duplicated regions, long reads are also required to increase the confidence of the reads alignment. The Sanger sequencing platform fulfils the requirement of long reads with large insert size and therefore, only primate species that have WGS data from the Sanger platform were considered during species selection for phylogenetic footprinting (Section 2.3.2).

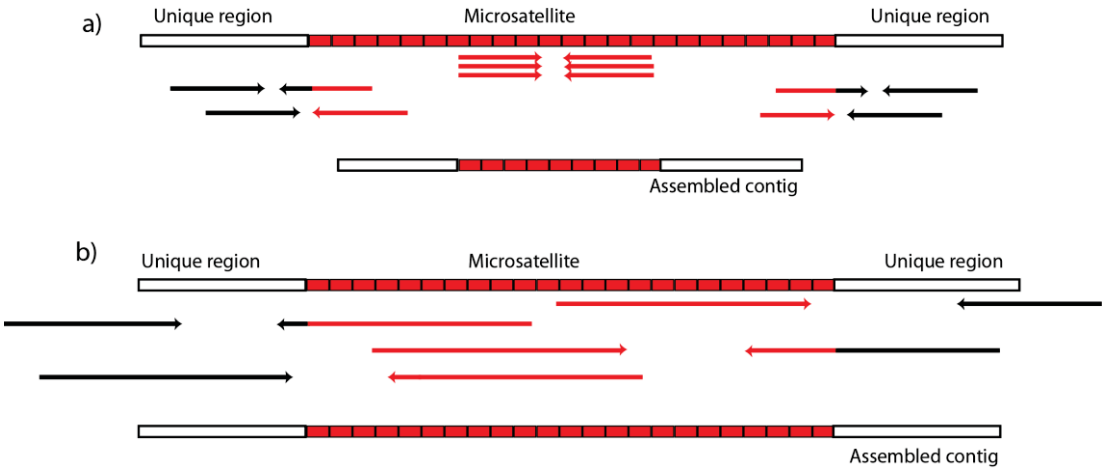


Figure 2.3: Microsatellite assembly using different size paired-end reads.

a) Microsatellite assembly using short reads with small insert size. b) Microsatellite assembly using long reads with large insert size. The microsatellite monomer is shown as a red box. The red arrows represent the reads from microsatellite, the black arrows represent the reads from a unique region and red-black arrows represent the reads that cross the microsatellite and a unique region. The forward arrows are reads from the 5' end, and the reverse arrows are reads from the 3' end. Paired reads are shown on the same horizontal level.

2.3.2. Selection of the primate species for the phylogenetic footprinting of human rDNA

The next step to start the phylogenetic footprinting of human rDNA was the selection of primate species for the sequence comparison. There are 261-377 living species of primates that are distributed among 13 families (Perelman *et al.* 2011). However, eight primate species, including both apes, monkeys and humans, have had their genomes sequenced on the Sanger platform (Marques-Bonet *et al.* 2009). From these eight species, I selected *Pan troglodytes*, *Gorilla gorilla*, and *Pongo abelii* from the Hominidae, *Nomascus leucogenys* from the Hylobatidae, *Macaca mulatta* from the old world monkeys and *Callithrix jacchus* from the new world monkeys for comparative analysis of the IGS in primates. Since *Papio cynocephalus* and *Macaca mulatta* are closely related, only *Macaca* was included in the study as a representative of the old-world monkeys. These primates all have had their genome completely sequenced on the Sanger platform, which is essential for the strategy to construct the rDNA sequences. Further, the range of relatedness of the species plays the most critical role in the phylogenetic footprinting. The selected primate species represent a group of closely related species (species from Hominidae and Hylobatidae) together with more distantly related species (species from the old world and new world monkeys) (Figure 2.4). Since the selected primates are used as the representative members of their genus, from this point onwards they are referred to by their common name, i.e. *Pan troglodytes* as chimpanzee, *Gorilla gorilla* as gorilla, *Pongo abelii* as orangutan, *Nomascus leucogenys* as gibbon, *Macaca mulatta* as macaque and *Callithrix jacchus* as common marmoset.

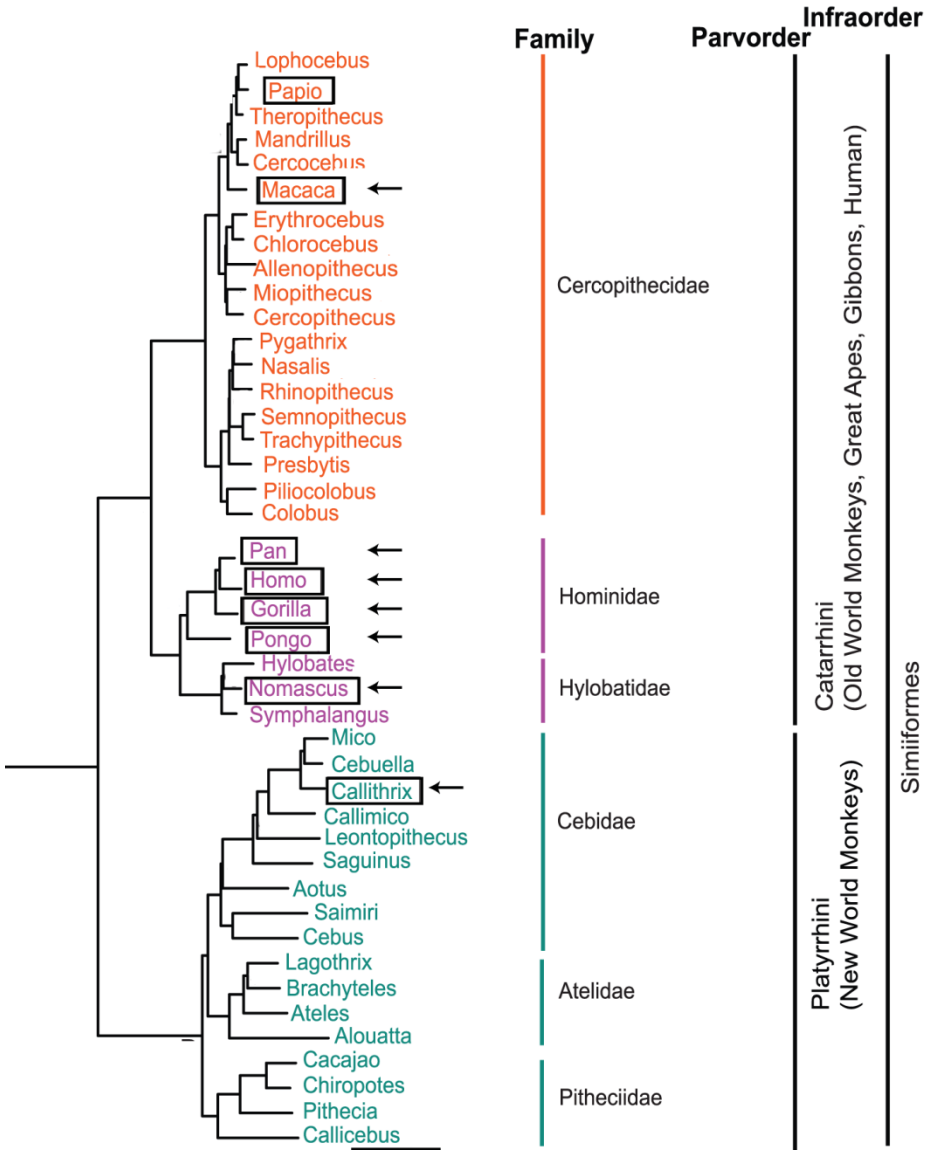


Figure 2.4: Primate phylogenetic tree showing the genera selected for human rDNA phylogenetic footprinting.

The phylogenetic tree shows the relationships of the known ape genera (purple text), old world monkeys (orange text) and new world monkeys (blue text), as well as their taxonomic classification (to the right). The genera of species for which whole genome Sanger sequence data are available are boxed. Arrows indicate the genera selected for human rDNA phylogenetic footprinting. (The figure is adapted from Perelman et al. 2011).

2.3.3. Comparison of sequence assemblers to determine the ability to assemble the rDNA sequence

Reads from repeat regions have identical sequences, hence it is not possible to differentiate between them, and therefore, to reduce assembly complexity, assemblers ignore reads from the repeat regions. Most Sanger read assemblers are based on the overlap-layout-consensus (OLC) algorithm for read assembly. OLC based assemblers mark the reads as repeats based on read coverage. To start the assembly, overlaps are determined among the reads. The reads from repeat regions pile up to form high coverage contigs because of their sequence identity. Contigs that have read coverage higher than a certain cutoff value are marked as repeat contigs and are discarded before proceeding further. This is also true for the rDNA in primates: multiple rDNA units are present in the genome and have almost identical sequence; the assembler cannot differentiate between the reads from different units and piles them together to form high coverage contigs. As a result of high coverage, the rDNA contigs are treated by the assemblers as repeat contigs and are discarded in the initial phase of assembly. Hence, although the scheme to obtain the rDNA unit sequence using WGA seems straightforward, it is challenging to implement. To find an efficient assembler that can assemble the rDNA unit accurately, I performed a comparative analysis of the Sanger read assemblers.

2.3.3.1. Dataset to assess the efficiency of the Sanger assemblers

The goal of performing WGA for the selected primates was to obtain the rDNA sequence. Therefore, to search for an efficient assembler that can assemble the rDNA sequence accurately, I prepared a dataset that has a high number of human rDNA reads. I used human WGS data from the JCVI sequencing center for the comparative analysis of sequence assemblers. Reads were grouped according to their insert size into eleven groups (Section 2.2.1.1). These groups were mapped to the GenBank human rDNA sequence (Acc. No. U13369). The human rDNA consensus sequence obtained from lib_12500 (group with insert size of 12,500 bp) mapped reads gave a complete rDNA sequence while the consensus rDNA sequence for the mapped reads from the other groups cover ~80-95% of the rDNA sequence. Since lib_12500 has reads for the entire rDNA, it was used to evaluate the repeat assembly efficiency of different assemblers.

2.3.3.2. *De novo* assembly comparison of sequence assemblers using lib_12500 dataset:

All the reads from lib_12500 were used as input to perform *de novo* assemblies. Default parameters were used for the selected assemblers. The rDNA containing contigs were identified by screening the obtained assemblies with the GenBank human rDNA sequence (Acc. No. U13369) as the query sequence, using BLAST. The contigs that have >95% identity with U13369 and >2 kb were marked as rDNA containing contigs. The list of the assemblers and the outcome of the assemblies are summarized in Table 2.5**.** After testing different assemblers, Arachne was selected for the assembly of primate rDNA sequences because it was able to assemble the entire human rDNA IGS and its computational memory requirement was within the limits of the available computing facilities.

Table 2.5: Comparative analysis of assemblers to evaluate their efficency to assemble the h

Assembler name	Assembler details	Reason for selection	Result
Newbler (gsAssembler)	Roche	Part of Roche GS data analysis software package.	The largest rDNA containing contig in length. Several regions of the rD were missing from the assembly.
Celera assembler	Celera genomics	Has been used in major genome projects e.g. for *Drosophila* (the first multicellular organism whole genome assembly) (Myers *et al.* 2000) and for a diploid human genome assembly (Levy *et al.* 2007).	rDNA contigs 1 kb to 40.12 kb in le obtained. These contigs do not repr entire rDNA sequence. Further, lon have large poly-Ns tracks. Therefor of contigs after removing poly-Ns i considerably less.
Phrap	University of Washington Genome	Successfully assembled repeated regions for several organisms including nested repeats in Maize, for which 67% of the genome is made of transposable elements http://www.phrap.org/phredphrap/phrap.html)	The memory requirement for *de no* assembly exceeded the limit of the computing server (512 GB RAM).
MIRA (Mimicking Intelligent Read Assembly)	Deutsches Krebsforsch-ungszentrum Heidelberg	The literature for MIRA claims that it can resolve repeat regions (http://sourceforge.net/projects/mira-assembler).	rDNA contigs of length 5 kb-33.17 97-99% identity with U13369 sequ obtained. The contigs together repr entire rDNA sequence. The memor requirement was enormous: ~150 G memory was used for lib_12500 rea (3,225,868 reads).
Arachne	The Broad Institute	The assembler was used to assemble the mouse genome (Batzoglou *et al.* 2002; Jaffe *et al.* 2003) and efficiently resolve repeat regions in mouse genome.	rDNA contigs 24.57 kb and 30.24 k length were obtained that represent IGS but only a fraction of the rDNA region.

2.3.4. Reference human rDNA unit sequence

The GenBank entry for the only reference human rDNA unit (U13369) that is currently available was constructed by assembling partial sequences obtained from different experiments (Gonzalez and Sylvester 1995) and therefore, it is possible that this sequence does not accurately represent the true rDNA sequence. To establish an accurate reference rDNA sequence I extracted a complete human rDNA unit from an unannotated GenBank BAC entry AL353644 (described in detail in chapter 3). The extracted complete rDNA unit is 43,972 bp in length and includes a 13,357 bp 45S rRNA coding sequences (referred as coding sequence from here on) and a 30,615 bp IGS sequence. I refer to this extracted BAC human rDNA sequence as the "human rDNA", and this was used as the human rDNA reference for all comparisons with the primate rDNA sequences. The coding sequences of the human rDNA and U13369 are 98.2% identical with nucleotide variants distributed throughout the sequence. The IGS sequences are 88.3% identical. The reason for the lower identity in the IGS is copy number variation in R-repeats and microsatellite regions. There are only two R-repeat blocks present in the human rDNA unit compared to three in U13369 (the R1-repeat block at coordinates 13,481-14,279 in U13369 is absent in the human rDNA). The 2 kb $[TCTC]_n$ microsatellite at 21,894-23,859 and the 50 bp $[TCT]_n$ microsatellite at 40,625-40,677 in the human rDNA are absent in the U13369 IGS. Similarly, a 200 bp $[TCTC]_n$ microsatellite at 26,341-26,542 and a 165 bp $[CTTG]_n$ microsatellite at 40,106-40,271 are present in U13369 IGS but are absent in the human rDNA IGS. Excluding these repeat copy number variations, the identity between the IGS of the human rDNA and U13369 is 98.1%.

2.3.5. Construction and verification of primate rDNA unit sequences

2.3.5.1. Construction of primate rDNA sequences using whole genome assembly strategy

The next step after selecting the appropriate assembler was to obtain the rDNA sequences of the selected species by using the *de novo* WGA strategy. The details of the whole genome assemblies and construction of the rDNA sequence for each primate species using WGA are as follows:

2.3.5.1.1. Chimpanzee reference rDNA unit sequence

A WGA for chimpanzee was performed using publicly available WGS data from a male individual from West Africa named "Clint" (NCBI Project Accession: PRJNA13184;

Mikkelsen *et al.* 2005) using Arachne (Sections 2.2.1.2 and 2.2.1.2.2; assembly statistics is described in Appendix Table 1). Seven contigs containing 6 kb to 58 kb of chimpanzee rDNA sequence were identified by screening this WGA with the human rDNA sequence using BLAST (Table 2.6). The rDNA contigs were named Pt_1 to Pt_7 in order of decreasing length (where Pt stands for *Pan troglodytes*). I derived the chimpanzee rDNA sequence from contig Pt_1 because it has one complete and one partial rDNA unit. The partial sequence was removed to give a complete reference chimpanzee rDNA sequence of length 41,773 bp (Figure 2.5). This WGA chimpanzee rDNA unit consists of a 13,279 bp coding sequence and a 28,494 bp IGS sequence. The obtained chimpanzee rDNA sequence is 93.1% to 98.3% identical to the six other chimpanzee rDNA containing contigs (Figure 2.5). The variation between the contigs is mainly because of copy number variation in microsatellites.

Table 2.6: Statistics of the potential chimpanzee rDNA contigs.

Contig name	Length	Number of reads	Average Coverage
Pt _1	46,151	7,686	84.5
Pt _2	30,263	3,616	79.9
Pt _3	21,021	4,646	136.7
Pt _4	16,569	2,153	83.6
Pt _5	11,156	548	32.8
Pt _6	6,540	252	26.7
Pt _7	6,501	229	23.0

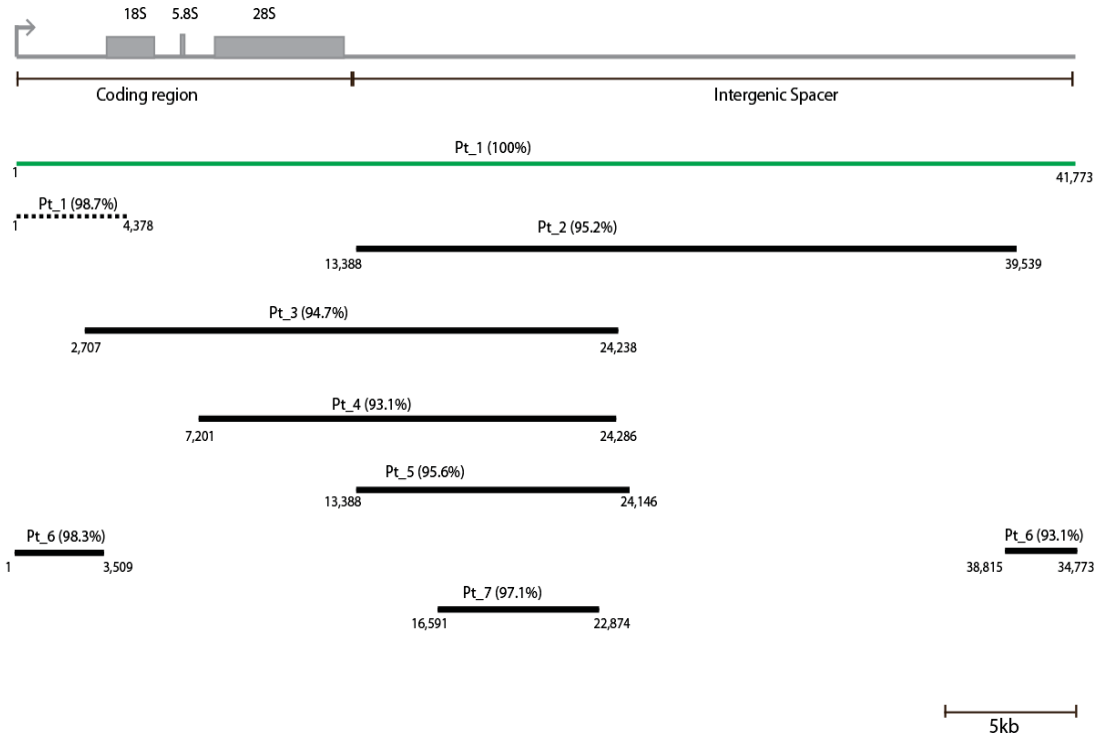


Figure 2.5: WGA contigs containing chimpanzee rDNA sequence.

The rDNA containing contigs were mapped to the chimpanzee WGA rDNA sequence (grey line). The contigs that have IGS followed by coding region were split into two at the IGS-coding joint before mapping. Such contigs are represented as two lines on the same horizontal plane instead of a single continuous line. Both parts were compared to the rDNA separately but named as the same contig. The contig Pt_1 that was used to extract the chimpanzee rDNA sequence is represented as a green line and the partial sequence of a second rDNA unit that was removed to obtain the rDNA sequence is represented as a dotted black line. The remaining contigs are represented as black lines. Contig names and their sequence identity with the chimpanzee rDNA sequence (in parentheses) are indicated on the top of each contig. The positions of the contigs in the reference chimpanzee rDNA sequence are indicated at their ends. Scale is shown at the bottom.

2.3.5.1.2. Gorilla reference rDNA unit sequence

A WGA for gorilla was performed using publicly available WGS data from a western lowland female individual named "Kmilaha" (Project Accession: PRJNA169344; Scally *et al.* 2012) using Arachne (assembly statistics is described in Appendix Table 1). Four contigs containing 3 kb to 34 kb of gorilla rDNA sequence were identified by screening this WGA with the human rDNA sequence using BLAST (Table 2.7). The contigs were named Gg_1 to Gg_4 in order of decreasing length (where Gg stand for *Gorilla gorilla*). Contigs Gg_1,

Gg_3 and Gg_4 were merged to obtain a reference gorilla rDNA sequence of length 37,526 bp using Consed (Figure 2.6; see Section 2.2.1.2.3 Step-3). This WGA gorilla rDNA unit consists of a 12,871 bp coding sequence and a 24,655 bp IGS sequence. The WGA gorilla rDNA sequence is 95.9% identical to only remaining non-merged gorilla rDNA containing contig, Gg_2 (Figure 2.6). The variation between this remaining contig and the WGA is mainly because of copy number variation in microsatellites.

Table 2.7: Statistics of the potential gorilla rDNA contigs.

Contig name	Length	Number of reads	Average Coverage
Gg_1	33,494	8,636	162.1
Gg_2	34,320	3,897	67.1
Gg_3	11,471	1,986	108.7
Gg_4	3,044	804	126.1

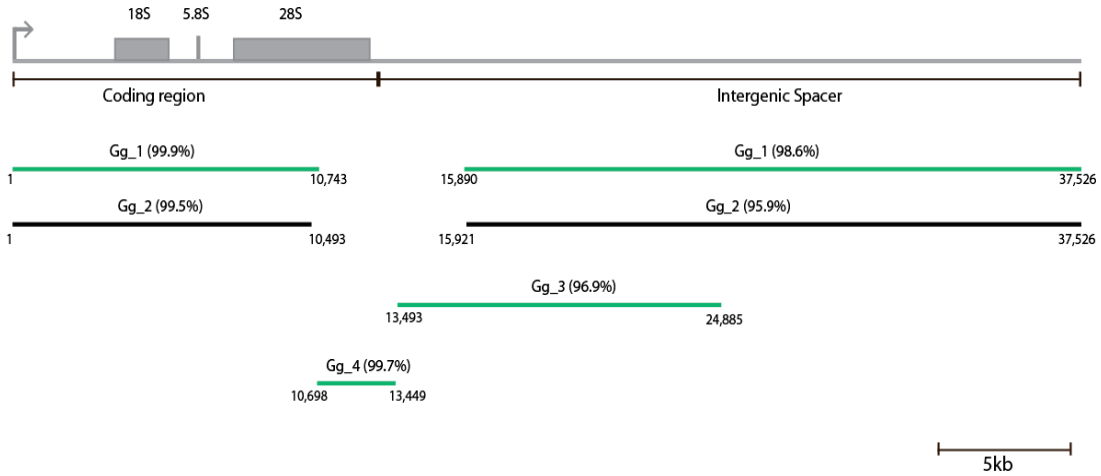


Figure 2.6: WGA contigs containing gorilla rDNA sequence.

The rDNA containing contigs were mapped to the gorilla WGA rDNA sequence (grey line). The contigs that have IGS followed by coding region were split into two at the IGS-coding joint before mapping. Such contigs are represented as two lines on the same horizontal plane instead of a single continuous line. The contigs that were merged to obtain the gorilla rDNA sequence are represented as green lines and remaining contigs are represented as black lines. Other notations are same as in Figure 2.5. Scale is shown at the bottom.

2.3.5.1.3. Orangutan reference rDNA unit sequence

A WGA for orangutan was performed using publicly available WGS data from a female Sumatran individual named "Susie" (Project Accession: PRJNA20869; Locke *et al.* 2011) using Arachne (assembly statistics is described in Appendix Table 1). Nine contigs containing 5 kb to 39 kb of orangutan rDNA sequence were identified by screening this WGA with the human rDNA sequence using BLAST (Table 2.8). The contigs were named Pa_1 to Pa_9 in order of decreasing length (where Pa stands for *Pongo abelii*). Pa_1 and Pa_2 were merged to obtain a reference orangutan rDNA sequence of length 40,901 bp using Consed (Figure 2.7). The WGA orangutan rDNA unit consists of a 13,230 bp coding sequence and a 27,671 bp IGS sequence. The WGA orangutan rDNA sequence is 89% to 98.3% identical to the seven other orangutan rDNA containing contigs (Figure 2.7). The variation between the contigs and the WGA rDNA sequence is higher compared to other primates. The rDNA arrays in orangutan are distributed on nine chromosomes, therefore one possible explanation for this high level of variation could be variation between chromosomes.

Table 2.8: Statistics of the potential orangutan rDNA contigs.

Contig name	Length	Number of reads	Average Coverage
Pa_1	39,995	11,465	202.9
Pa_2	16,632	2,440	96.2
Pa_3	29,219	2,439	61.1
Pa_4	28,600	2,474	65.1
Pa_5	21,699	1316	45.9
Pa_6	18,584	2,271	92.8
Pa_7	16,965	573	26.4
Pa_8	10,686	369	25.5
Pa_9	5,511	172	25.3

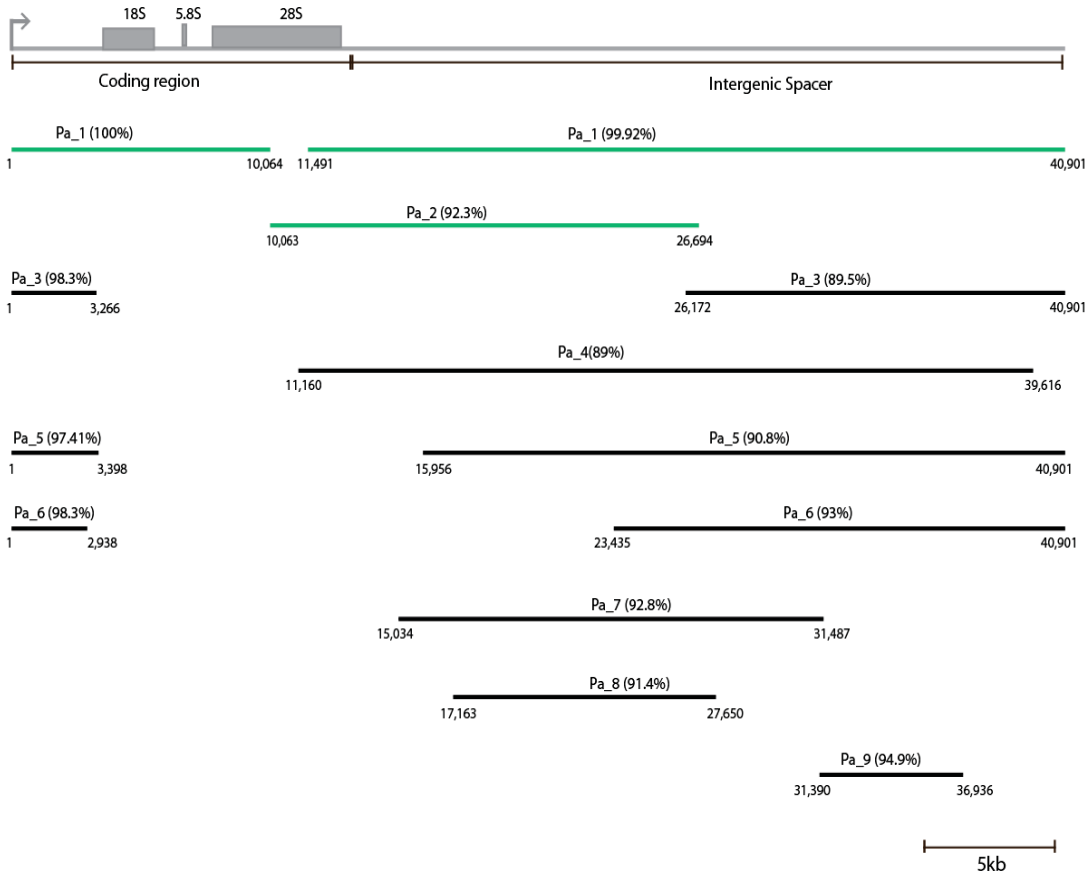


Figure 2.7: WGA contigs containing orangutan rDNA sequence.

The rDNA containing contigs were mapped to the orangutan rDNA sequence (grey line). The contigs that have IGS followed by coding region were split into two at the IGS-coding joint before mapping. Such contigs are represented as two lines on the same horizontal plane instead of a single continuous line. The contigs that were merged to obtain the orangutan rDNA sequence are represented as green lines and remaining contigs are represented as black lines. Other notations are same as in Figure 2.5. Scale is shown at the bottom.

2.3.5.1.4. Gibbon reference rDNA unit sequence

A WGA for gibbon was performed using publicly available WGS data from a northern white cheek female individual named "Asia" (Project Accession: PRJNA20869; Carbone *et al.* 2006) using Arachne (assembly statistics is described in Appendix Table 1). Five contigs containing 3 kb to 37 kb of gibbon rDNA sequence were identified by screening this WGA with the human rDNA sequence using BLAST (Table 2.9). The contigs were named Nl_1 to Nl_5 in order of decreasing length (where Nl stands for *Nomascus leucogenys*). Contigs Nl_1, Nl_3 and Nl_4 were merged to obtain a reference gibbon rDNA sequence of length 42,909 bp using Consed (Figure 2.8). The WGA gibbon rDNA unit consists of a 12,299 bp

coding sequence and a 30,610 bp IGS sequence. The gibbon WGA rDNA sequence is 97.5% to 98.1% identical to the two other gibbon rDNA containing contigs (Figure 2.8). The variation between the contigs is mainly because of copy number variation in microsatellites.

Table 2.9: Statistics of the potential gibbon rDNA contigs.

Contig name	Length	Number of reads	Average Coverage
Nl_1	37,580	5,026	93.9
Nl_2	32,946	6,284	143.5
Nl_3	30,751	2,797	68.7
Nl_4	7,639	138	11.9
Nl_5	3,820	620	85.1

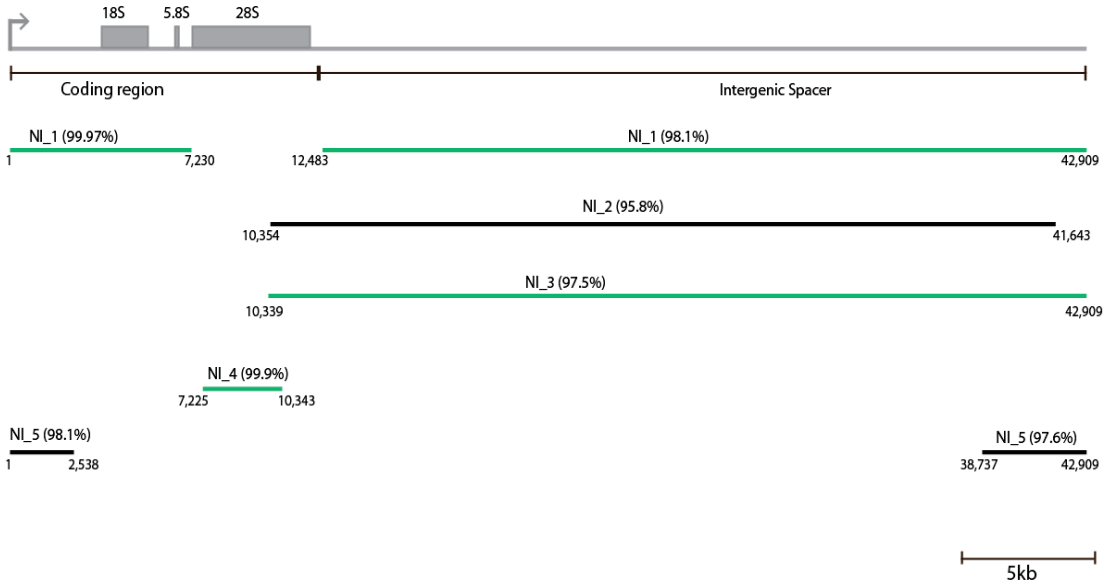


Figure 2.8: WGA contigs containing gibbon rDNA sequence.

The rDNA containing contigs were mapped to the Gibbon rDNA sequence (grey line). The contigs that have coding region attached to the end of IGS were split into two at IGS-coding region joint before mapping. Such contigs are represented as two lines on the same horizontal plane instead of continuous line. The contigs that were merged to obtain the gibbon rDNA sequence are represented as green lines and remaining contigs are represented as black lines. Other notations are same as in Figure 2.5. Scale is shown at the bottom.

2.3.5.1.5. Macaque reference rDNA unit sequence

A WGA for macaque was performed using publicly available shotgun sequencing data from a female individual (Project Accession: PRJNA12537; Gibbs *et al.* 2007) using Arachne (assembly statistics is described in Appendix Table 1). Five contigs containing 3 kb to 38 kb of the macaque rDNA sequence were identified by screening this WGA with the human rDNA sequence using BLAST (Table 2.10). The contigs were named Mm_1 to Mm_5 in order of decreasing length (where Mm stands for *Macaca mulatta*). Contigs Mm_1 and Mm_2 were merged to obtain a reference macaque rDNA sequence of length 41,735 bp using Consed (Figure 2.9). The sequence consists of a 12,979 bp coding sequence and a 28,756 bp IGS sequence. The WGA macaque rDNA sequence is 96.0% to 99.5% identical to the three other macaque rDNA containing contigs (Figure 2.9; Table 2.10). The variation between the contigs is mainly because of copy number variation in microsatellites.

Table 2.10: Statistics of the potential macaque rDNA contigs.

Contig name	Length	Number of reads	Average Coverage
Mm_ 1	38,051	4,889	88.5
Mm_ 2	29,874	3,760	89.9
Mm_3	27,560	1,677	45.0
Mm_4	7,383	531	50.4
Mm_5	3,375	158	29.9

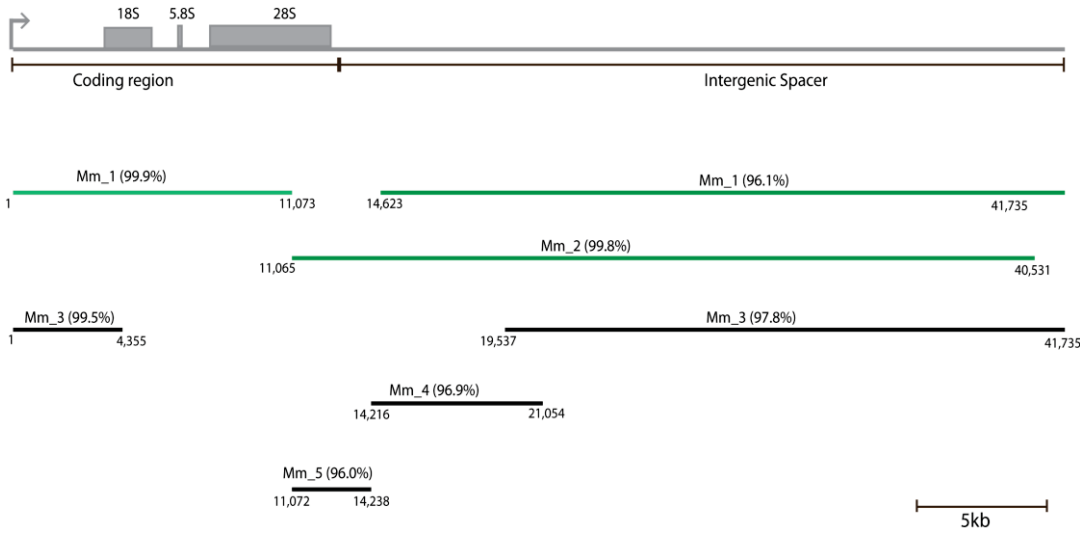


Figure 2.9: WGA contigs containing Macaque rDNA sequence.

The rDNA containing contigs were mapped to the macaque rDNA sequence (grey line). The contigs that have coding region attached to the end of IGS were split into two at IGS-coding region joint before mapping. Such contigs are represented as two lines on the same horizontal plane instead of continuous line. The contigs that were merged to obtain the macaque rDNA sequence are represented as green lines and remaining contigs are represented as black line. Other notations are same as in Figure 2.5 Scale is shown at the bottom.

2.3.5.1.6. Marmoset reference rDNA unit sequence

A WGA for marmoset was performed using publicly available WGS data from a female individual (Project Accession: PRJNA20401) using Arachne (assemble statistics is described in Appendix Table 1). Four contigs containing 13 kb to 75 kb of the common marmoset rDNA sequence were identified by screening this WGA with human rDNA sequence using BLAST (Table 2.11). The contigs were named Cj_1 to Cj_4 in the order of decreasing length (where Cj stands for *Callithrix jacchus*). I derived the callithrix rDNA sequence from Cj_1 because it has one complete and one partial rDNA unit. The partial sequence was removed to give a complete reference callithrix rDNA sequence of length 39,625 bp (Figure 2.10). This common marmoset rDNA unit consists of a 12,720 bp coding sequence and a 26,905 bp IGS sequence. The WGA marmoset rDNA sequence is 93.3% to 99.4% identical to the three other marmoset rDNA containing contigs (Figure 2.10). The variation between the contigs is mainly because of copy number variation in microsatellites.

Table 2.11: Statistics of the potential marmoset rDNA contigs.

Contig name	Length	Number of reads	Average Coverage
Cj_ 1	72,138	10,542	96.6
Cj_2	30,214	2,150	49.2
Cj_3	24,040	990	30.6
Cj_4	13,035	405	21.8

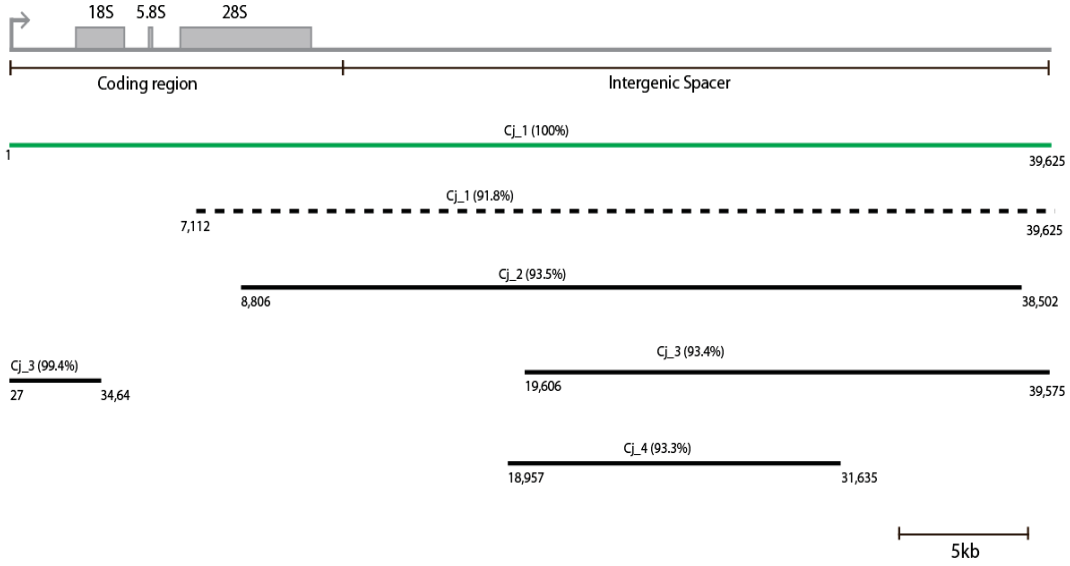


Figure 2.10: WGA contigs containing marmoset rDNA sequence.

The rDNA containing contigs were mapped to the common marmoset rDNA sequence (grey line). The contigs that have IGS followed by coding region were split into two at the IGS-coding joint before mapping. Such contigs are represented as two lines on the same horizontal plane instead of a single continuous line. The contig Cj_1 that was used to extract the common marmoset rDNA sequence is represented as a green line and the partial sequence of a second rDNA unit that was removed to obtain the rDNA sequence is represented as a dotted black line. The remaining contigs are represented as black lines. Other notations are same as in Figure 2.5 Scale is shown at the bottom.

2.3.5.2. Verification of primate rDNA sequences obtained from WGA strategy using BAC clones

The rDNA has two levels of repeat complexity. First, multiple copies are present in the genome and second, there are number of repeat elements within a single rDNA unit. Many repeats found in the rDNA are also present in other regions of the genome. Therefore, it is probable that reads from the other repeats regions may have misassembled with the rDNA reads. Although Arachne gave a fully assembled human IGS, we cannot rule out the possibility that it may not have assembled the rDNA accurately for the other primates. Further, since human rDNA was used as a reference sequence to search for the rDNA contigs in the primate WGAs, it is possible that regions present in other primates, but not in human, remain undetected. To eliminate these possibilities, I decided to verify the WGA rDNA sequences are correct by using BAC clones containing rDNA. To obtain such BAC clones, whole genome BAC libraries obtained from CHORI for the individual primates (except chimpanzee) were screened using high-density hybridization filters (Section 2.2.2.3). Chimpanzee was not included for the BAC sequencing because of the high level of sequence similarity between human and chimpanzee. The chimpanzee rDNA obtained from WGA is 89.3% identical to human rDNA (the 2kb $[TCT]_n$ microsatellite block at position 21,894-23,859 that is variable among the human IGS was excluded before the comparison). Once rDNA containing BAC clones had been identified, I used two different methods, to verify the primate rDNA sequences:

1) **Determination of rDNA unit length in the BAC clones:** I-P*po*I, a homing enzyme, cuts only at one site in the rDNA (in the 18S). The average insert size of a BAC clone is ~175 kb. Given that a primate rDNA unit is ~40 kb and assuming the BAC only contains rDNA units, each rDNA containing BAC should have approximately 4-5 copies of the rDNA. I performed I-P*po*I restriction digestion of the BAC clones for each species and determined the size of the rDNA unit by field-inversion gel electrophoresis (FIGE; Section 2.2.2.7). The presence of the rDNA unit band in the FIGE gel was confirmed by Southern hybridization (Section 2.2.2.3) and the length was calculated using a 5 kb ladder.

2) **Next generation sequencing of the BAC clones:** BAC clones were sequenced on the Illumina platform. *De novo* assembly was performed to obtain the sequences of these BACs (Section 2.2.1.3.3). However, the assemblies were fragmented. To merge the contigs first I mapped them to the WGA rDNA sequences and then combined them to obtain the sequence. D*e novo* assembly is a stringent way to verify the sequence as it is solely driven by the overlapping of reads and does not use any previous sequence information. To verify that the fragmented assembly is obtained not because of absence of data but due to limitation of NGS

data I also constructed the BAC rDNA sequences by mapping the reads to the corresponding WGA rDNA sequences. Complete rDNA unit sequences were obtained for each of the five primate species using this read mapping approach.

2.3.5.2.1. Identification of BAC clones by screening BAC libraries

A BAC library filter for each selected primate species *viz.* gorilla, orangutan, gibbon, macaque and marmoset was screened using a probe made from the human rDNA 18S rRNA coding region (Section 2.2.2.3). Several rDNA positive BACs were identified in the filter (e.g. Figure 2.11-Figure 2.13). Three positive BAC clones of different signal intensities were selected from each filter for further investigation. The BACs were ordered from CHORI as E. *coli* stabs, and the presence of rDNA in the BACs was confirmed by amplifying the 18S region using colony PCR (Figure 2.14; Section 2.2.2.4). BAC clones that gave expected amplicon size (indicated in **bold** in Table 2.12) were selected while those with no product were not included for further study.

Table 2.12: BAC filters screened to identify the rDNA containing BAC clones.

Primate species	Filter	High Intensity BAC clones	Medium Intensity BAC clones	Low Intensity BAC clones
Gorilla	CHORI-255 1A	**CH255-37M20**	**CH255-30A8**	**CH255-31L1**
Orangutan	CHORI-276 3F	**CH276-103L10**	**CH276-120P14**	CH276-126H17
Gibbon	CHORI-271 10A	**CH271-470I24**	**CH271-442M6**	CH271-446H17
Macaque	CHORI-250 1D	**CH250-26D15**	**CH250-46L14**	**CH250-701**
Marmoset	CHORI-259 3E	**CH259-119I6**	**CH259-137E18**	CH259-113H6

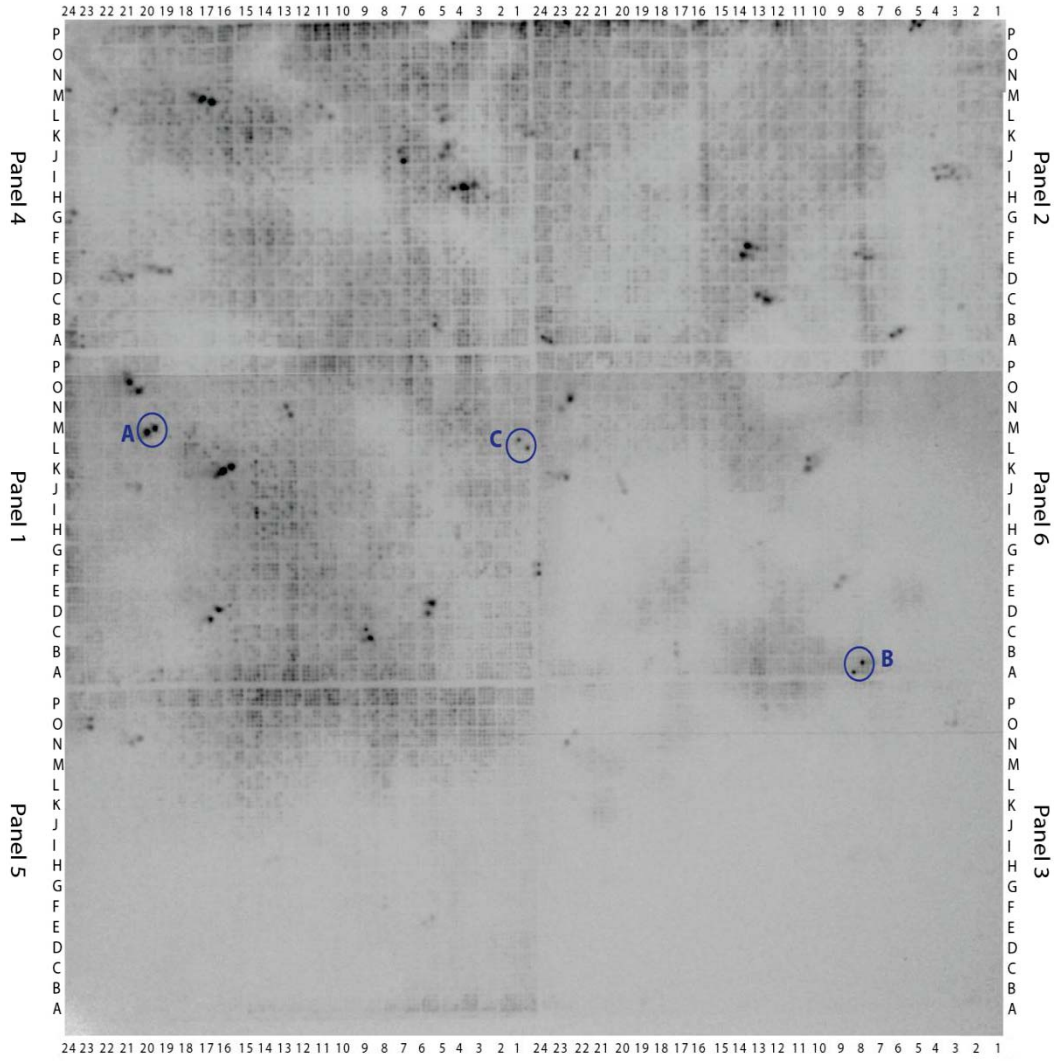


Figure 2.11: Gorilla BAC library filter CHORI-255 1A with signals for the rDNA positive BAC clones.

The filter was hybridized with 18S human rDNA probe. Each rDNA positive BAC clone is present as duplicated signals. The three BAC clones of different signal intensities A) CH255-37M20 (strong signal), B) CH255-30A8 (intermediate signal) and C) CH255-31L1 (weak signal) were selected for further investigation and are highlighted by circles. The filter is divided into six panels by thick lines and each panel is divided into 16x24 boxes by fine lines. The coordinates and panels used to identify the BAC clones are indicated on the sides of the filter.

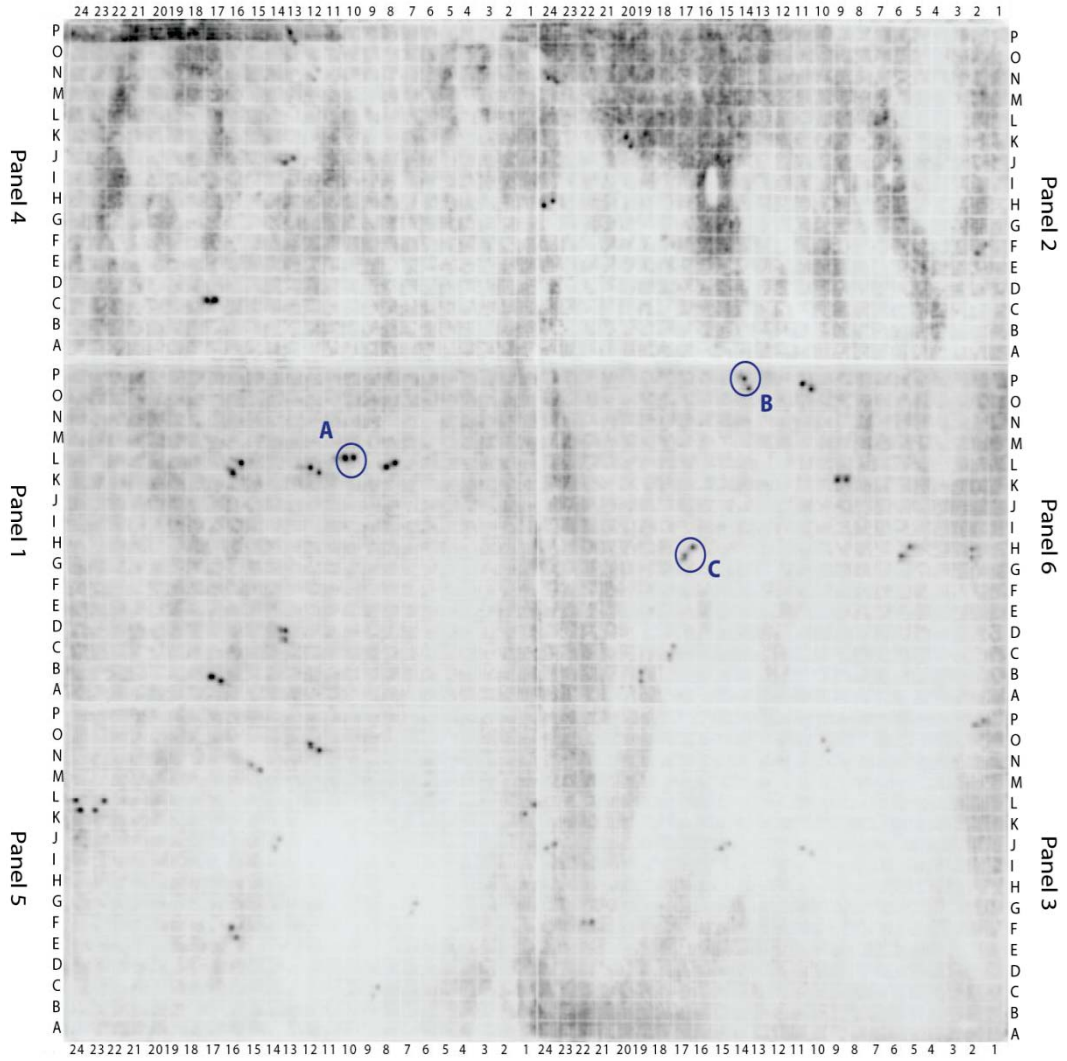


Figure 2.12: Orangutan BAC library filter CHORI-276 3F with signals for the rDNA positive BAC clones.

The filter was hybridized with 18S human rDNA probe. Each rDNA positive BAC clone is present as duplicated signals. The three BAC clones of different signal intensities A) CH276-103L10 (strong signal), B) CH276-120P14 (intermediate signal) and C) CH276-126H17 (weak signal) were selected for further investigation and are highlighted by circles. Other notations are same as in Figure 2.11

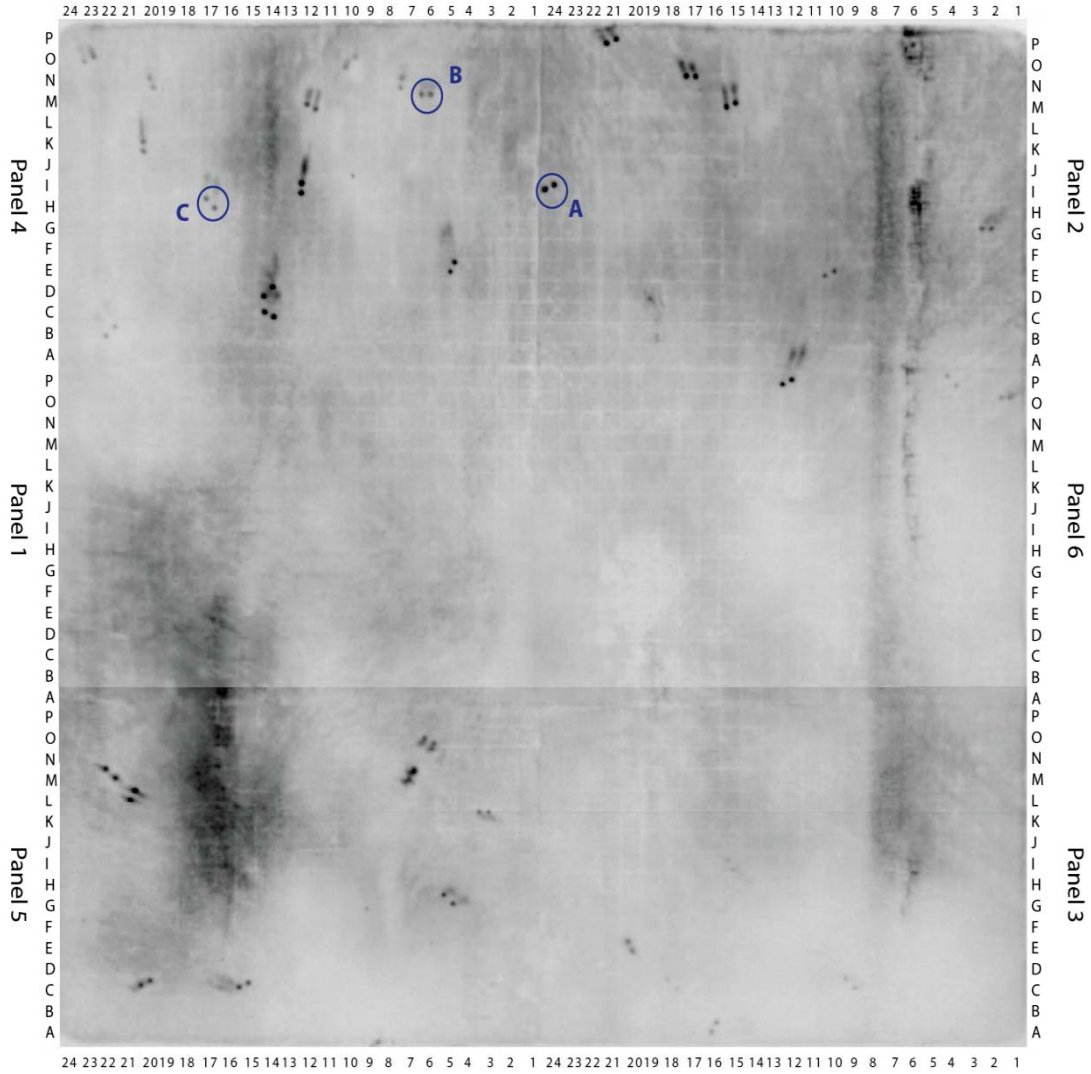


Figure 2.13: Gibbon BAC library filter CHORI-271 10A with signals for the rDNA positive BAC clones.

The filter was hybridized with 18S human rDNA probe. Each rDNA positive BAC clone is present as duplicated signals. The three BAC clones of different signal intensities A) CH271-470I24 (strong signal), B) CH271-442M6 (intermediate signal) and C) CH271-446H17 (weak signal) were selected for further investigation and are highlighted by circles. Other notations are same as in Figure 2.11

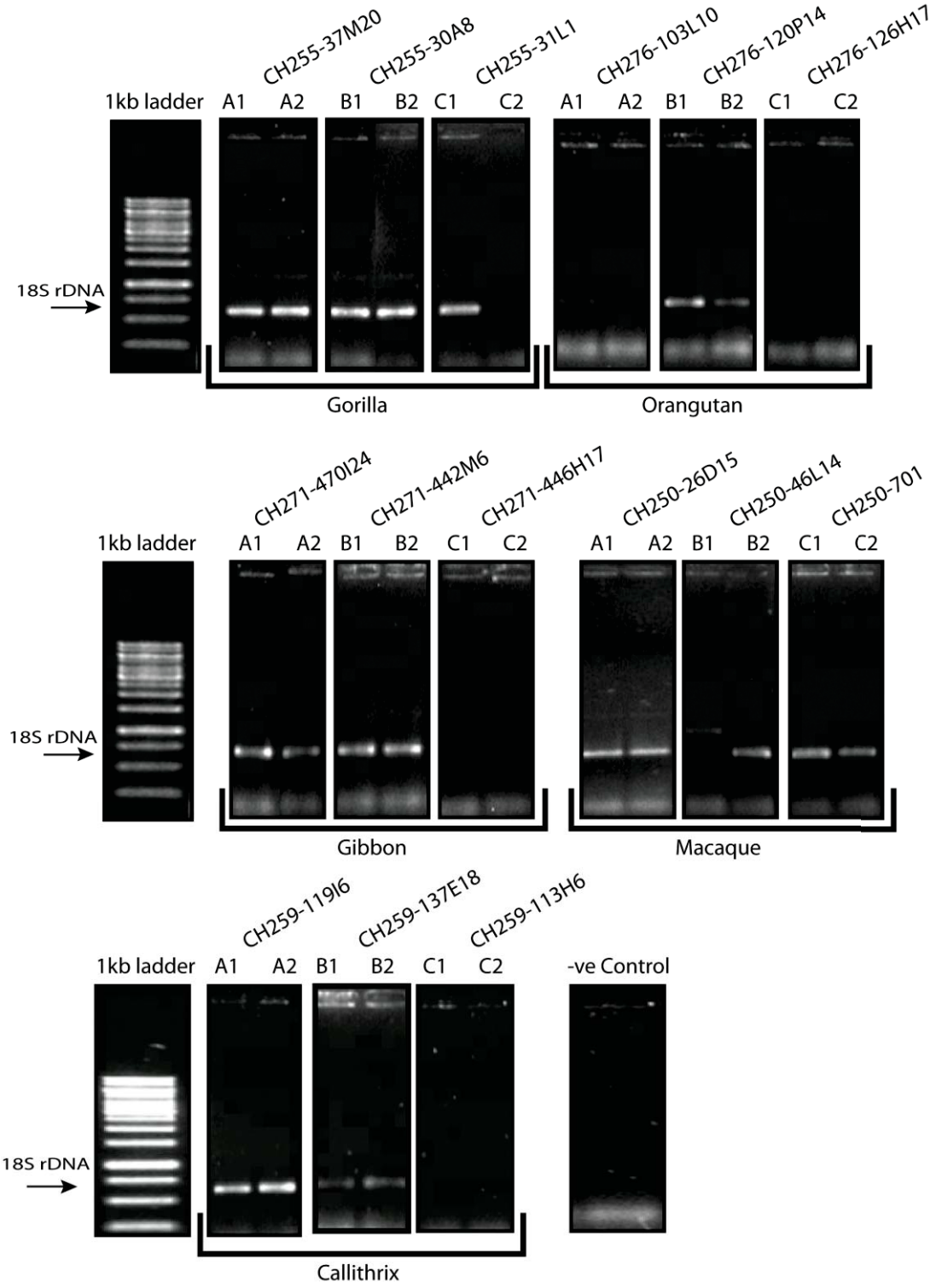


Figure 2.14: Verification of the presence of the rDNA unit in BAC clones.

A 563 bp region of the 18S rDNA (indicated by an arrow) was amplified by colony PCR and separated by electrophoresis on agarose gels to verify the presence of the rDNA in gorilla, orangutan, gibbon, macaque and marmoset BAC clones selected by the BAC filter screening. Two colony isolates were used for each BAC clone. The BAC clone name and isolate number are indicated above each lane.

2.3.5.2.2. Verification of the gorilla rDNA using BAC clones

Three gorilla BAC clones CH255-37M20, CH255-30A8 and CH255-31L1 were selected for further analysis after verifying that the BAC clones selected by the BAC filter screening are rDNA positive (Section 2.3.5.2.1). The estimated rDNA unit size obtained by I-P*po*I digestion of the three BAC clones is ~38 kb (Figure 2.15), which is close to the WGA rDNA length of 37,526 bp. *De novo* assembled rDNA sequences for CH255-37M20, CH255-30A8 and CH255-31L1 are respectively 93.9%, 92.2% and 88.4% identical to the gorilla WGA rDNA (Figure 2.16). To examine if the differences between the BAC assembled rDNA sequence and gorilla WGA rDNA are real or if they are a limitation of NGS assembly, I compared the gorilla WGA rDNA with the mapped BAC rDNA sequences. The sequence identities between the CH255-37M20, CH255-30A8 and CH255-31L1 mapped sequences and the WGA gorilla rDNA are 98.4%, 96.3% and 94.8% respectively. I found that the variation between the assembled and mapped BAC rDNA sequences is because of two reasons: low read coverage and limitation of NGS data to resolve the long stretch of repetitive regions. Certain gaps in the BAC mapped rDNA are either smaller than the corresponding gaps in the BAC assembled rDNA, or are absent from the BAC mapped rDNA (Figure 2.16) because of low read coverage in these regions and therefore they are unable to be resolved by the assembler for e.g. gaps in ITS-2 and 28S in the BAC clone CH255-37M20. The gaps in the BAC assembled rDNA that corresponds to the microsatellites that are absent in the BAC mapped rDNA were unable to be resolved by NGS assembly because of highly repetitive sequence. Specifically comparison between the WGA rDNA and the BAC assembled and mapped rDNA shows that certain gaps in the BAC assembled rDNA coding region are also present in the BAC mapped rDNA coding region (Figure 2.16), verifying that these gaps are present because of the absence of reads.

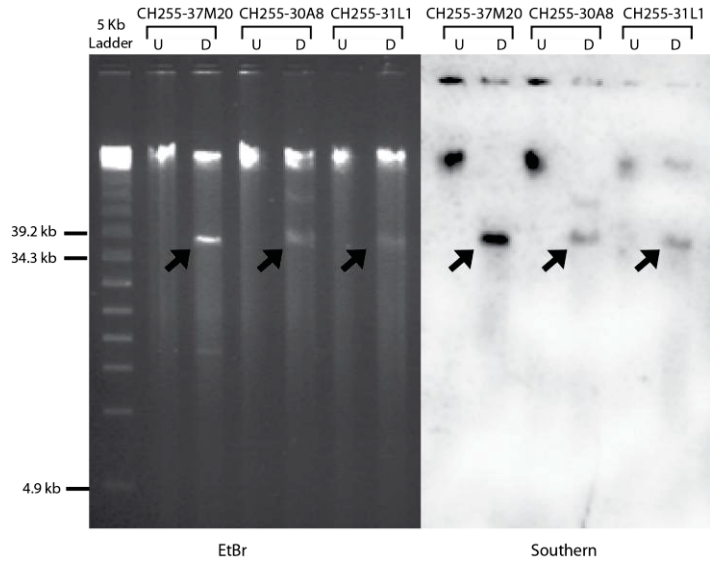


Figure 2.15: Estimating the length of rDNA units in the gorilla BAC clones.

Undigested (U) and I-PpoI digested (D) BACs were ran on a FIGE gel (on left) to determine the size of the rDNA unit in the selected gorilla BAC clones. All three bands are ~38 kb in size. The arrows indicate the rDNA bands. The gel was probed with an 18S rDNA fragment (Southern blot on right) to verify the bands contain rDNA. The bands in the undigested lanes (U) are the BAC clones and E. coli genomic DNA (contamination). In the digested lane of CH255-30A8 the first band (~46 kb) above the rDNA band is probably a rDNA unit with small part of 45S rRNA coding region from the adjacent rDNA unit and the second band is E. coli genomic DNA (no corresponding signal in the Southern blot). In the digested lane of CH255-31L1, the band above the rDNA band is undigested DNA, as it is same size as the band in the corresponding undigested lane (U). The numbers on the left next to the 5 kb ladder are the sizes of the bands used to estimate the rDNA unit size.

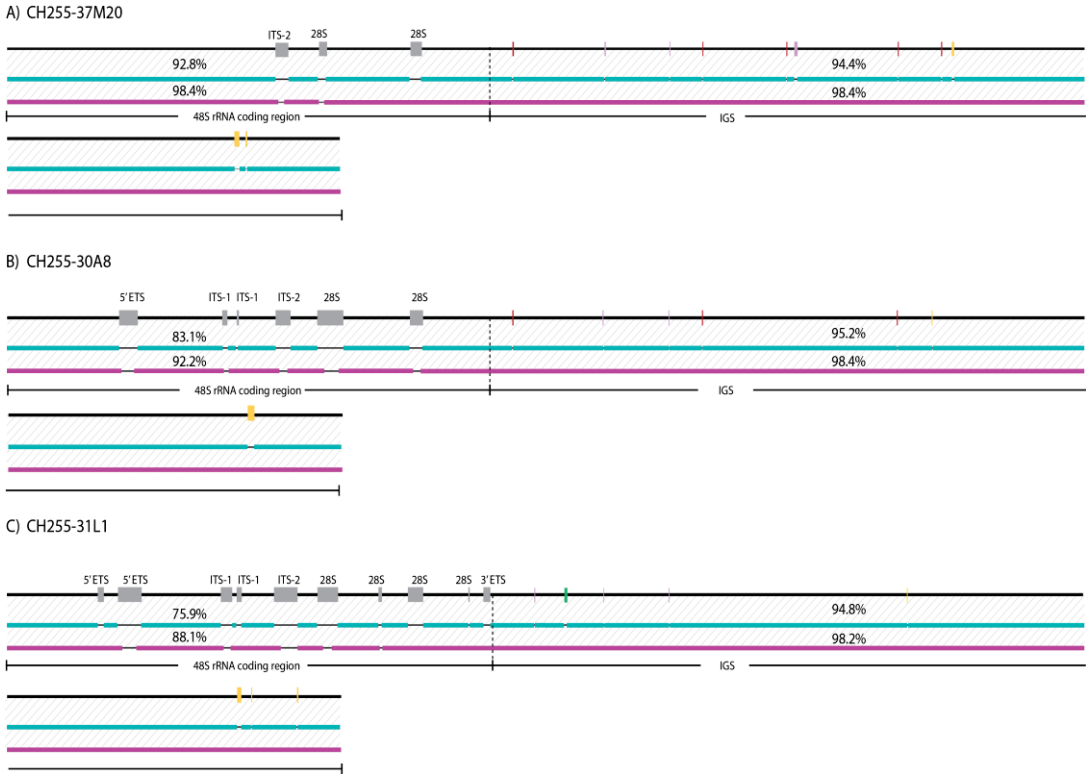


Figure 2.16: Variation between gorilla WGA rDNA and BAC clones gorilla rDNA.

The WGA rDNA sequence (black line) was compared with the assembled (cyan line) and mapped (pink line) BAC rDNA sequences of three BAC clones A) CH255-37M20 B) CH255-30A8 and C) CH255-31L1. The positions of the coding region and the IGS in WGA rDNA are indicated for each BAC clone (the junction of coding and IGS is indicated as a dotted vertical line). The sequence identities for both regions are indicated above the sequence. The gaps in the BAC rDNA sequences are represented as thin grey lines. The regions in the WGA rDNA sequence corresponding to gaps in the assembled and mapped BAC rDNA sequences are represented as boxes (grey for coding region, green for Alu, yellow for microsatellite, purple for low complexity and red for unique region).

2.3.5.2.3. Verification of the orangutan rDNA using BAC clones

Two orangutan BAC clones CH276-103L10 and CH276-120P14 were selected for further analysis after verifying that the BAC clones selected by the BAC filter screening are rDNA positive. The estimated rDNA unit size obtained by I-P*po*I overnight digestion of the three BAC clones is ~42 kb, which is close to the WGA rDNA length of 40,901 bp (Figure 2.17). *De novo* assembled rDNA sequences for CH276-103L10 and CH276-120P14 are respectively 88.9% and 89.8% identical to the orangutan WGA rDNA (Figure 2.18). To examine if the difference between the BAC assembled rDNA sequence and orangutan WGA

rDNA are real or if they are a limitation of NGS assembly, I compared the orangutan WGA rDNA with the BAC mapped rDNA sequences. The sequence identities between the CH276-103L10 and CH276-120P14 mapped sequences and the orangutan WGA rDNA are 97.3% and 97.6% respectively (Figure 2.18). Similar to the case of gorilla the difference in the orangutan BAC rDNA and the WGA rDNA is because of absence of reads from the coding region, low read coverage and limitation of NGS data to resolved repetitive sequences (Figure 2.18).

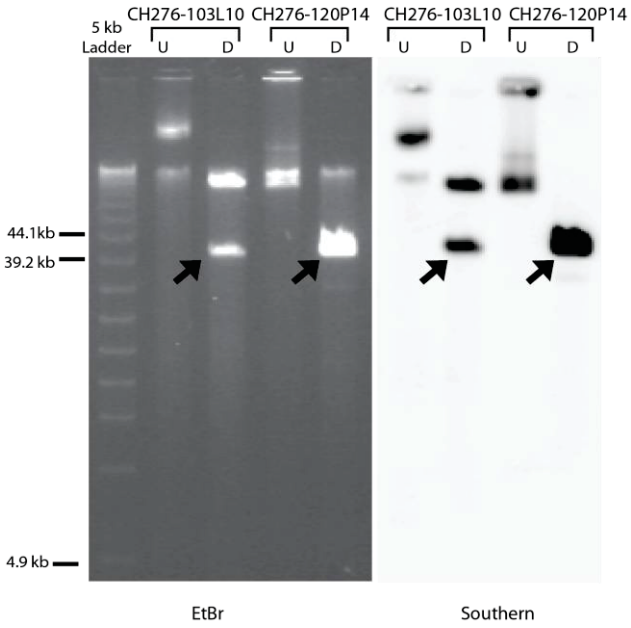


Figure 2.17: Estimating the length of rDNA units in orangutan BAC clones.

Undigested (U) and I-PpoI digested (D) BACs were ran on a FIGE gel (on left) to determine the size of the rDNA unit in orangutan BAC clones. All three bands are ~42 kb in size. The arrows indicate the rDNA bands. The gel was probed with an 18S rDNA fragment (Southern blot on right) to verify that band contains rDNA. The bands in the undigested lanes (U) are the BAC clones and E. coli genomic DNA (contamination). In digested lane of CH276-103L10, the band above the rDNA band is undigested DNA, as it is same size as the band in the corresponding undigested lane (U). In digested lane of CH276-120P14, the band above the rDNA band in the gel is E. coli genomic DNA as it is same size as the band in the corresponding undigested lane and has no corresponding signal in the Southern blot (D). The numbers on the left next to the 5 kb ladder are the sizes of the bands used to estimate the rDNA unit size.

Figure 2.18: Variation between the orangutan WGA rDNA and BAC clones orangutan rDNA.

The WGA rDNA sequence (black line) was compared with the assembled (cyan line) and mapped (pink line) BAC rDNA sequences of two BAC clones A) CH276-103L10 and B) CH276-120P14. Other notations are same as in Figure 2.16.

2.3.5.2.4. Verification of the gibbon rDNA using BAC clones

Two gibbon BAC clones CH271-470I24 and CH271-442M6 were selected for further analysis after verifying that the BAC clones selected by the BAC filter screening are rDNA positive. The estimated rDNA unit size obtained by I-*Ppo*I overnight digestion of the two BAC clones is ~44 kb, which is close to the assembled rDNA length of 42,909 bp (Figure 2.19). *De novo* assembled rDNA sequences for CH271-470I24 and CH271-442M6 are respectively 92.6% and 93.7% identical to the gibbon rDNA (Figure 2.20). To examine if the difference between the BAC assembled rDNA sequence and WGA gibbon rDNA are real or if they are limitation of NGS assembly, I compared the WGA gibbon rDNA with the BAC mapped rDNA sequences. The sequence identities between the CH271-470I24 and CH271-442M6 mapped sequences and the gibbon WGA rDNA are 98.5% (Figure 2.20). Similar to the case of gorilla the difference in the gibbon BAC rDNA and the WGA rDNA is because of absence of reads from the coding region, low read coverage and limitation of NGS data to resolved repetitive sequences (Figure 2.20).

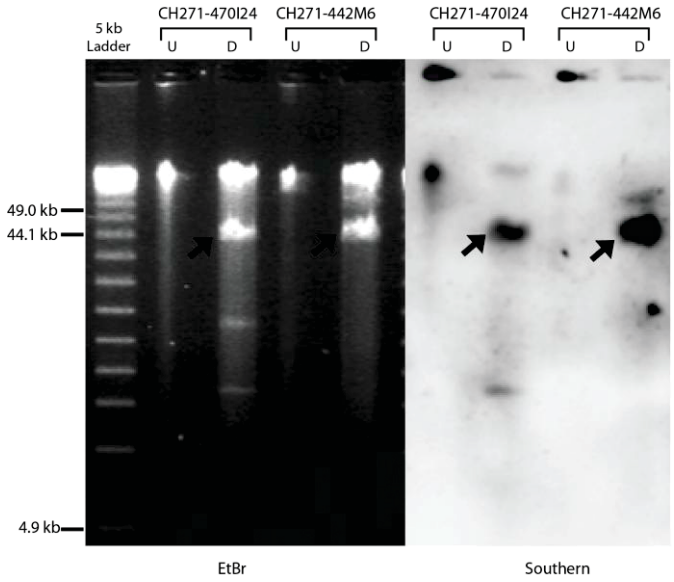


Figure 2.19: Estimating the length of rDNA units in gibbon BAC clones.

Undigested (U) and I-PpoI digested (D) BACs were ran on a FIGE gel (on left) to determine the size of the rDNA unit in gibbon BAC clones. All three bands are ~44 kb in size. The arrows indicate the rDNA bands. The gel was probed with an 18S rDNA fragment (Southern blot on right) to verify that band contains rDNA. The bands in the undigested lanes (U) are the BAC clones and E. coli genomic DNA (contamination). In digested lanes, the band above the rDNA band is the undigested DNA as it is same size as the band in the corresponding undigested lane (U). The bands below the rDNA band in CH271-470I24 are probably from the backbone of the BAC. The numbers on the left next to the 5 kb ladder are the sizes of the bands used to estimate the rDNA unit size.

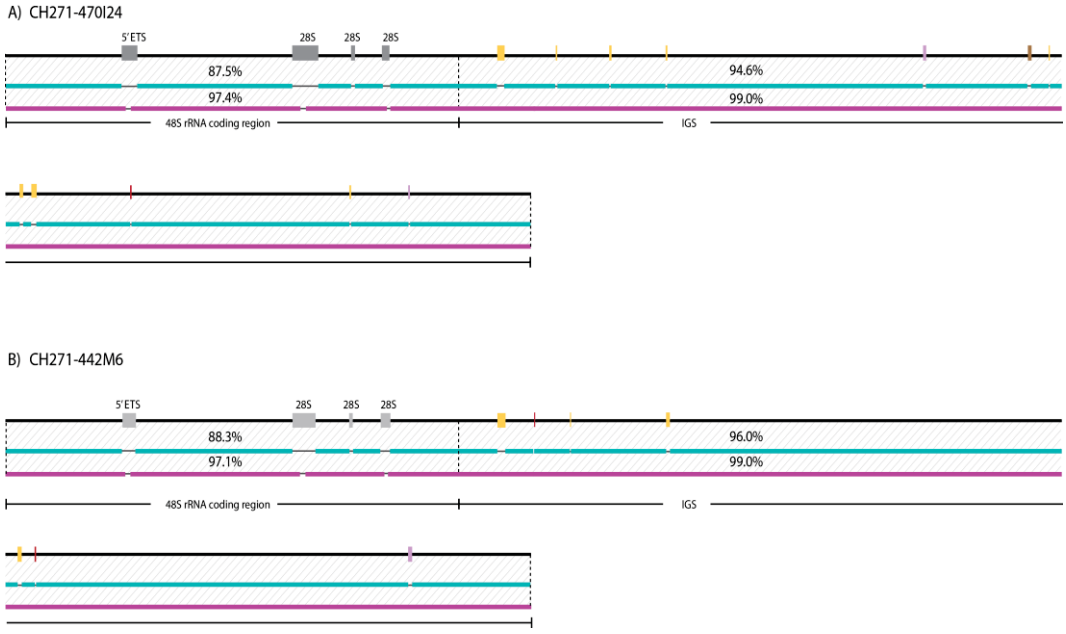


Figure 2.20: Variation between gibbon WGA rDNA and BAC clones gibbon rDNA.

The WGA rDNA sequence (black line) was compared with the assembled (cyan line) and mapped (pink line) BAC rDNA sequences of two BAC clones A) CH271-470I24 and B) CH271-442M6. Other notations are same as in Figure 2.16.

2.3.5.2.5. Verification of the macaque rDNA using BAC clones

Three macaque BAC clones CH250-26D15, CH250-46L14 and CH250-701M6 were selected for further analysis after verifying that the BAC clones selected by the BAC filter screening has are rDNA positive. The estimated rDNA unit size that was obtained by I-*Ppo*I overnight digestion of the three BAC clones is ~42.5 kb, which is close to the WGA rDNA length of 41,735 bp (Figure 2.21). *De novo* assembled rDNA sequences for CH250-26D15, CH250-46L14 and CH250-701 are respectively 92.7%, 91.7% and 93.1% identical to the WGA macaque rDNA (

Figure 2.22). To examine if the difference between the BAC assembled and WGA macaque rDNA are real or if they are a limitation of NGS assembly, I compared the macaque WGA rDNA with the BAC mapped rDNA sequences. The sequence identities between the CH250-26D15, CH250-46L14 and CH250-701 mapped sequences and the macaque rDNA are 97.9%, 97.8% and 98.4% respectively. Similar to the case of gorilla the difference in the macaque BAC rDNA and the WGA rDNA is because of absence of reads from the coding region, low read coverage and limitation of NGS data to resolved repetitive sequences (

Figure 2.22).

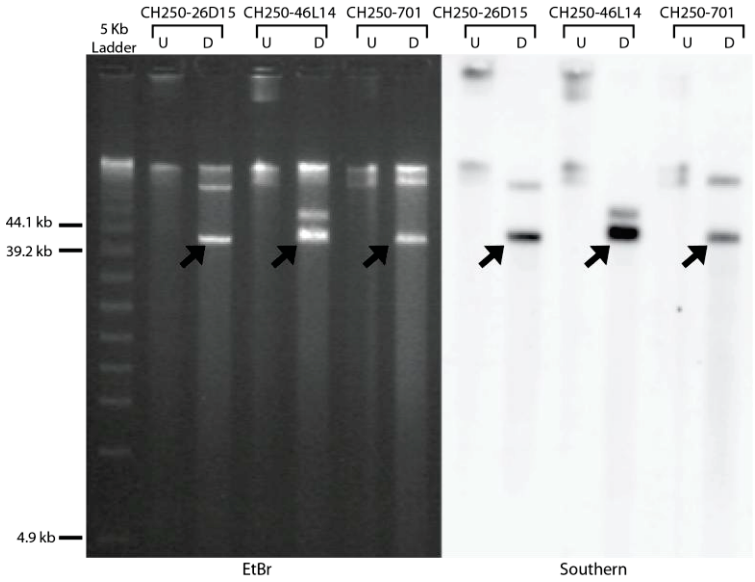


Figure 2.21: Estimating the length of the rDNA units in the macaque BAC clones.

Undigested (U) and (D) I-PpoI digested BACs were ran on a FIGE gel (on left) to determine the size of the rDNA unit in macaque BAC clones. All three bands are ~42.5 kb in size. The arrows indicate the rDNA bands. The gel was probed with an 18S rDNA fragment (Southern blot on right) to verify that band contains rDNA. The bands in the undigested lanes (U) are the BAC clones and E. coli genomic DNA (contamination). In digested lane of CH250-26D15 the two bands above the rDNA band in the gel are complete rDNA unit with a partial unit (lower band) and E. coli genomic DNA (upper band; same size of band in the corresponding undigested lane and have no corresponding signal in the Southern blot). In digested lane of CH250-46L14, the two bands above the rDNA band in the gel are complete rDNA unit with a partial unit. In the digested lane of CH259-119I6, the band above the rDNA band is undigested DNA, as it is same size as the band in the corresponding undigested lane (U). In digested lane of CH250-701, the two bands above the rDNA band in the gel are complete rDNA unit with a partial unit (lower band) and E. coli genomic DNA (upper band; same size of band in the corresponding undigested lane and have no corresponding signal in the Southern blot). The numbers on the left next to the 5 kb ladder are the sizes of the bands used to estimate the rDNA unit size.

Figure 2.22: Variation between macaque WGA rDNA and BAC clones macaque rDNA.

The WGA rDNA sequence (black line) was compared with the assembled (cyan line) and mapped (pink line) BAC rDNA sequences of three BAC clones A) CH250-26D15 B) CH250-46L14 and C) CH250-701. Other notations are same as in Figure 2.16.

2.3.5.2.6. Verification of the marmoset rDNA using BAC clones

Two marmoset BAC clones CH259-119I6 and CH259-137E18 were selected for further analysis after verifying that the BAC clones selected by the BAC filter screening has are rDNA positive. The estimated rDNA unit size that was obtained by I-*Ppo*I overnight digestion of the two BAC clones is ~40 kb, which is close to the WGA rDNA length of 39,625 bp (Figure 2.23). *De novo* assembled rDNA sequences for CH259-119I6 and CH259-137E18 are respectively 91.5%, and 91.8% identical to the marmoset WGA rDNA (Figure 2.24). To examine if the difference between the BAC assembled rDNA sequence and marmoset rDNA are real or if they are limitation of NGS assembly, I compared the marmoset rDNA with the BAC mapped rDNA sequences. The sequence identities between the CH259-119I6 and CH259-137E18 mapped sequences and the marmoset rDNA are

97.2% and 97.5% respectively. The variations between the WGA rDNA and BAC assembled and mapped rDNA sequences are mainly due to the limitation of NGS data to resolve the repeat regions. Similar to the case of gorilla the difference in the marmoset BAC rDNA and the WGA rDNA is because of absence of reads from the coding region, low read coverage and limitation of NGS data to resolved repetitive sequences (Figure 2.24).

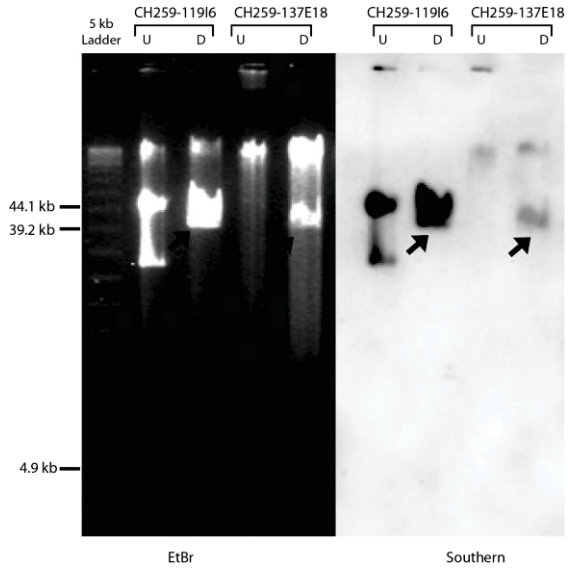


Figure 2.23: Estimating the length of the rDNA units in marmoset BAC clones.

Undigested (U) and (D) I-PpoI digested BACs were ran on a FIGE gel (on left) to determine the size of the rDNA unit in marmoset BAC clones. All three bands are ~40 kb in size. The arrows indicate the rDNA bands. The gel was probed with an 18S rDNA fragment (Southern blot on right) to verify that the bands contain rDNA. The bands in the undigested lanes (U) are the BAC clones and E. coli genomic DNA (contamination). In the digested lane of CH259-137E18, the band above the rDNA band in the gel is E. coli genomic DNA as it is same size of band in the corresponding undigested lane and has no corresponding signal in the Southern blot (D). In the digested lane of CH259-119I6, the band above the rDNA band is the undigested DNA as it is the same size as the band in the corresponding undigested lane (U). The numbers on the left next to the 5 kb ladder are the sizes of the bands used to estimate the rDNA unit size.

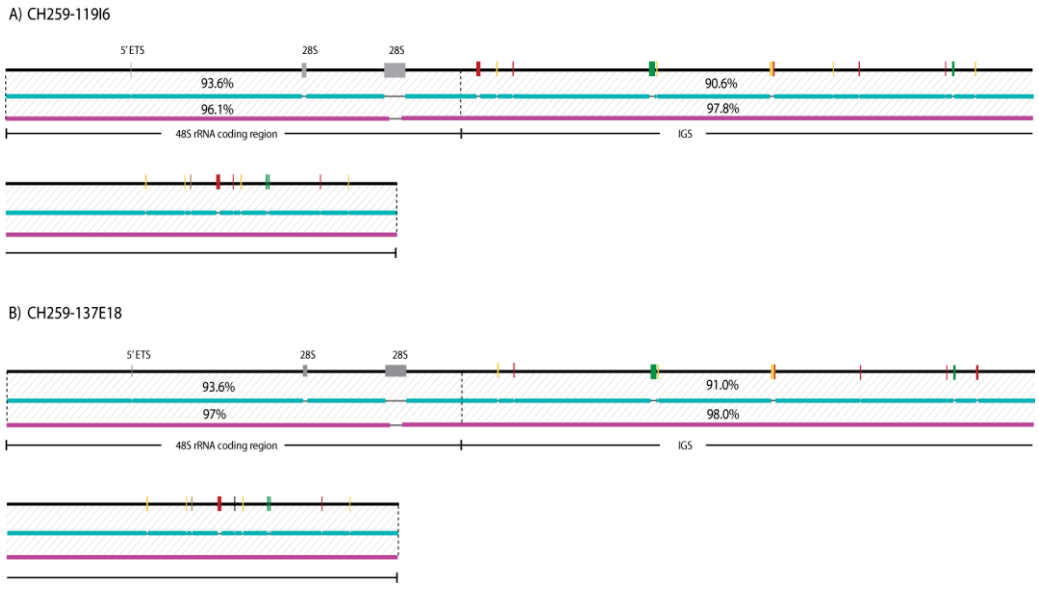


Figure 2.24: Variation between marmoset WGA rDNA and BAC clones marmoset rDNA.

The WGA rDNA sequence (black line) was compared with the assembled (cyan line) and mapped (pink line) BAC rDNA sequences of two BAC clones A) CH259-119I6 and B) CH259-137E18. Other notations are same as in Figure 2.16

2.3.5.3. The primate reference rDNA sequences

The contigs used to construct the rDNA sequences and the ones with greater sequence similarity (>95% similarity) with the WGA rDNA sequence represent more than 60% of the rDNA reads, except for orangutan (Table 2.13). This suggests that the WGA rDNA sequences are appropriate consensus sequences for the majority of the rDNA units. Further, the WGA rDNA sequences and the BAC rDNA units are approximately the same size (Table 2.14). The differences in size are similar for all the primate rDNA units and are always bigger from the FIGE results than the sequenced size (Table 2.14), suggesting the size is slightly overestimated using FIGE gels. The BAC mapped rDNA sequences are >97% identical to the WGA except for gorilla BAC clones CH255-30A8 and CH255-31L1 where identity is slightly lower because of larger gaps in the coding regions due to the absence of reads from the sequencing data. Compared to the BAC mapped rDNA sequences, the BAC assembled rDNA sequences have slightly lower identities with the WGA rDNA sequences, mainly because of gaps in the coding region and the microsatellites. Other labs have also observed the absence of reads from parts of the rDNA coding region in human and mouse rDNA NGS data (Prof. Ross Hannan and Prof. Brian McStay personal communication). The reason for the missing rDNA coding region reads is unknown. One possibility is that the

rDNA coding regions form secondary structures that prevent these sequences from being represented in the sequencing libraries.

Secondly, the gaps in the microsatellite region result from limitations of highly repetitive regions to be resolved with short read NGS data (Section 2.3.1). Overall, the BAC assembled and mapped rDNA units are similar to WGS rDNA sequence, demonstrating that the repeats present in the WGA sequence are actually from the rDNA and that no region is missing from the sequence. Together, the evidence supports the rDNA sequence obtained from the WGA being an accurate representation of the true rDNA sequence. Therefore, I used the WGA sequences as the reference rDNA sequences for all species, and the chimpanzee, gorilla, orangutan, gibbon, macaque and marmoset WGA rDNA sequences are here on referred to as "chimpanzee rDNA", "gorilla rDNA", "orangutan rDNA", "gibbon rDNA", "macaque rDNA", and "marmoset rDNA" respectively.

Table 2.13: The number of rDNA reads represented by the WGA rDNA sequence.

Primate name	Consensus sequence rDNA reads*	Total rDNA reads in WGS data	Consensus reads percentage
Chimpanzee	12,331	19,130	64.5%
Gorilla	15,323	15,323	100%
Orangutan	13,905	23,519	59.1%
Gibbon	14,865	14,865	100%
Macaque	11,015	11,015	100%
Marmoset	10,542	14,087	74.8%

* The reads for the contigs used to construct the rDNA sequence and the contigs that have >95% sequence identity with the constructed rDNA sequence.

Table 2.14: The length variation between the WGA and BAC rDNA sequences of the six primate species.

Primate name	Length of rDNA obtained from WGA (bp)	Estimated Length of rDNA using FIGE (bp)	Length difference (bp)
Chimpanzee	41,773	Not determined	Not determined
Gorilla	37,526	38,169	643
Orangutan	40,901	41,928	1,027
Gibbon	42,909	44,231	1,322
Macaque	41,735	42,681	946
Marmoset	39,625	40,467	842

2.3.6. Characterization of the human and six primate rDNA sequences

After constructing and verifying the reference chimpanzee, gorilla, orangutan, gibbon, macaque and marmoset rDNA sequences, the next step was to characterize them. To start the characterization first I defined the 45S rRNA coding and IGS regions in the sequences by comparing them with the human 45S rRNA coding sequence. Human 45S rRNA coding sequence align completely (end to end) to all the rDNA sequences except marmoset. In marmoset, the 5' ETS is 272 bp shorter than the human 5' ETS. This could be either because marmoset 5' ETS actually being shorter, or failure of the WGA to properly assemble this region. In case of marmoset, only 72 bp of the 5' ETS matches to the rDNA sequence therefore reported marmoset rDNA coding region possibly slightly smaller than the actual coding region. Next, I searched for different repeat elements (Alu elements, microsatellites, LINEs, LTRs and satellites) in the rDNA sequences using the Dfam database and RepeatMasker (Section 2.2.1.4). Several Alu elements were found in the primate IGS and were numbered according to their position in the IGS. Further, I looked for the R-repeat, long repeat and butterfly repeat (Section 1.4.1) previously reported from the human rDNA in the primate rDNA sequences. The different features that are present in the primate rDNA sequences are represented in Figure 2.25-Figure 2.30 and are summarized in Table 2.15 and Table 2.16. The salient characteristics of the different primate rDNA sequences, in comparison to human, are described in following subsections:

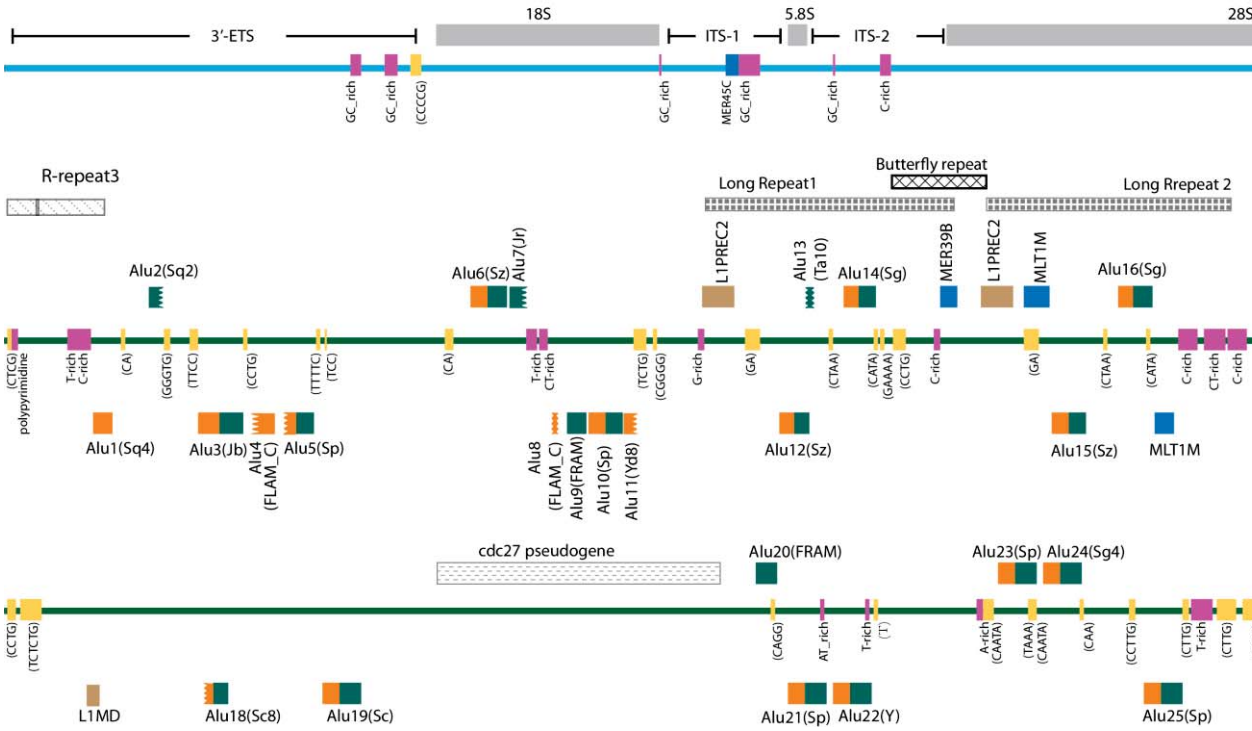


Figure 2.25: The complete chimpanzee rDNA repeat unit.

The rDNA unit has a 13,279 bp coding region (blue line) and a 28,494 bp IGS (green line). The coding regio boxes). The IGS contains several repeat elements including Alu elements (green-orange boxes), microsatellite (brown boxes) and low complexity regions (purple boxes). The names of the repeat elements are indicated sequence features: R-repeat (R-repeat 1-3), long repeats and butterfly repeat that are indicated by diagonally boxes respectively. The cdc27 pseudogene is indicated by as a dashed box. The partial ends of the fragmented The elements above the rDNA are on forward strand and elements below the rDNA are on reverse strand. Th based on the sequence of forward strand and for clarity of the figure, microsatellites are placed on the rDNA. Sca

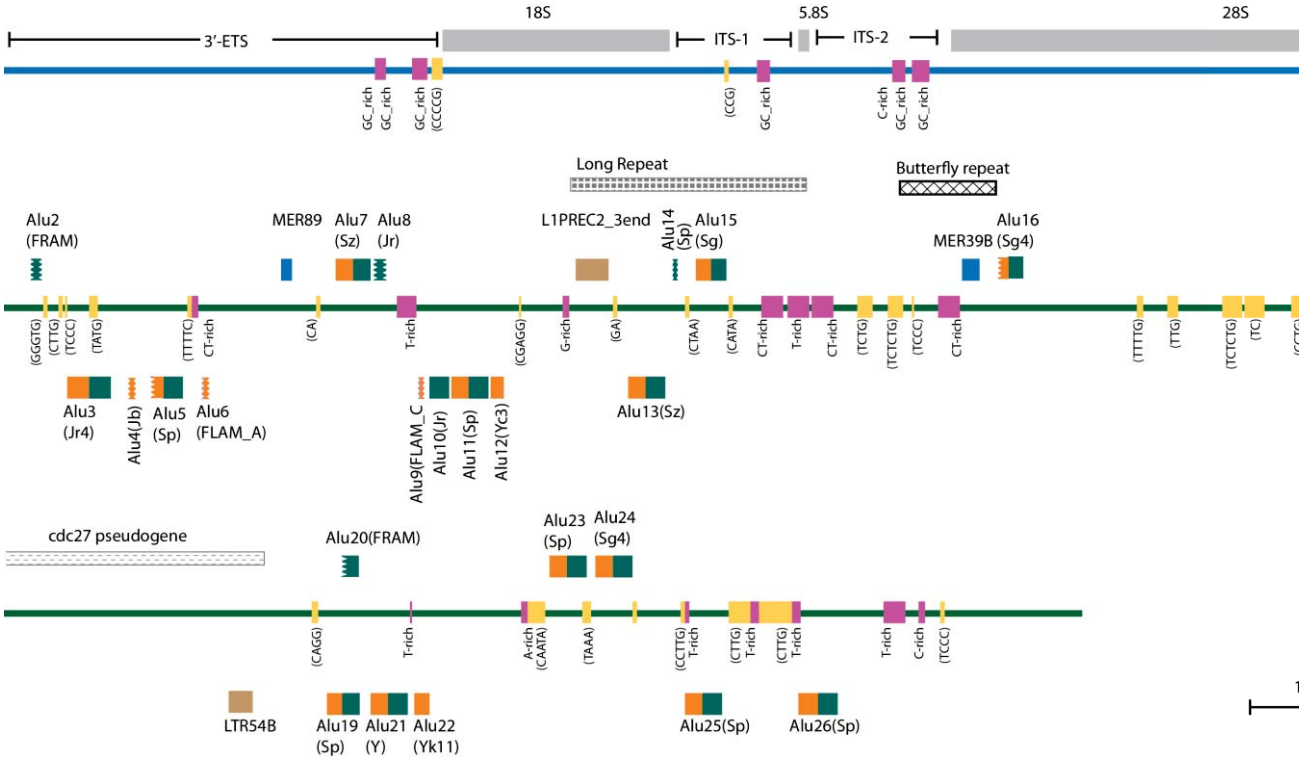


Figure 2.26: The complete gorilla rDNA repeat unit.

The rDNA unit has 12,871 bp a coding region (blue line) and a 24,655 bp IGS (green line). The remain chimpanzee rDNA unit in Figure 2.25.

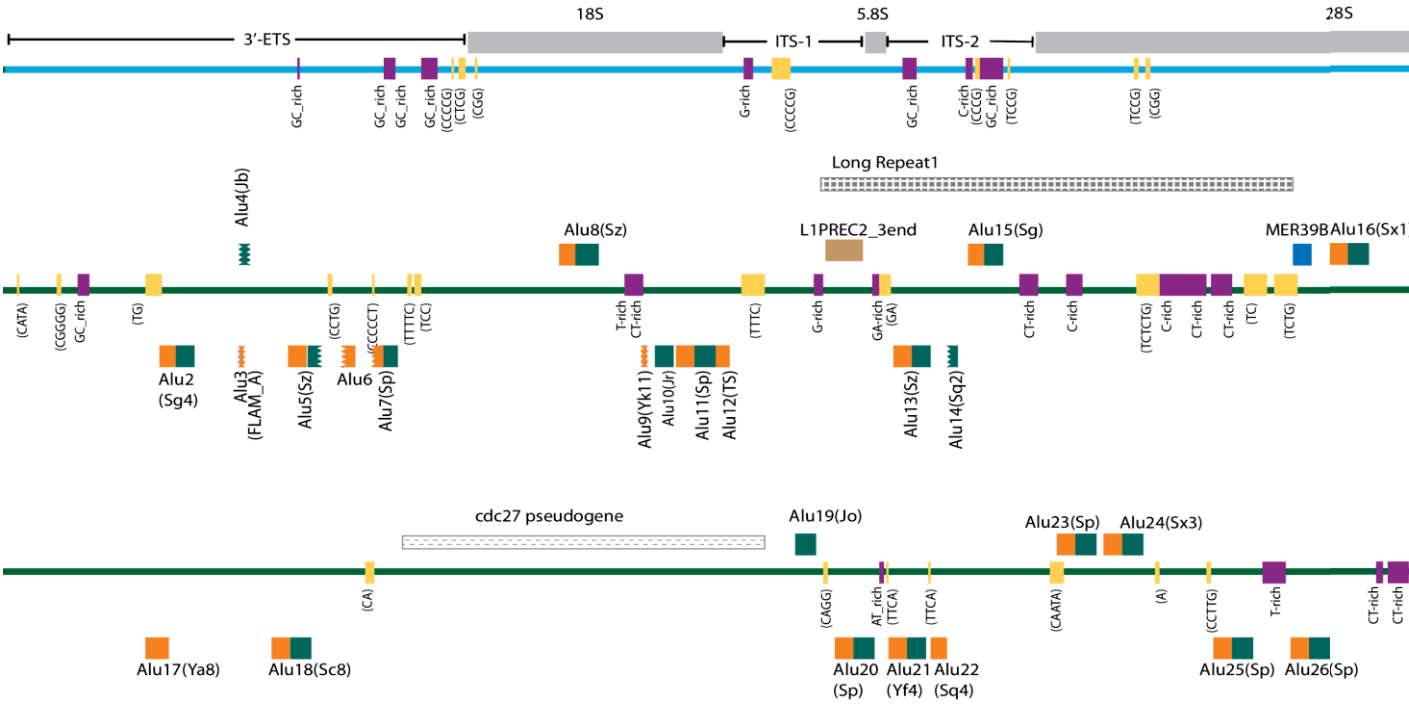


Figure 2.27: The complete orangutan rDNA repeat unit.

The rDNA unit has a 13,230 bp coding region (blue line) and a 27,671 bp IGS (green line). Other notation

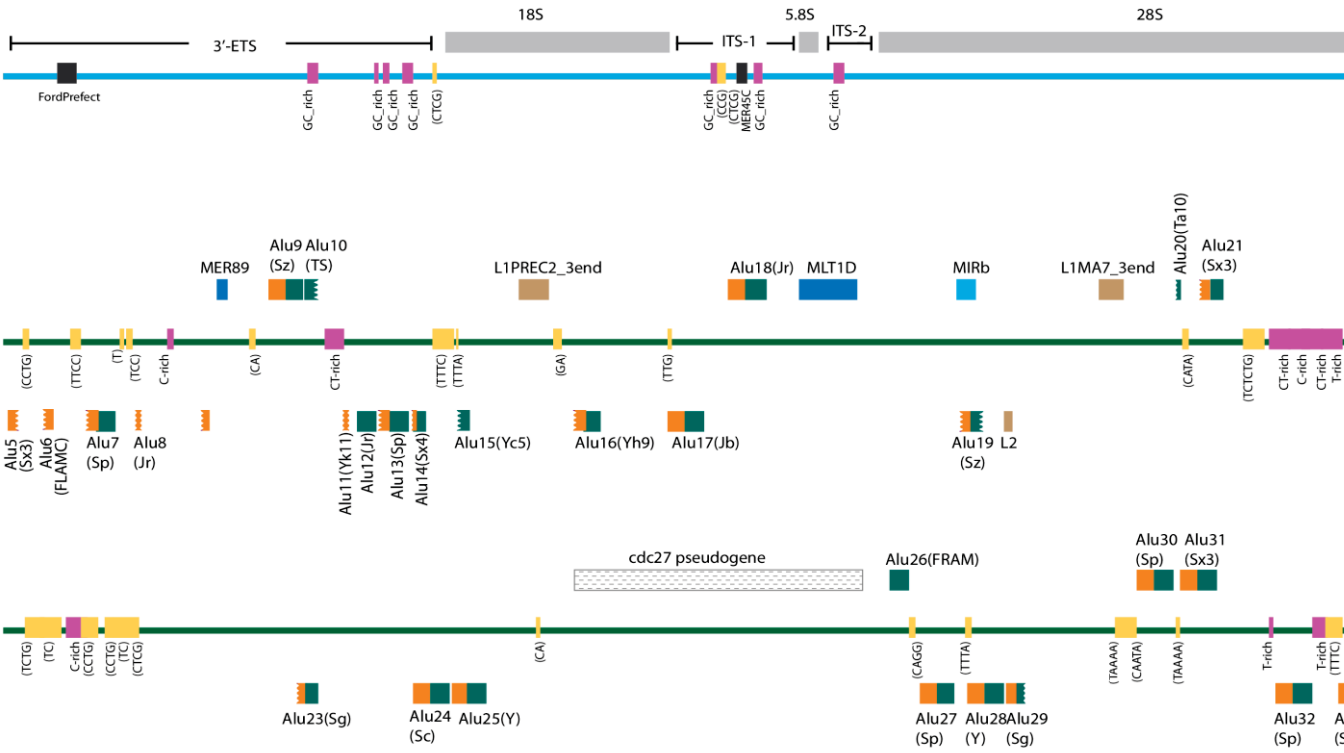


Figure 2.28: The complete gibbon rDNA repeat unit.

The rDNA unit has a 12,299 bp coding region (blue line) and a 30,610 bp IGS (green line). Other notat

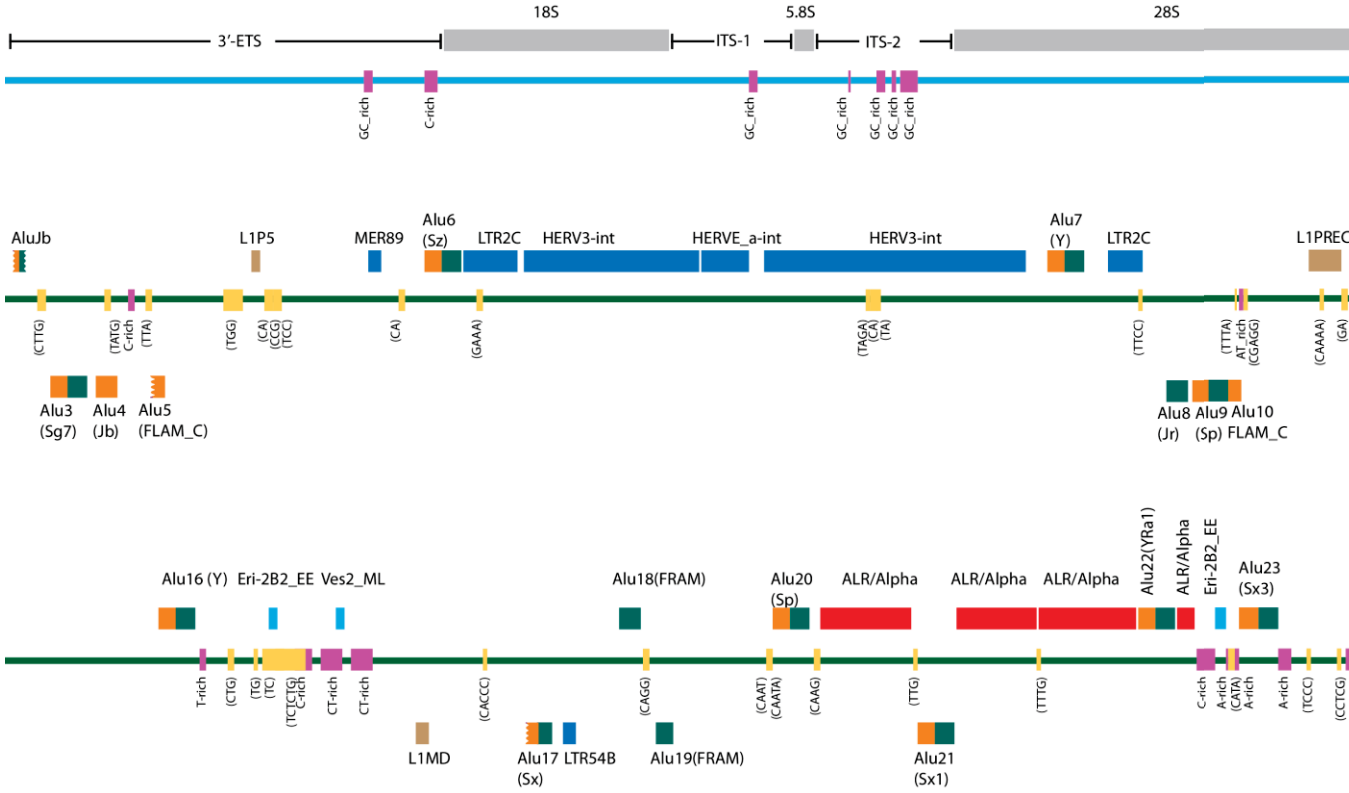


Figure 2.29: The complete macaque rDNA repeat unit.

The rDNA unit has a 12,979 bp coding region (blue line) and a 28,756 bp IGS (green line). The IGS c (red boxes). Other notations are same as in Figure 2.25.

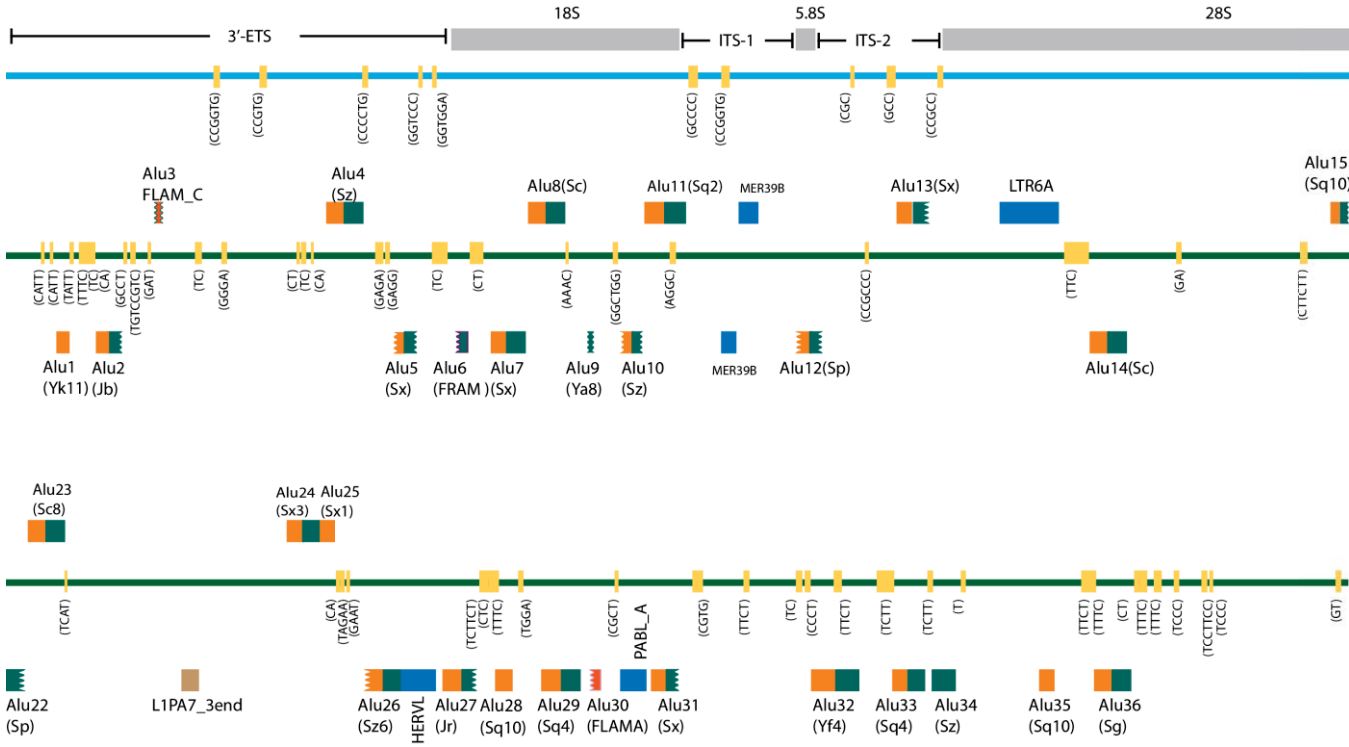


Figure 2.30: The complete marmoset rDNA repeat unit.

The rDNA unit has a 12,720 bp coding region (blue line) and a 26,905 bp IGS (green line). Other notations ar

2.3.6.1. Coding region

The length of the 45S rRNA coding sequence in the six primate species is similar to human *i.e.* approximately 13 kb, except for gibbon, which has slightly smaller coding region (Table 2.15). As expected, as we move from chimpanzee to marmoset, the pairwise sequence identity with human decreases for the coding region (Table 2.15). As expected, the 18S and 5.8S rDNA primate sequences are almost identical to the human. However, because of variable regions, the 28S rRNA gene sequence is variable in length (4.6 kb to 5 kb) and has a lower sequence identity (Table 2.15) than the 18S and 5.8S. As anticipated, overall sequence variation in the coding region of the primates is mainly because of differences in the ITS and ETS regions.

Table 2.15: rDNA sequence comparison between human and the six primate species.

Features		Human	Chimpanzee	Gorilla	Orangutan	Gibbon	Macaque	Marmoset
Total length (bp)		43,972	41,773	37,526	40,901	42,909	41,735	39,625
Coding region length in bp (Identity compared to human)	**Total**	13,357	13,279 (94.2%)	12,871 (92.0%)	13,230 (88.7%)	12,299 (83.1%)	12,979 (86.8%)	12,720 (79.7%)
	18S	1,869	1,867 (99.8%)	1,864 (99.7%)	1,964 (98.9%)	1,868 (99.8%)	1,868 (99.8%)	1,864 (99.7%)
	5.8S	157	157 (100%)[a]	97 (100%)	157 (99.4%)	157 (100%)	157 (100%)	157 (100%)
	28S	5,070	5,056 (97.4%)	4,796 (93.3%)	4,970 (92.3%)	4,667 (88.9%)	4,785 (92.7%)	4,997 (91.1%)
IGS region length in bp (Identity compared to human)		30,615	28,494 (79.8%)	24,655 (73.2%)	27,671 (61.9%)	30,610 (63.4%)	28,756 (32.7%)	26,905 (25.4%)

2.3.6.2. Microsatellites:

Similar to human, the microsatellite component of the rDNA unit in all six primate species is higher than the genome wide average (Table 2.16). Compared to the human rDNA, the microsatellite content of other primate rDNA sequence is lower because they lack the two long $[TC]_n$ repeat blocks that are present in the human (Figure 1.3). The microsatellites are not identical between the human and primate rDNA sequences, however the microsatellites located near to Alu elements are more often the same type than those in other parts of the IGS. Studies have shown that the microsatellites present in the promoter regions are relatively more conserved compared to those from other parts of the genome (Sawaya *et al.*

2013), particularly microsatellites GA/TC and CTT/GAA (Morris *et al.* 2010). The majority of the microsatellites that are conserved and in close proximity to Alus are GA/TC and CTT/GAA microsatellites (Figure 1.3). It has been previously shown that changes in the copy number of microsatellites in promoter regions can modulate gene expression (Iglesias *et al.* 2004; Li *et al.* 2004). The mechanism by which this occurs is still not known. It is thought that the property of microsatellites to alter the structure of the DNA from the general B-form to Z-form, H-form and G-quadruplex result in the changes of the expression level (Palumbo *et al.* 2008; Shklover *et al.* 2010; Sawaya *et al.* 2012). Since the microsatellites found in close proximity to Alus are usually of the same type, it is possible that these microsatellites regulate Alu expression in the IGS through a similar mechanism.

2.3.6.3. Satellites:

A unique feature of the macaque rDNA is the presence of alpha satellites in the IGS. Three blocks of alpha satellites are present in the macaque IGS, covering 2,386 bp (Figure 2.29). Alpha satellites are characteristic features of centromeres (Manuelidis 1978; Wevrick and Willard 1989). In macaque, the rDNA is localized pericentromerically, therefore, the alpha satellites may have been inserted into the macaque IGS and then fixed as a result of rDNA homogenization.

Table 2.16: Repeat composition of the primate rDNA sequences.

Repeat Elements	Human	Chimpanzee	Gorilla	Orangutan	Gibbon	Macaque	Marmoset
Micro-satellites	20.3[a] (0.8)[b]	8.7 (0.8)	6.6 (1.1)	7.7 (0.8)	7.7 (0.8)	6.2 (0.8)	10.4(0.9)
Alus (SINE)	13.3 (10.6)	13.1 (10.3)	13.3 (8.3)	13.6 (9.8)	16.0 (10.6)	14.2 (10.1)	18.2 (11.0)
LINE	4.3 (20.4)	1.6 (21.6)	1.3 (19.8)	1.1 (22.2)	1.6 (21.8)	1.5 (19.1)	0.4 (21.8)
LTR	1.2 (8.3)	0.7 (9.0)	0.4 (8.4)	0.9 (9)	1.7 (8.7)	12.2 (8.4)	3.60 (1.0)

[a] 9.34% if TC track is removed. [b] The number in the parenthesis is the repeat percentage of entire genome.

2.3.6.4. Alu elements:

Alu elements are most abundant repeat element of the IGS in all the primates (Table 2.16). Previous studies have shown that several Alu elements in the human rDNA are conserved in apes and macaque (Gonzalez *et al.* 1989; Gonzalez *et al.* 1993). These studies used Southern hybridization to demonstrate Alu element conservation in the IGS, hence they can only confirm the positional conservation of the Alu elements but not the level of sequence identity. To determine the sequence identity, I compared the human Alu elements with their corresponding primate IGS Alu elements. Most of the Alu elements have high sequence identity from 89%-97% demonstrating that not only the position but also sequences of the Alu elements are conserved. The results of pairwise comparison are summarized in Table 2.17. Interestingly, $Alu_{human}27$ and 28 that were reported to be conserved in macaque (Gonzalez *et al.* 1993) are actually are on the opposite strand in macaque and therefore are not orthologous. The pairwise comparisons show that as we move from human to gibbon the level of sequence identity for certain Alu elements ($Alu_{human}6$, 10, 13, 15 and 17) decreases more rapidly than others, suggesting that a selective constrain is being applied to some of the Alu elements. This suggests that these Alu elements may have some functional role. Previous study have shown that about 30% of the gorilla genome has a higher DNA sequence identity to human or chimpanzee than human and chimpanzee have to each other, and this is more pronounced for regions surrounding the genes than the genes themselves (Scally *et al.* 2012). Interestingly, $Alu_{human}1$, 10, 15, 19, 20, 23 and 26 have this same pattern of sequence identity between human, chimpanzee and gorilla. The role of these Alu elements in the IGS is not known. Several Alu elements in the human genome are transcribed to form noncoding RNAs therefore, it is possible that these highly conserved human Alu elements may have a similar role.

Table 2.17: Pairwise sequence comparisons showing the level of sequence conservation between

Human	Chimpanzee	Gorilla	Orangutan
$Alu_{human}1$	$Alu_{chimp}1$ (92.1, 7.9, 0)[a]	$Alu_{gorilla}1$ (93.4, 6.6, 0)	[b]
$Alu_{human}2$			
$Alu_{human}3$			
$Alu_{human}4$			
$Alu_{human}5$			
$Alu_{human}6$	$Alu_{chimp}5$ (94.2, 5.4, 0.4)	$Alu_{gorilla}5$ (92.6, 5.8, 1.7)	$Alu_{orang}7$ (88.0, 10.7, 1.4)
$Alu_{human}7$	$Alu_{chimp}6$ (97.3, 2.4, 0.3)	$Alu_{gorilla}7$ (90.9, 3.9, 5.7)	$Alu_{orang}8$ (89.6, 8.6, 2.0)
$Alu_{human}8$			
$Alu_{human}9$			
$Alu_{human}10$	$Alu_{chimp}9$ (94.2, 2.6, 0.7)	$Alu_{gorilla}10$ (94.9, 5.1, 0)	$Alu_{orang}10$ (89.9, 7.8, 2.4)
$Alu_{human}11$	$Alu_{chimp}10$ (97.5, 2.5, 0)	$Alu_{gorilla}11$ (94.7, 4.9, 0.3)	$Alu_{orang}11$ (89.0, 9.2, 1.3)
$Alu_{human}12$		$Alu_{gorilla}12$ (96.7, 3.3, 0)	$Alu_{orang}12$ (90.3, 6.6, 3.3)
$Alu_{human}13$	$Alu_{chimp}12$ (94.9, 4.8, 0.3)	$Alu_{gorilla}13$ (92.4, 3.7, 4.1)	$Alu_{orang}13$ (89.8, 8.6, 1.7)
$Alu_{human}14$			
$Alu_{human}15$	$Alu_{chimp}14$ (94.6, 5.0, 0.4)	$Alu_{gorilla}15$ (95.4, 4.2, 0.4)	
$Alu_{human}16$	$Alu_{chimp}15$ (95.5, 3.8, 0.7)		
$Alu_{human}17$	$Alu_{chimp}16$ (94.6, 5.02, 0.4)		
$Alu_{human}18$	$Alu_{chimp}17$ (97.2, 2.3, 0.5)	$Alu_{gorilla}16$ (95.3, 4.7, 0)	
$Alu_{human}19$	$Alu_{chimp}18$ (91.1, 1.1, 1.1)	$Alu_{gorilla}17$ (98.1, 1.1, 0.6)	
$Alu_{human}20$	$Alu_{chimp}19$ (97.7, 2.3, 0)	$Alu_{gorilla}18$ (98.1, 1.6, 0.3)	$Alu_{orang}17$ (89.0, 8.2, 2.9)
$Alu_{human}21$	$Alu_{chimp}20$ (96.9, 3.1, 0)		$Alu_{orang}18$ (94.4, 4.9, 0.6)
$Alu_{human}22$	$Alu_{chimp}21$ (97.6, 2.3, 0)	$Alu_{gorilla}19$ (96.7, 2.5, 0.7)	$Alu_{orang}19$ (94.2, 3.3, 2.3)
$Alu_{human}23$	$Alu_{chimp}22$ (92.1, 4.1, 0.3)	$Alu_{gorilla}21$ (96.7, 3.3, 0)	$Alu_{orang}20$ (92.5, 4.1, 2.7)
$Alu_{human}24$		$Alu_{gorilla}22$ (99.2, 0.8, 0)	$Alu_{orang}21$ (88.1, 8.1, 3.2)
$Alu_{human}25$	$Alu_{chimp}23$ (97.7, 2.3, 0)	$Alu_{gorilla}23$ (96.1, 2.6, 1.3)	$Alu_{orang}22$ (93.5, 4.6, 2.0)
$Alu_{human}26$	$Alu_{chimp}24$ (97.1, 2.9, 0)	$Alu_{gorilla}24$ (97.4, 2.9, 0)	$Alu_{orang}23$ (91.9, 5.3, 2.9)
$Alu_{human}27$	$Alu_{chimp}25$ (97.1, 2.9, 0)	$Alu_{gorilla}25$ (95.1, 4.2, 0.7)	$Alu_{orang}24$ (90.2, 7.8, 1.6)

[a.] Percent sequence identity, percent mismatches and percent gaps are indicated in the parentheses [b.] Grey boxe
the specific primate that is orthologous to human Alu element.

2.3.6.5. Additional repeat elements:

The additional repeat elements (R-repeats, Long repeats and Butterfly repeats) that have been reported from the human IGS by Gonzalez *et al.* (Gonzalez and Sylvester 1995) are also present in the IGS of other apes. In chimpanzee as in to human, two copies of the Long repeats and Butterfly repeats are present while in gorilla and orangutan only one copy. This suggests that these repeat elements duplicated in the IGS after the common ancestor of chimpanzee and human diverged.

2.3.7. Phylogenetic footprinting to identify potential noncoding functional elements in the IGS

In order to identify potential functional elements in the human IGS, I next compared the human rDNA with the primate rDNA sequences to identify phylogenetic footprints (regions that are highly conserved during evolutionary timeline) in the human rDNA IGS. To achieve this goal, I initially made a multiple sequence alignment (MSA) by aligning human rDNA with the rDNA sequences of chimpanzee, gorilla, orangutan, gibbon, macaque and marmoset. However, the MSA was poorly aligned because of the low sequence identity between human and marmoset rDNA. Therefore, marmoset rDNA was removed from the analysis and rDNA sequences were realigned to obtain the $MSA_{human\text{-}macauqe}$ that was used for the further study. Next, to observe the level of sequence conservation in the human IGS a similarity plot was generated from $MSA_{human\text{-}macauqe}$ (Figure 2.31**)** using Synplot (Section 2.2.1.5). The human rDNA sequence in $MSA_{human\text{-}macauqe}$ has long runs of gaps that are predominantly the result of the satellite blocks in the macaque rDNA. Because the goal of this study is to search for the conserved regions in the human IGS, all the columns in the MSA with gaps in the human rDNA were removed before generating the similarity plot. This facilitated visualization of the positions of conserved regions (obtained in the next step) relative to the human rDNA. A 75 bp sliding window with 1 bp increment was used to generate the similarity plot (Section 2.2.1.5). The plot represents the sequence conservation between the human and primate rDNA sequences. To demarcate the conserved regions in the human IGS, a cutoff of 80% identity with a minimum length of 10 bp was used (Section 2.2.1.5). The average rDNA sequence identity among the selected primates is 61.1%, and this decreases to 51.6% when just the IGS is considered. This arbitrary 80% cutoff mark was chosen as it is much higher than the average sequence identity (61.1%), and therefore represents a conservative cutoff value. Further, a comparison between different database studies showed that the average minimum sequence length of the binding sites of transcription factors is ~16 bp (Kulakovskiy *et al.* 2013). Therefore, the smaller cutofff of 10

bp was used to ensure most the potential protein binding sites in the IGS could be identified. Conserved regions less than 10 bp apart were merged together to obtain fifty-three conserved regions that represent the potential noncoding functional elements in the human IGS (Figure 2.31; Appendix Table 2). The conserved regions are referred to as ConR-1 to ConR-53 in the order of distance from the last base of the 5' ETS. The conserved regions can be grouped into three clusters: the first between the coding region and the long track of $[TC]_n$ satellites, the second between the $[TC]_n$ tracks and the cdc27 pseudogene, and the third between the cdc27 pseudogene and the end of the IGS. The conserved patterns appearing in the rDNA are outlined below:

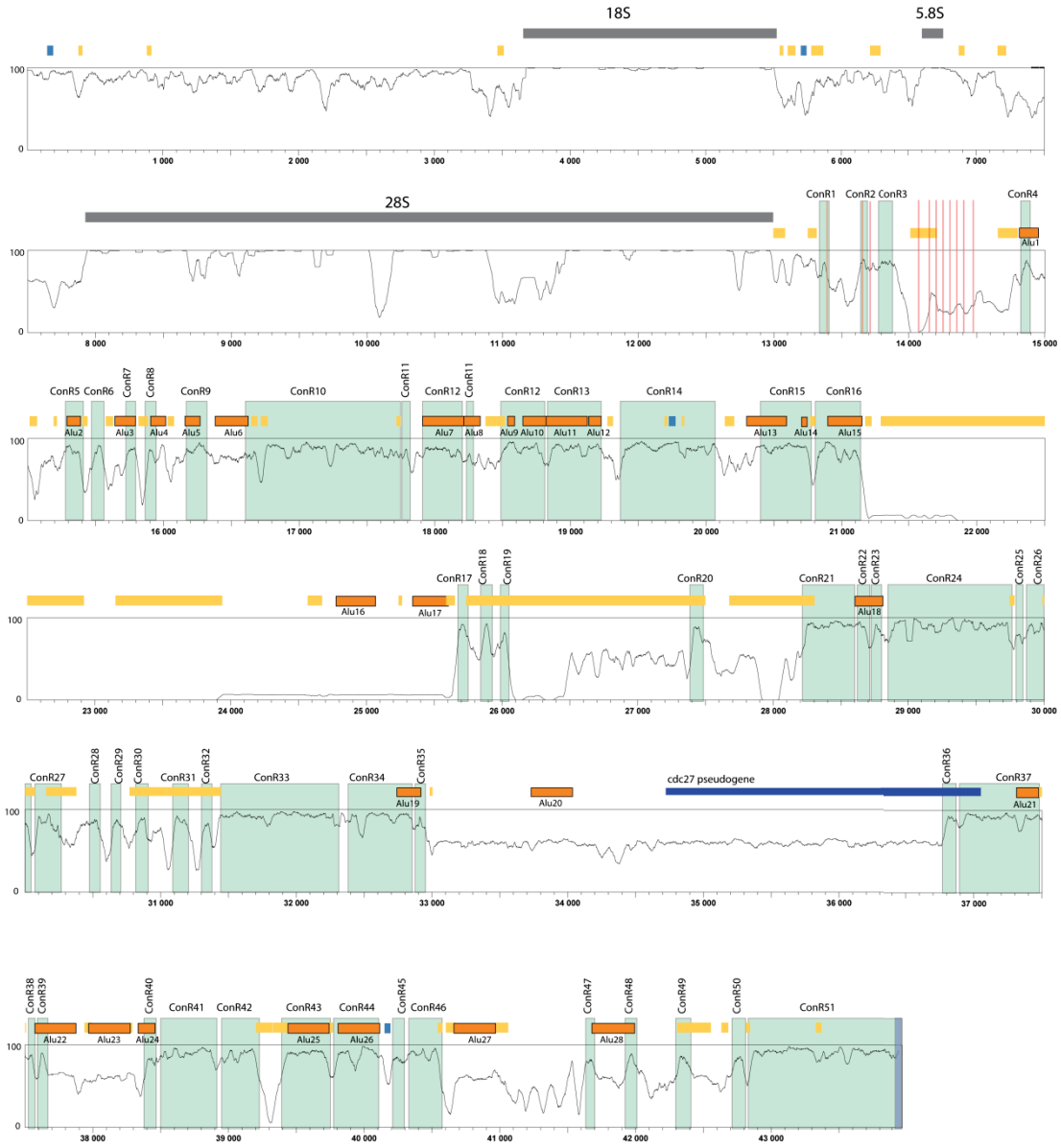


Figure 2.31: Sequence similarity plot for human rDNA with five different primate species *viz.* chimpanzee, gorilla, orangutan, gibbon and macaque.

The horizontal axis represents the position in the human rDNA. The vertical axis represent the level of sequence similarity between 0 (no identity) and 1 (all the bases are same in the column). The conserved regions (green shaded regions) were identified using a 10 bp minimum length and >= 0.8 sequence identity score. The name of the conserved regions is indicated on green box. Annotations of the human rDNA representing different functional elements and repeat elements are mapped to the similarity plot and shown above it. Orange boxes represent Alu elements, yellow boxes represent microsatellites and blue box represents cdc27 pseudogene. The purple vertical bar represents the promoter and the red vertical bars represent terminator elements.

2.3.7.1. Conservation of previously known features in the human IGS:

To verify that the phylogenetic footprinting is capable of identifying functional elements in the human rDNA, I first looked whether known functional elements are present in the conserved regions.

2.3.7.1.1. rRNA coding regions

As anticipated the 18S and 5.8S rDNA are highly conserved across the primates (Hillis and Dixon 1991) (Figure 2.32). For 28S rDNA, conserved region appears as strong peaks and the variable regions as region of relatively low conservation. Similarly as reported previously (Netchvolodov *et al.* 2006), the core promoter element (-45 bp to +18 bp) and the upstream element (-156 bp to -107 bp) for rDNA transcription are highly conserved. Eleven rDNA transcription terminators are present 390 bp downstream of the 28S rRNA coding region (Pfleiderer *et al.* 1990). Of these eleven putative terminators, the first three are conserved while the others are not. It has been reported previously that the termination efficiency of first three terminators is higher than the remaining terminators (Pfleiderer *et al.* 1990), and this is supported by the higher conservation of the first three terminators.

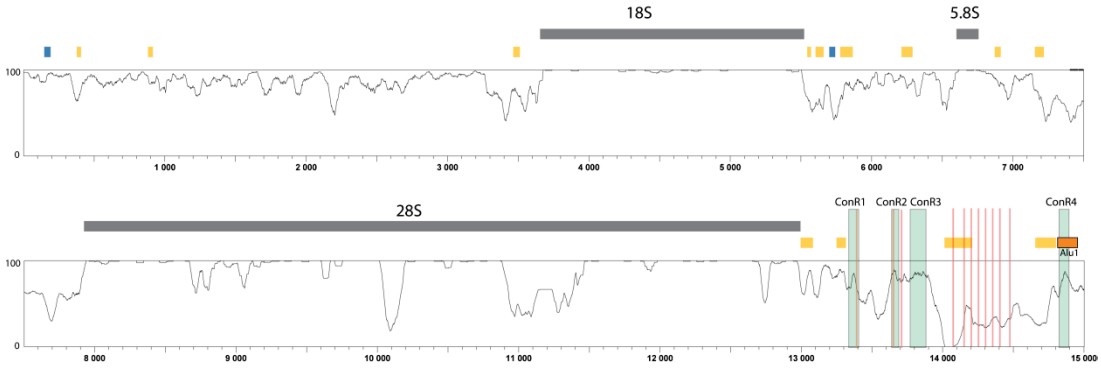


Figure 2.32: Sequence conservation plot for the rRNA coding regions.

Notations are the same as in Figure 2.31.

2.3.7.1.2. c-Myc and p53 binding sites

c-Myc is an oncogene and is associated with the rDNA transcriptional upregulation in many cancerous cells (Grandori *et al.* 2005). c-Myc binding site in the human rDNA is present proximal to the rDNA promoter (Grandori *et al.* 2005) and are conserved among the primates. Another crucial protein known to be associated with the rDNA is p53. p53 is a tumour suppressor gene (Oren 2003) and is involved in the rDNA transcription suppression (Budde and Grummt 1999). A putative binding site of p53 has been reported in the IGS

(Kern *et al.* 1991) and are conserved among the primates. Overall, this suggests that phylogenetic footprinting is capable of identifying protein binding sties in the IGS.

2.3.7.1.3. Noncoding transcripts

Two noncoding transcripts from the rDNA IGS region have been previously characterized, the pRNA (plays a role in rDNA silencing) (Mayer *et al.* 2006) and the IGS_{28}RNA (involved in nucleolar protein sequestration during acidosis) (Audas *et al.* 2012). Interestingly both of the regions encoding these transcripts correspond to conserved regions in the IGS. The pRNA is transcribed from conR-53 while IGS_{28}RNA is transcribed from a region of the IGS that overlaps with conserved regions conR-23 to conR-25 demonstrating that the phylogenetic footprinting capable of identifying noncoding transcripts.

2.3.7.1.4. Alu elements conservation

Alu elements constitute 13.3% of the human IGS. Several studies have shown the conservation of the Alu elements in the IGS. Therefore, next I looked for the conservation of Alu elements. Several Alu elements present in the IGS are highly conserved and correspond to various conserved regions. The high conservation of the Alu elements in the human IGS suggests that they may have some biological role. Further, this also demonstrates that the phylogenetic footprinting is able to identify conservation in the IGS.

2.3.7.1.5. Conservation of cdc27 pseudogene in apes

The phylogenetic footprinting demonstrate that cdc27 pseudogene is conserved in human and apes but is absent in the monkeys (Figure 2.33). This is same as has been reported previously by Gonzalez *et al.* (1993). The average identity of cdc27 pseudogene is 89.2%. Recent studies have shown that the pseudogenes are not inert and have several biological roles (Pink *et al.* 2011; Poliseno 2012; Johnsson *et al.* 2013). High conservation of the cdc27 pseudogene in human and apes suggests that probably they have some biological function.

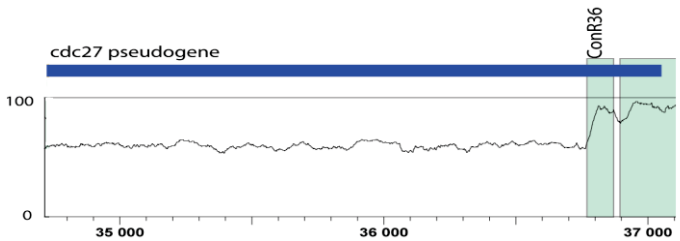


Figure 2.33: Sequence conservation plot for the cdc27 pseudogene.

Notations are the same as in Figure 2.31.

Together, these functional elements account for four of the 53 conserved regions, suggesting that the remaining conserved regions represent novel functional elements.

2.3.8. Search for potential functionality of conserved regions of unknown function:

To proceed further, I focused on characterizing the conserved regions. I decided to cross-reference RNA-seq and ChIP-seq data from the ENCODE project with the conserved regions to obtain some indication on the potential function of these elements (Section 2.1.1). Initially, six cell types were selected for the analysis: three cancerous cell types (A549, K562, and HeLa-S3) and three noncancerous cell types (GM12878, H1-hESC, and HUVEC) (Table 2.3). Hepatocellular carcinoma cell type HepG2 was included in the analysis for searching the long poly(A)+ transcripts from the IGS.

2.3.8.1. Conservation of transcriptionally active regions

To identify the potential transcripts from the IGS, I mapped RNA-seq reads from the ENCODE project to the IGS. Data from ENCODE for cell types GM12878, H1-hESC, HUVEC, K562, and HeLa-S3 were used for the RNA-seq analysis [long poly(A)- RNA-seq data for A549 are not available in the ENCODE database]. To identify the long RNA transcripts from the IGS I performed reference based RNA-seq assembly by mapping the RNA-seq reads using STAR and assembling the mapped reads using Cufflinks for long poly(A)+/- RNA-seq ENCODE data (Section 2.2.1.6.4). The assembly resulted in three long poly(A)- RNA transcripts from GM12878, two from HUVEC, two from HeLa-S3 and seven from K562 (Figure 2.34) while no long poly(A)+ transcript was obtained for the IGS. The transcripts were labelled corresponding to their cell type, followed by a transcript number. The small RNA-seq data from ENCODE for cell lines GM12878, H1-hESC, HUVEC, K562, and HeLa-S3 were mapped to the modified human genome assembly hg19 using bowtie (Section 2.2.1.6.4). The mapped reads of small RNA cannot be assembled because of only initial 36 bp of the RNA transcripts are sequenced. To avoid background noise, signals with a read coverage higher than five were considered as transcript peaks (Figure 2.34). The obtained long poly(A)- transcripts and small RNA signals were then mapped to the IGS to identify which transcripts derive from conserved regions. The transcripts $transcript_{HeLa}$-2 and $transcript_{K562}$-6 corresponding to the conserved region conR-53 are antisense to the coding region of the pRNA. Since these transcripts are present only in the cancerous cell lines but not in the noncancerous cell types included in this study, it is possible, that their transcription is restricted to cancerous cells. To further verify the presence of these transcripts in cancerous cells, I mapped HepG2 long poly(A)- RNA-seq data to the human rDNA. The

assembly gave five long poly(A)- RNA transcripts from HepG2. Similar to $transcript_{HeLa}$-2 and $transcript_{K562}$-6 a transcript named $transcript_{HepG}$-5 obtained from the assembly is antisense to the pRNA coding region. This suggests that this transcript is likely to be cancerous cell specific. Although further, analysis is required to establish that the transcript is cancerous cell specific.

Several other long poly(A)- transcripts that derive from conserved regions conR-34 to conR-52 are present in all the cell types except H1-hESC. These long poly(A)- transcripts overlap each other suggesting that conR-34 to conR-52 are transcribed to form a single long noncoding RNA that is further processed to produce different RNA isoforms in a cell type specific manner. However, it is also possible that each transcript is transcribed independently in a cell type specific manner or a combination of the two, with some transcripts being transcribed independently while others are isoforms of a long noncoding RNA transcript. A strong peak of poly(A)+ small RNA is present in conserved region conR-26 in H1-hESC cells. This peak is not present in any other cell types included in this study, indicating that the corresponding transcript may be a specific feature of embryonic stem cells.

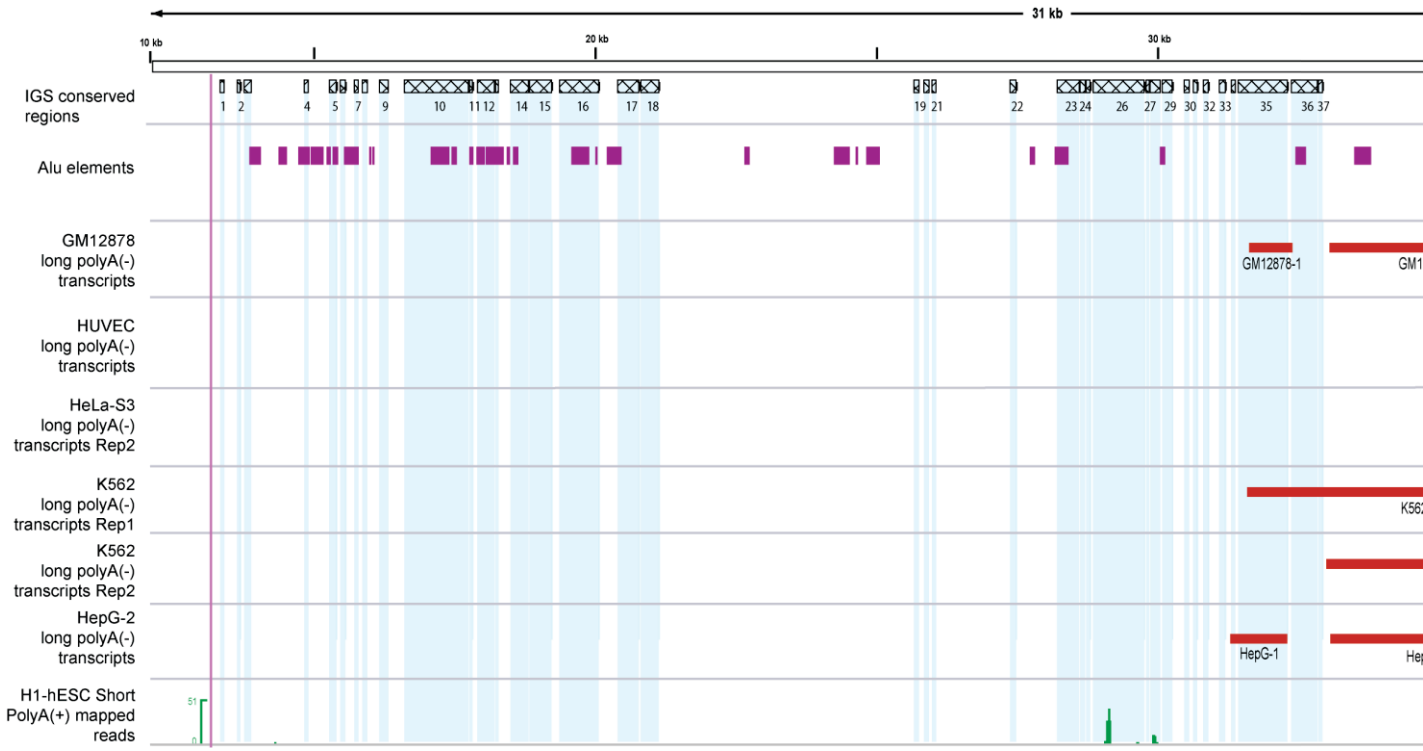


Figure 2.34: The long poly(A)- and small poly(A)+ transcripts in the human IGS from different cell types.

The hatched boxes with the blue shaded regions below represent the conserved regions with their name indicated bel IGS is shown as diagonally shaded region. The purple boxes in the "Alu element" row represent the Alu elements presen element" row represent long poly(A)- transcripts from different cell types. The last row represent small poly(A)+ sig boxes represent long poly(A)- transcripts and the name of the transcripts are indicated below the boxes. The small po peaks with the scale (green vertical line on the left) representing the number reads. The name of the cell type and the lane. The scale above shows the position in the rDNA and the start of the IGS is demarcated by the pink vertical line.

2.3.8.2. Conserved regions as potential transcriptional regulators

The presence of several transcripts that originate from the human IGS implicates that transcriptional regulators (promoters, enhancers and insulators) of these IGS transcripts are present in the IGS and these may lie within the conserved regions. Therefore, to identify potential transcription regulators in human IGS, I mapped ChIP-seq data from ENCODE for various histone modifications and TFs for six cell types GM12878, H1-hESC, HUVEC, K562, A549 and HeLa-S3 and then intersect these with the conserved peaks. The list of factors I mapped is given in Table 2.1 and Table 2.2. All the replicates of selected histone modifications and TFs were mapped to modified human genome assembly hg19 using Bowtie and peaks were called in the mapped data using MACS2 (Section 2.2.1.6). The results of ChIP-seq mapping for embryonic cell H1-hESC is shown in Figure 2.35 while for remaining cell types are shown in Appendix Figure 1 - Appendix Figure 5. The peaks for the histone modifications associated with active transcription are distinct and sharp while those associated with transcriptional repression are comparatively broad and merged together. However, in HeLa-S3 cell type the histone modifications associated with active transcription also have continuous peaks possibly representing a different transcriptional pattern. Based on the positions of peaks for active histone modifications in difference cell types (excluding HeLa-S3), the conserved regions were divided into two groups: 1) conserved regions that have peaks of active histone modifications and 2) conserved regions with no peaks. Peaks for active histone modifications span more than one conserved region regions and can be group into three clusters depending on enrichment for histone modification and TFs as described below:

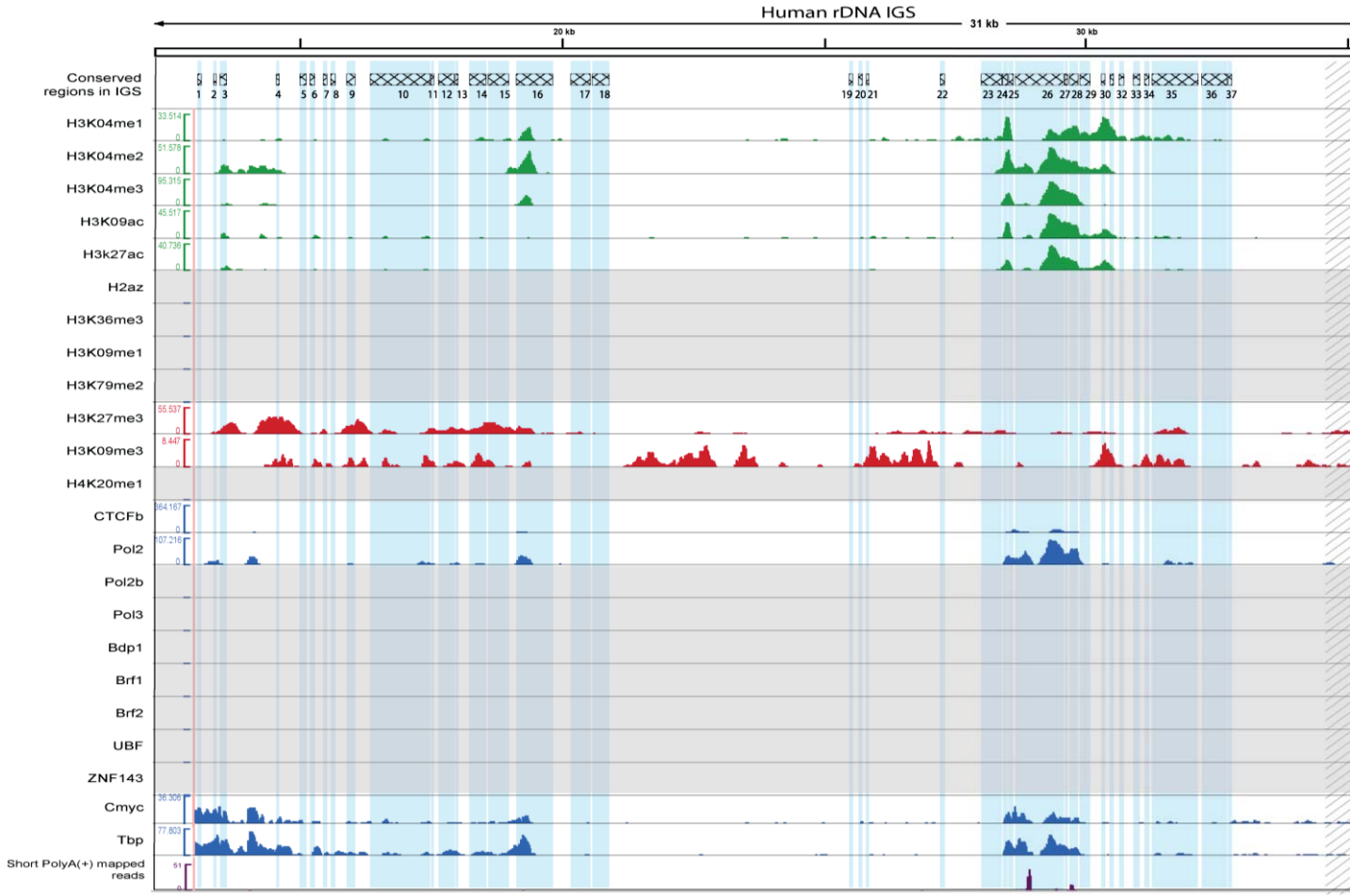


Figure 2.35: Chromatin, transcription factor and transcript landscape of the IGS in embryonic cells H1-h

The hatched boxes with the blue shaded regions below represent the conserved regions with their name indicated bel IGS is shown as diagonally shaded region. Each row represents the enrichment for an active histone modifications (gr (red signals), transcription factor (TF; blue signals) or small poly(A)+ signal. The name of the histone modifications an left side of each row represents the level of enrichment. The scale above shows the position in the rDNA and the start line. The grey rows represents the absence of the data for the histone modification or TF in the ENCODE project for the

Cluster-1: This consists of conserved region conR-53. The cluster corresponds to the region that transcribes pRNA and the transcripts antisense to the pRNA coding region that are found in the cancerous cell types $transcript_{HeLa}$-2, $transcript_{K562}$-6 and $transcript_{HepG}$-5. The conR-53 region shows enrichment for histone modifications that are associated with active transcription. This includes peaks of H3K4me2/3 (associated with promoters), H3K9ac (transcription initiation), and H3K27ac and H2A.Z (open chromatin) (Figure 2.36). Further, this region also shows enrichment for TFs Pol II, UBF, ZNF143, c-Myc, and TBP and the pattern differ among cell types. ZNF143 is known to be associated with both the Pol II and Pol III machineries (Schaub *et al.* 1997; Schuster *et al.* 1998). Since in cluster-1 there is a peak for Pol II but not Pol III, it is likely that the ZNF143 in this region is associating with Pol II. All the cell types, except GM12878, have enrichment of transcription factor CTCF. It has been previously reported that H3K4me2/3, H2A.Z and CTCF are enriched in the region and are associated with the pRNA transcription (van de Nobelen *et al.* 2010). The identification of a previously described chromatin state in cluster-1 demonstrates that this approach is capable of identifying transcriptional regulators in the human IGS. Further, peaks of certain histone modification and TFs show two separate or fused peaks in conR-53 indicating that region may have closely placed transcription regulators.

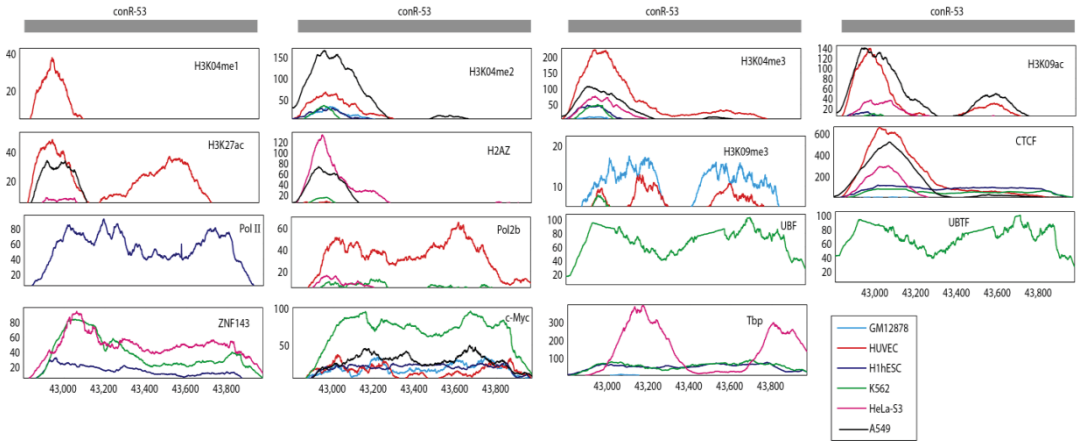


Figure 2.36: Chromatin marks and TFs associated with conR-53.

Chromatin marks and TFs that are present in conR-53 from GM12878, HUVEC, H1-hESC, A549, K562 and Hela-S3 cell types are shown, with each mark/factor in a separate panel. Grey bars at the top of show the location of the conserved region. The horizontal axis shows the positions (in bp) in the rDNA and the vertical axis shows the normalized signal level for the chromatin mark.

Cluster-2: This consists of conserved regions conR-23 to conR-31**.** This region shows enrichment for chromatin marks and TFs for all cell types except the lymphoblastoid cell type GM12878 (Figure 2.37). The specific enriched histone modifications and TFs are: H3K4me1/2/3, H3K09ac and H3K27ac (transcription initiation and open chromatin), CTCF (insulator/promoter protein), and Pol-II, UBF, ZNF143, c-Myc and TBP. A genome wide mapping study has shown that H3K4me1 is relatively more enriched than H3K4me2/3 at enhancers, while H3K4me3/2 is relatively more enriched than H3K4me1 at promoter (Barski *et al.* 2007). Different cell types show different ratios of H3K4me1/2/3 suggesting that this region may be acting as an enhancer or promoter depending on cell type. Although all cell types show enrichment of factors in this cluster, the most significant peaks are for the embryonic cell type H1-hESC and the leukaemia cell type K562. H1-hESC has higher enrichment of H3K4me3 compared to H3K4me1 together with open chromatin histone modifications (H3K9ac and H3K27ac), indicating that region may be acting as a promoter in this cell type. Further, H1-hESC is also enriched for Pol II and polymerase cofactor TBP, suggesting that the region is potentially transcribed. Strikingly, the RNA-seq analysis revealed a strong peak of short poly(A)+ transcripts in H1-hESC cells at conR-26 that coincide with the enrichment of factors in this region. Together these results suggest that this region is transcriptionally active, conserved, and being driven by an active chromatin structure around these regions. The alveolar carcinoma cell type A549 shows a higher enrichment of H3K4me2/1 than H3K4me3 as well as with open chromatin histone modifications (H3K9ac and H2A.Z). These features suggest that this region may be acting as enhancer in this cell type.

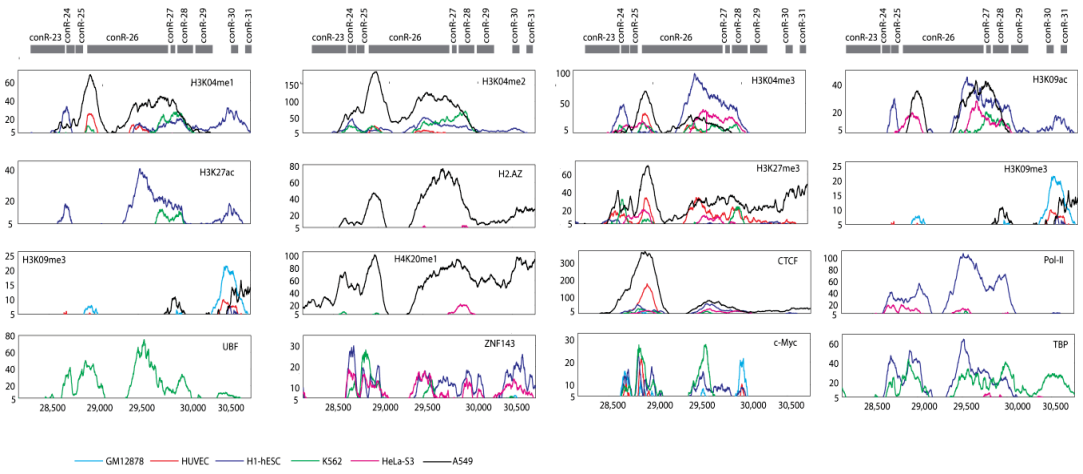


Figure 2.37: Chromatin marks associated with conR-23 to conR-31.

Chromatin marks and TFs that are present in conR-23 to conR-31 from HUVEC, H1-hESC, A549, K562 and Hela-S3 cell types are shown, with each mark/factor in a separate panel. Grey bars at the top of show the location of the conserved region. The horizontal axis shows the positions (in bp) in the rDNA and the vertical axis shows the normalized signal level for the chromatin mark.

Cluster-3: This consists of conserved region conR-16. In the embryonic stem cell type H1-hESC, this cluster has a higher enrichment of H3K4me3/2 than H3K4me1 (Figure 2.38). Genome wide mapping studies have shown that higher enrichment of H3K4me3/2 over H3K4me1 is found around promoters, suggesting that cluster 3 may acting as poised promoter in the H1-hESC cell type. Further, this cluster shows enrichment of factors associated with the polymerase machinery i.e. TBP, Pol II and ZNF143, as well as CTCF, providing further evidence that it may be behaving as a promoter. In alveolar cancerous cell A549, the cluster has enrichment of H3K4me2 than H3K4me1/3, suggesting that cluster-3 may act as silent promoter in the A549 cell type. Interestingly however, no transcript was found in the region around cluster-3 in the RNA-seq assembly, therefore the role of this region remains unclear.

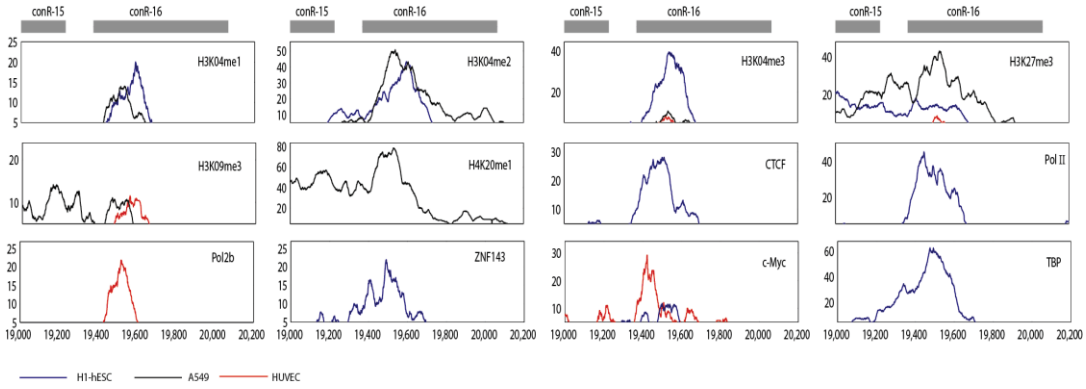


Figure 2.38: Chromatin marks associated and TFs with conR-16.

Chromatin marks and TFs that are present in conR-16 from HUVEC, H1-hESC and A549 cell types are shown, with each mark/factor in a separate panel. Grey bars at the top of show the location of the conserved region. The horizontal axis shows the positions (in bp) in the rDNA and the vertical axis shows the normalized signal level for the chromatin mark.

2.3.8.3. Origin of replication

The presence of origin of replication in the human IGS has been reported by various studies (Coffman *et al.* 1993; Little *et al.* 1993; Gencheva and Russev 1996; Coffman *et al.* 2006; Dimitrova 2011). However, the position of origin of replication was found variable in these studies and therefore the exact position of origin of replication in the IGS is not known. To identify the position of the potential origin of replication in the IGS, I mapped ORC (origin replication complex) ChIP-seq data for the HeLa-S3 cell type to the rDNA and looked for overlap in the conserved regions. The data used for the analysis are from the study by

Dellino *et al.* (Dellino *et al.* 2013). The data show two prominent peaks: one spans conR-1, 2 and 3 while the other is in conR-53 (Figure 2.39). Both peaks are consistent with the finding of Gencheva *et al.* (Gencheva and Russev 1996) where they have reported the presence of an origin of replication near to the rDNA promoter, which corresponds to conR-53 and another in the region next to the terminator which corresponds conR-1 to conR-3. Further, a small third peak for ORC around conR-24 and conR-25 is also present and a potential candidate of an origin of replication.

Figure 2.39: Origin replication complex (ORC) binding in the HeLa-S3 cell type.

The hatched boxes with the blue shaded regions below represent the conserved regions with represents the enrichment for ORC in Hela-S3 cells. The scale on left of the row represents th shows the position in the rDNA and the start of the IGS is demarcated by the pink vertical line.

2.3.9. Transcription machinery associated with the rDNA

Although the rDNA is transcribed by Pol I, an increase in the number of transcriptionally active rDNA copies can lead to Pol II transcription in the human rDNA (Gagnon-Kugler *et al.* 2009). Since cancerous cells are marked by increases in the number of active rDNA units (Maggi and Weber 2005; Bywater *et al.* 2012) it is possible that other polymerase machineries may also be transcribing the rDNA. To search whether Pol II and Pol III also play a role in the transcription of the human IGS, I performed ChIP-seq mapping using data for Pol II, Pol III and associated cofactors. ChIP-seq mapping of Pol II shows distinctive peaks in the different cell types. In the embryonic cell type H1-hESC, the Pol II is associated with conserved regions conR-16, conR-53, and conR 23-31, while in epithelial HUVEC and leukaemia K562 cells, and the Pol II peaks are restricted to conR-53 (Figure 2.36). In HeLa-S3, Pol II has closely placed peaks that overlap to several conserved regions of the IGS. The ChIP-seq data for Pol III together with associated cofactors Bdp1, Brf1 and Brf2 were mapped to the rDNA. Peaks for Pol III are restricted to the cdc27 pseudogene and the regions that contains Alu20, that is present upstream of the cdc27 pseudogene (Figure 2.40)

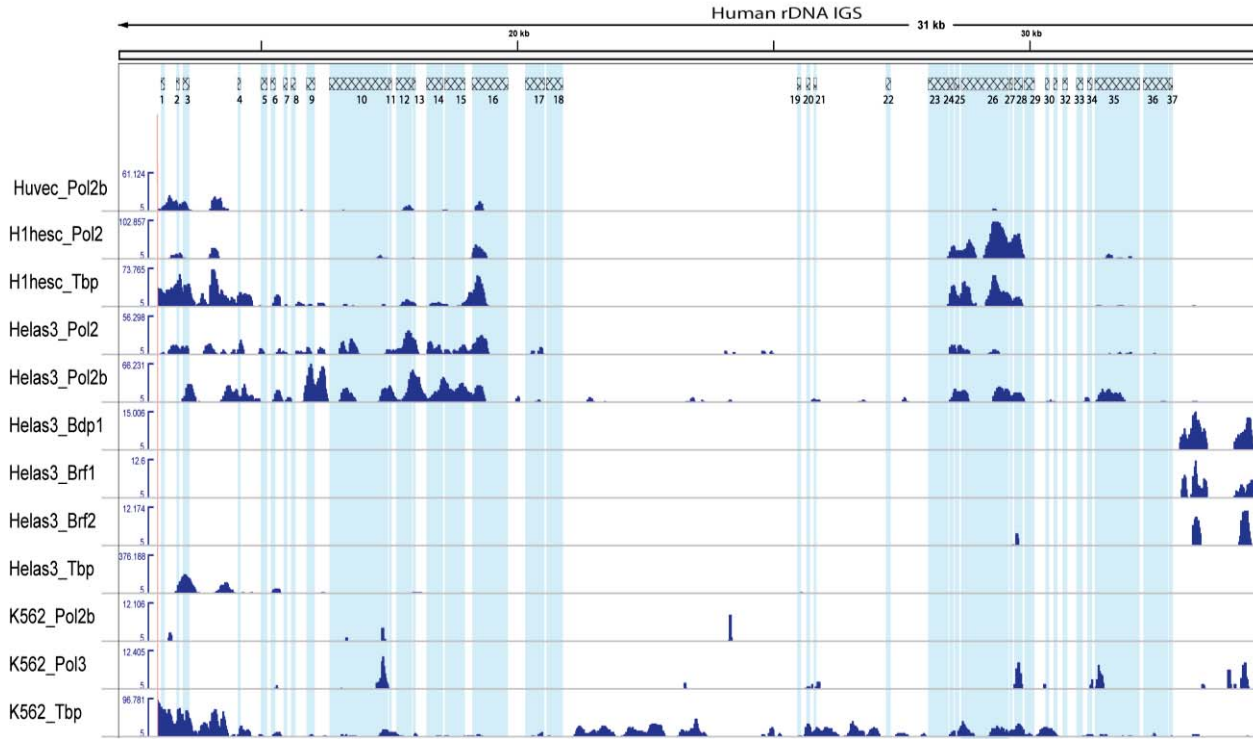


Figure 2.40: Pol machineries and related transcription factors that associate with the human

The hatched boxes with the blue shaded regions below represent the conserved regions w position of cdc27 pseudogene in the IGS is shown as diagonally shaded region. Each row rep polymerase machinery component or associated transcription factor. The name of the ce transcription factor is indicated to the left. The scale on left of each row represents the level the position in the rDNA and the start of the IGS is demarcated by the pink vertical line.

2.4. Discussion

The highly repetitive structure of the human rDNA imposes a challenge to demarcate the potential functional elements in the IGS. In this work, I have combined phylogenetic footprinting with RNA-seq assembly and ChIP-seq data for chromatin markers to identify the potential functional elements in the human IGS. To initiate the analysis I constructed six primate rDNA sequences using WGA and compared them with the human rDNA to identify conserved regions in the human IGS. This analysis has provided a comprehensive dataset for potential candidates of the functional elements in the human IGS. The analysis identified the known functional elements, including the rDNA promoter, terminator, and protein binding sites as highly conserved regions (Section 1.4.3) demonstrating that the technique is capable of identifying functional elements. Overall 53 conserved regions (named as conRs; Section 2.3.7) were identified in the IGS. The RNA-seq assembly identified several transcripts that overlap with the conserved regions (Section 2.3.8.1). Since, chromatin structure helps to determine the functional characteristics of a genomic region, potential functions of the conserved regions were searched for by mapping publically available, ENCODE ChIP-seq data for a variety of histone modifications and TFs to the human rDNA. This identified three clusters that act as potential transcription regulators in the human IGS (Section 2.3.8.2). The pairwise comparison between the Alu elements of human and primates together with phylogenetic footprinting analysis also demonstrated high conservation of Alu elements in the IGS indicating they may have a functional role. Overall, this study demonstrates that the human IGS contains several potential functional elements but further experiments are required to characterize them.

Potential transcripts and transcription regulatory elements in the human IGS

Integrating the transcripts identified from RNA-seq analysis and the enrichment of histone modifications and transcription factors with the level of sequence conservation revealed two potential functional hotspots in the human IGS. The chromatin profiling revealed three regions enriched in active histone modifications in the IGS corresponding to a region near to the promoter (Cluster-1) and a second region in the middle of the IGS (Cluster-2), together with a cell type specific enrichment in H1-hESC (Cluster-3). These results are the same as reported by Zentner *et al.* (2011) using four cell types (H1-hESC, HUVEC, K562 and NHEK) despite the difference in experimental protocols. Zentner *et al.* (2011) mapped the reads without trimming the low quality ends and used the tool F-seq (Boyle *et al.* 2008) to call the peaks, while I trimmed the low quality reads, filtered the short reads, and used the tool MACS2 (Zhang *et al.* 2008) to call the peaks. The reproducibility of results despite the

differences in the strategy used to prepare the reads for aligning and the underlying algorithm of the tool used to call the peaks between the two studies suggests that the enrichments reported here are rigorous and are not an artefact of the tools used. The first functional hub is just upstream of the rDNA promoter and is represented by conR-53. The most striking feature of this hotspot is the presence of a transcript in all cancerous cell types included in the study. This transcript is antisense to the coding region of the pRNA transcript. The pRNA transcript plays a crucial role in the silencing of rDNA units by recruiting the nucleolar remodelling complex (NoRC) to the promoter region (Mayer *et al.* 2006; Mayer *et al.* 2008). In many cancerous cells, the demand for protein and consequently ribosomes is high as a result of increased proliferation (Hadjiolov 1985; Montanaro *et al.* 2008). To meet the high demand for ribosomes the number of active rDNA units increases to produce more rRNA. The role of the pRNA in repressing rDNA transcription suggests that if it is prevented from recruiting NoRC, the number of active rDNA units will increase. Recent studies have revealed the role of antisense transcripts of several genes in regulating the level of expression by sense-antisense RNA pairing (Yelin *et al.* 2003; Katayama *et al.* 2005a; Faghihi and Wahlestedt 2009; Werner 2013). Therefore, I propose that the antisense transcript I identified here forms an RNA-RNA duplex with the pRNA, and this prevents pRNA from recruiting NoRC to the rDNA. The outcome of such a system would be an increase in the number of active rDNA copies and this is consistent with my observation that this antisense transcript is only present in the cancer cell types assayed here. In support of this, Hwang *et al.* (2011) have shown that production of a transcript that is antisense to the coding region and the IGS across the promoter region increases the 47S rRNA concentration in lung cancer cells. Therefore, there may be a system of sense-antisense noncoding RNAs, including the pRNA/antisense transcript described in this thesis that control the transcriptional activity of the rDNA.

The second functional hotspot in the IGS is represented by cluster-2 (conR-26 to conR-29). The most interesting feature associated with this region is the presence of a novel small poly(A)+ transcript in the H1-hESC cells that is absent in other cell types. While the function of this transcript in human is unknown it is interesting to note that in mouse 6.76% of rDNA units are methylated in embryonic stem cells compared to 22.57% in embryonic fibroblasts (Zheng *et al.* 2012). Therefore, I suggest that one possible function of the small poly(A)+ transcript from the region in H1-hESC could be prevention of methylation of rDNA units. The mechanism by which this transcript may prevent methylation of rDNA units in embryonic cells might be similar to the way that the pRNA promotes acetylation of promoter regions by facilitating binding of NoRC. Interestingly, in contrast to all the other cell types included in this study, H1-hESC does not have any long poly(A)- transcripts from the IGS.

This cell line does have a small poly(A)+ transcript from cluster-2 specific to it. The presence of a small poly(A)+ transcript but absence of the long poly(A)- transcripts from the IGS in H1-hESC suggests that expression of the identified small poly(A)+ and the long poly(A)- transcripts may have a reciprocal relationship. Further, less H3K4me3 in other cell types compared to H1-hESC suggests that cluster-2 may be epigenetically suppressed in different cell types and probably does not transcribe this small poly(A)+ transcript following embryonic stem cell differentiation .

2.4.2. Cdc27 pseudogene as potential regulator

The cdc27 pseudogene present in the human IGS is highly conserved among the apes (section 2.3.7.1.5 and Figure 2.31) suggesting it may have a functional role. The CDC27 gene itself encodes a subunit of the anaphase-promoting complex (APC) (Tugendreich *et al.* 1993). The APC is well studied in yeast and is highly conserved across the eukaryotes. It is involved in three different stages of the cell cycle: a) it facilitates the separation of sister chromatids at the metaphase-anaphase transcription; b) it promotes the exit from mitosis at the end of anaphase; and c) it prevents premature entry into S-phase by regulating the proteins involved in the progression of the cell cycle (Thornton *et al.* 2006). Cdc27 is involved in these different stages of APC regulation by promoting interactions between the other APC components (Tugendreich *et al.* 1995).

For long time it has been thought that pseudogenes are inactive, and they have been considered to be a source of transcriptional noise (Li *et al.* 1981; Harrison *et al.* 2005). However, recent studies have shown that pseudogenes are not necessarily "junk DNA" but can play regulatory roles (Pink *et al.* 2011; Poliseno 2012). For example, transcripts from the pseudogenes of several genes, including PTEN, OCT4, NANOG and Jun, act as transcriptional and post-transcriptional regulators of their parent genes (Pain *et al.* 2005; Poliseno *et al.* 2011; Poliseno 2012). Furthermore, pseudogenes are also known to have isoforms that have different functionalities (Johnsson *et al.* 2013). The RNA-seq assembly I performed identified transcripts from the cdc27 pseudogene region in GM12878, HUVEC, K562 and HepG2. The function of this transcript is unknown. I propose that the rDNA cdc27 pseudogene transcripts may regulate the level of cdc27 protein and thus alter the behaviour of the APC. This assumption is based on the role of pseudogene transcripts in regulating other genes, and on the conservation of cdc27 pseudogene in the primate rDNA. The transcripts identified from the cdc27 pseudogene in the RNA-seq assemblies differ between cell types, therefore these isoforms may have cell type specific functions in regulating the APC. However, experimental testing is required to establish the role of the rDNA cdc27 pseudogene transcript. A RT-PCR based heteroduplex screening assay identified a transcript

from the cdc27 pseudogene to be present in colon cancer cells (Wang *et al.* 2003). One possibility is that this transcript and the one found in this study from the RNA-seq assembly that corresponds to the IGS cdc27 pseudogene is the same. Since, the sequence of the transcript reported by Wang *et al.* was not reported, it cannot be ruled out that this transcript is from one of the two other cdc27 pseudogenes that are present on chr 2 and chr Y. The length of cdc27 pseudogene in the human rDNA is 1,951 bp. In comparison, the pseudogene on chr 2 is 774 bp and is 90% identical to the rDNA pseudogene, while the chr Y cdc27 pseudogene is 1,921 bp and is 88% identical to the rDNA pseudogene. This level of sequence identity means it is possible to distinguish among potential transcripts from these three different regions (rDNA, chr 2 and chr Y). Considering the stringent mapping parameters used in this study for the RNA-seq assembly, it is unlikely that the reported transcript is from the cdc27 pseudogenes on chr 2 and chr Y.

2.4.3. RNA Polymerase II and III machineries are associated with the human IGS

ChIP-seq mapping of Pol II in this study shows distinct peaks of Pol II enrichment in the IGS of embryonic H1-hESC cell and umbilical vein endothelial HUVEC cell while HeLa-S3 have closely placed peaks of Pol II that cover a large portion of the IGS. Together, these results indicate that Pol II may also be involved in the transcription of regions in the IGS, and that the activity of Pol II varies between different cell types. It has been previously reported by Bierhoff *et al.* (2010) that Pol II transcribes a region antisense to the mouse rDNA coding region, extending ~4 kb into the IGS in the rDNA promoter region. The presence of this antisense transcript was also reported in human Hela-S3 and HaCat cell types suggesting that Pol II potentially transcribes an antisense rDNA transcript in human as well (Bierhoff *et al.* 2010). The identification of enrichment of the Pol II in the IGS strongly suggests that Pol II may also transcribe certain regions of the human IGS.

The human IGS contains a number of Alu elements and some of them are highly conserved. Since Alu elements are transcribed by Pol III (Deininger 2011), I mapped the data for Pol III and associated cofactors (Bdp1, Brf1 and Brf2) from Hela-S3 and leukaemia K562 cells to the rDNA. The peaks for enrichment of Pol III and associated factors in the IGS are restricted to the cdc27 pseudogene and Alu20 (located just upstream of the cdc27 pseudogene), suggesting there is region-specific activity of Pol III. However, the enrichment for Pol III is very weak suggesting its contribution to transcribing the human IGS is minor. Previously it has been reported that in budding yeast all three Pol machineries transcribe different regions of the rDNA. Pol I and Pol III have their canonical roles of transcribing 35S and 5S rRNA respectively (Kassavetis *et al.* 1990; Paule and White 2000), while Pol II

transcribes various noncoding transcripts in the IGS (Houseley *et al.* 2007; Mayán 2013). Compared to budding yeast, the roles of the three Pol machineries in the human IGS have not been explored in detail. The ChIP-seq mapping results presented in this chapter reveals that not just Pol I but also the other two RNA polymerase machineries may be associated with the human IGS. However, it is not possible with these ChIP-seq data to determine if Pol II and Pol III are associated with human rDNA units that are localized in the nucleolus or outside the nucleolus in inactive NORs.

2.4.4. Potential origin of replication in IGS

Several organisms, including budding yeast, fission yeast and mouse, have had a defined origin of replication identified in their IGS (Linskens and Huberman 1988; Gögel *et al.* 1996; Sanchez *et al.* 1998). However, for humans the position of the rDNA origin of replication is not well established. Different studies have shown various positions, from the promoter region to the entire rDNA, as potential sites of origin of replication activity in the human rDNA. The inter-origin spacing in the human genome is ~30 kb (Guilbaud *et al.* 2011) and since the human rDNA repeat units are ~43 kb in length, we might expect each rDNA unit to contain an origin of replication. The demarcation of origin of replications in human is difficult because human origins lack a fixed sequence pattern for origin of replication complex binding. ChIP-seq data for ORC from HeLa-S3 cells show peaks around conR-53 (near to the promoter) and around conR-1, 2 and 3 (near to the terminator) (Figure 2.39) which suggest that these regions might be acting as an origin. Conserved region conR-53 corresponds to the region transcribing pRNA and antisense transcripts. A recent genome wide mapping study of ChIP-seq data for ORC from Hela-S3 cells has shown that in human, regions where ORC binding coincides with transcriptionally active regions are more likely to fire as an origin (Dellino *et al.* 2013). Additionally an association between Pol II and ORC has been found in budding yeast, suggesting that Pol II may help to recruit ORC (Mayan 2013). All together i.e. the presence of Pol II along with the transcription of pRNA and antisense from the conR-53 region supports that the region may acting as an rDNA origin of replication. Overall, this supports the results of Gencheva *et al.* (1996) who found that the rDNA origin of replication is located in the IGS near to the promoter and terminator.

2.4.5. CTCF association is not restricted near to the rDNA promoter but also present in the other regions of the IGS

I also searched for association of CTCF in the IGS. The results show enrichment of CTCF at the promoter region as previously reported (van de Nobelen *et al.* 2010; Zentner and Scacheri 2012). However, I also identified three new regions of enrichment that correspond

to cluster-2, cluster-3 and a region around ~38.5 kb - ~39.5 kb in the IGS. The CTCF association with IGS shows variation among the cell types (Figure 2.35; Figure 2.11-Figure 2.13). This suggests that CTCF plays a dynamic role depending on the cellular conditions. The presence of CTCF in other regions of the IGS suggests that its role is not restricted to rRNA transcriptional regulation, and therefore I proposed that the CTCF has a broader role in the regulation of IGS transcription as well.

2.4.6. Conservation of Alu elements in the primate rDNA

The pairwise comparison between the Alu elements (Section 2.3.6.4) in the IGS shows that some of the Alu elements are more conserved then the others in the primates. The high conservation of some Alu elements suggests these Alus may have some function. It is unlikely that this Alu conservation results from their reverse transcription activity, as this should apply to all the Alus in the IGS. Table 2.17 and Figure 2.31 show that the level of conservation is not the same across all Alus present in the human IGS, with only a few Alus being conserved across the apes and monkeys. This suggests that these conserved Alus are likely to be subject to selective constraint.

The role of these highly conserved Alu elements is elusive. Studies have shown that more Alu elements are enriched in alternatively spliced internal exons from the human genome (Sorek *et al.* 2002). Therefore, one possible role of the Alu elements could be alternative splicing of IGS transcripts. The RNA-seq assembly identified transcripts from several cell types that span several Alu elements (Figure 2.34). Some of these transcripts overlap each other between the cell types, suggesting that these transcripts may be alternatively spliced, depending on the cellular conditions. The presence of Alu elements in these transcripts is consistent with the involvement of these Alu elements in the alternative splicing of the IGS long poly(A)- transcripts. A possible example of an alternatively spliced transcript that spans a conserved Alu is the three transcripts from the leukaemia cell type K562 ($transcript_{K562}$-3, $transcript_{K562}$-4 and $transcript_{K562}$-5) and a transcript from the endothelial cell type HUVEC ($transcript_{HUVEC}$-2) that overlap each other (Figure 2.34). Since these transcripts are from the same region, it is possible that they are alternatively spliced, and interestingly conserved Alu elements are present on the 5' end of both $transcript_{K562}$-4 and $transcript_{K562}$-5.

2.4.7. The limitations of ENCODE data to predict the function of the rDNA IGS

The ENCODE project has provided a massive set of genomic resources that includes ChIP-seq data for various chromatin modifications and transcription factors, and transcriptome data for various cell types. The data is publically accessible and provides an opportunity to identify potential transcripts and transcription regulators from unexplored regions of in the human genome. Since the project includes data for several chromatin modifications and transcription factors, an integrative study can be perform to study the association of the various factors in any region of the human genome. By using various cell lines representing different tissues it is possible to study the variation in the association of different epigenetic factors among different tissues. Further, the data also provide the opportunity to identify novel cancer cell markers by performing comparative analysis between cancerous and non-cancerous cell types.

However, care must be taken when attempting to interpret biological function from ENCODE project data. One of the biggest drawbacks of the ENCODE data is that they are from primary or immortalized cell types that are cell types from the tissue samples from different individuals that have been modified to keep them viable in laboratory conditions, including growth in culture media, that do not mimic the cellular conditions inside a human body. The association of different regulatory epigenetic factors to a genomic region varies depending on the cellular and other environmental conditions (Koch *et al.* 2007; Dunham *et al.* 2012). Therefore it is possible that transcripts and the epigenetic factors corresponding to a genomic region that were identified using ENCODE data is an effect of the laboratory conditions. Of relevance here, the activity of the rDNA varies depending upon the cellular conditions (Mayer *et al.* 2005; Drygin *et al.* 2010), therefore some of the IGS transcripts and transcriptional regulators identified in this study may be artefacts of the laboratory conditions. To rule out the influence of the laboratory conditions, further *in vivo* investigations will be required to verify the identified transcripts and regulators.

The results characterizing the human rDNA IGS using ENCODE data described in Sections 2.3.8.1 and Section 2.3.8.2 have limitations because of the presence of multiple copies of the rDNA units. Based on their transcriptional activity rDNA units can divided into two states: active and inactive (Section 1.5). Further, active rDNA units can be either transcriptionally active or poised to be active (Hamperl *et al.* 2013). The sequences of the rDNA units are nearly identical to each other and therefore it is not possible differentiate signals between different rDNA units. Thus, the chromatin and transcriptional landscape of the human IGS described in this study is the consensus of all the active and inactive rDNA units. Since it is

not possible to differentiate these different states based on sequences, it is difficult to correlate between peaks for epigenetic factors associated with different transcription states that coincide at the same position in the IGS.

In addition, it is difficult to estimate the signal-to-noise ratio for ChIP-seq data where a histone modification or transcription binding factor is associated with only a few rDNA units and therefore the signal peak is very small. Thus, ENCODE analysis gives an indication of the probable function of a region in the IGS and more sophisticated experiments that can differentiate rDNA repeats will be required to assign function to the region.

2.4.8. Comparison between the human and the yeast IGS phylogenetic footprinting analysis

Similar to the human IGS study, phylogenetic footprinting analysis has been reported previously for 6 *Saccharomyces* species (Ganley *et al.* 2005). Similar to the human IGS several conserved regions representing potential functional elements are present in the *Saccharomyces* IGS. Although, sequence identity between human and yeast is negligible yet, some of the identified conserved regions in yeast and human have similar function. Similar to the IGS of *Saccharomyces* species the rRNA promoter is highly conserved among the primates. Another functional region conserved in the IGS of *Saccharomyces* species is the origin of replication. The potential origin of replication sites identified in the human IGS using ORC ChIP-seq data are also conserved among the primates, although the function of these sites still needs to be verified. This demonstrates that despite of no sequence conservation the function associated with the IGS remain conserved between yeast and human. It is highly probable that other identified potential functional elements in the yeast and human have similar function.

2.4.9. Correlation between the increase size of the IGS and evolution of amniotes

In human and other amniotes the nucleolus is tripartite (Lamaye *et al.* 2011). Lamaye *et al.* (2011) have shown that the transition from bipartite to tripartite nucleolus occurred between the turtles and lizards, and coincides with the emergence of amniotes. The factors facilitating this transition are not known. The IGS size in the non-amniotes is ~2-3kb which is thought to be increases up to ~30kb in amniotes (Thiry and Lafontaine 2005). Thiry and Lafontaine (2005) have proposed the hypothesis that the transition from the bipartite to tripartite nucleolus is related to the increase in the IGS size. To test this hypothesis it is important to know when the increase in size of the IGS occurred and if the increase was gradual or there

was a sudden jump in the IGS size. However, because of the absence of the complete rDNA sequences in the database for most of the organisms including turtles and lizards it is not possible to verify this assumption. Complete rDNA sequences of different organisms including turtles and lizards to identify will be required to study the change in the size of the IGS during the evolution. The phylogenetic footprinting study of the human IGS demonstrates that several potential functional regions are present in the region. Further, integration of ENCODE analysis provides evidence that some of the identified functional regions are potentially transcribed. The function of identified novel transcripts from the human IGS is not known. It is possible that the identified transcripts are amniotes specific and probably have role in the development of amniotes-specific features. However, in the absence of sequence it is difficult to verify this hypothesis.

The integration of the evidence from different sources has identified a variety of potentially functional elements in the human IGS, including RNA transcripts, potential non-coding RNA promoters and an origin of replication. It also identified two potential functional hubs in the IGS as well as cell type specific functional regions. Further investigation will be required to delineate the functions of these regions. In concise, this study provides a comprehensive dataset of the potential functional elements in the human IGS.

[Blank page]

Chapter 3

Characterization of the regions surrounding the human rDNA array: The human rDNA flanking regions

[Blank page]

3.1. Introduction

Completion of the human genome project is one of the major achievements of the biological sciences. It has provided the reference sequence of the human genome that has facilitated to a better understanding of human biology. However, despite the much publicized fact that the human genome sequence is complete it contains several significant gaps that still need to be filled (Eichler *et al.* 2004). One of the most prominent gaps is the absence of the short arms of all five acrocentric chromosomes (chr 13, chr 14, chr 15, chr 21 and chr 22). In human, the short arms of all the acrocentric chromosomes contain nucleolar organizer regions (NORs) (Henderson *et al.* 1976). Each NOR consists of the rDNA array and the region surrounding the rDNA array, which we term the rDNA flanking regions. The NOR is the site of nucleolus formation, and the nucleolus is the site of ribosome biogenesis, which plays a critical role in cell survival. After the end of cell division, small nucleoli are formed around the NORs, which are later fused to form a nucleolus. This process is known as nucleolar fusion. The number of nucleoli in a cell usually varies between one and three (Anastassova-Kristeva 1977). The function of the nucleolus as a ribosomal factory has been studied in detail. However, in contrast far less is known about the underlying genomic sequence around which the nucleolus is formed.

3.1.1. The rDNA flanking regions sequences

The sequences of the rDNA flanking regions have been identified in the lab of Prof. Brian McStay (NUI, Galway, Ireland). To do this, they identified cosmids and BAC clones from the rDNA flanking regions by screening chromosome specific cosmid libraries and GenBank using the 8.3 kb distal flanking sequence identified by Gonzalez and Sylvester (Gonzalez and Sylvester 1997) and the 493 bp proximal flanking sequence identified by Sakai *et al.* (1995) (Section 1.6). They identified three cosmids (LA13 133H12 [chr 13], LA14 138F10 [chr 14] and LA15 64C10 [chr 15]) from the distal flanking region and four cosmids (LA13 165F6 [chr 13], LA14 101B3 [chr 14], LA15 25H3 [chr 15] and N 29M24 [chr 22]) from the proximal flanking region. They sequenced the LA14 138F10 distal region cosmid and the N 29M24 proximal region cosmid using the Sanger platform, and then used these sequences as queries to screen the GenBank database to identify BAC clones from the rDNA distal and the proximal flanking regions. The sequences of the flanking regions were then extended by BAC walking (searching overlapping BAC clones using BAC from the previous search) using BLAST. They identified 15 BAC clones (Accession no. AL592188, AL353644, CT476834, CT476837, CU633904, CU633906, CU633967, CU633971, CU634019, CU638689, FP236241, FP236315, FP236383, FP671120 and AC011841) from the distal

flanking region and three BAC clones (Accession no CR392039, CR381535 and AL354822) from the proximal flanking region. Further, they verified the positions of several of these BAC clones relative to the rDNA array using rDNA and BAC clones as probes for fluorescent *in situ* hybridization (FISH). BAC clone AC011841 is annotated in GenBank as being from chr 17 but the McStay lab FISH results using AC011841 as probe show that it is not from chr 17 but from an acrocentric chromosome short arm and is located in the distal rDNA flanking region. Lyle *et al.* (2007) had previously reported that the BAC clones CR392039 and CR381535 are from the rDNA proximal region, but had placed these midway between the rDNA array and the centromere rather than adjacent to the rDNA. This work is all described in Floutsakou *et al.* (2013). The identification of these BAC clones and cosmids from the distal and proximal flanking regions provides the opportunity to explore the rDNA flanking regions and to identify potential functional elements present within them.

3.1.2. Experimental strategy to characterize the rDNA flanking regions

To characterize the rDNA flanking regions I designed three analysis phases (Figure 3.1). Sakai *et al.* (1995) have reported the proximal-rDNA junction position at position 6,229 in ITS-1 on chr 15, chr 21 and chr 22 (Section 1.6). However, they were not able to identify the same junction in chr 13 and chr 14, suggesting that the proximal-rDNA junction position is not constant among the chromosomes. Further, in the same study the proximal junction in ITS-1 in chr 21 was found in one cell type but not another, suggesting that the proximal junction position may vary between individuals and/or among different tissues of an individual. Therefore, in the first analysis phase I decided to verify the known proximal-rDNA junction position and search for other potential junction positions using three different data sources (Section 3.3.1). First, to ensure that the proximal junction sequences do actually adjoin the rDNA array and are not just rDNA fragments duplicated elsewhere in the genome, I extended the length of rDNA sequence linkage with the proximal junction position using cosmids identified by the McStay lab. Second, to search for further evidence of the proximal junction position I designed a mapping based pipeline that uses BAC clones and cosmids from the proximal flanking region as reference sequences to identify WGS reads that span the rDNA junction position. Third, identified junctions were verified by PCR amplification. The restriction map of region around the distal-rDNA junction position is constant among all the five acrocentric chromosomes (Worton *et al.* 1988) and therefore distal-rDNA junction position at 39,029 also thought to be conserved among the acrocentric chromosomes. Therefore, to search if the distal junction position in the rDNA is constant among all five acrocentric chromosomes I decided to perform the distal-rDNA junction verification using same strategy as for the proximal junction position (Section 3.3.1).

Worton *et al.* (1988) demonstrated that the distal flanking region is conserved among the acrocentric chromosomes by comparing restriction maps from the distal region. Further, Gonzalez and Sylvester (1997; 2001) have reported that the distal region is conserved among the acrocentric chromosomes by amplifying regions near to the distal junction position. However, none of these studies has been able to quantify the level of inter-chromosomal sequence conservation. Therefore, in the second analysis phase I set out to quantify the level of intra- and inter- chromosomal sequence conservation of the rDNA distal and proximal flanking regions by comparing the overlapping regions of BAC clones and cosmids sequences from the same and different chromosomes (Section 3.3.2).

The lack of extensive rDNA flanking region sequences until now has meant that the sequence characteristics of these regions are not known. Therefore, in the third analysis phase I decided to construct representative consensus sequences for the distal and the proximal flanking regions using the BAC clones identified by the McStay lab (Section 3.3.3) and used these representative sequences to determine the repeat content, segmental duplicates and putative genes in the flanking regions (Section 3.3.4).

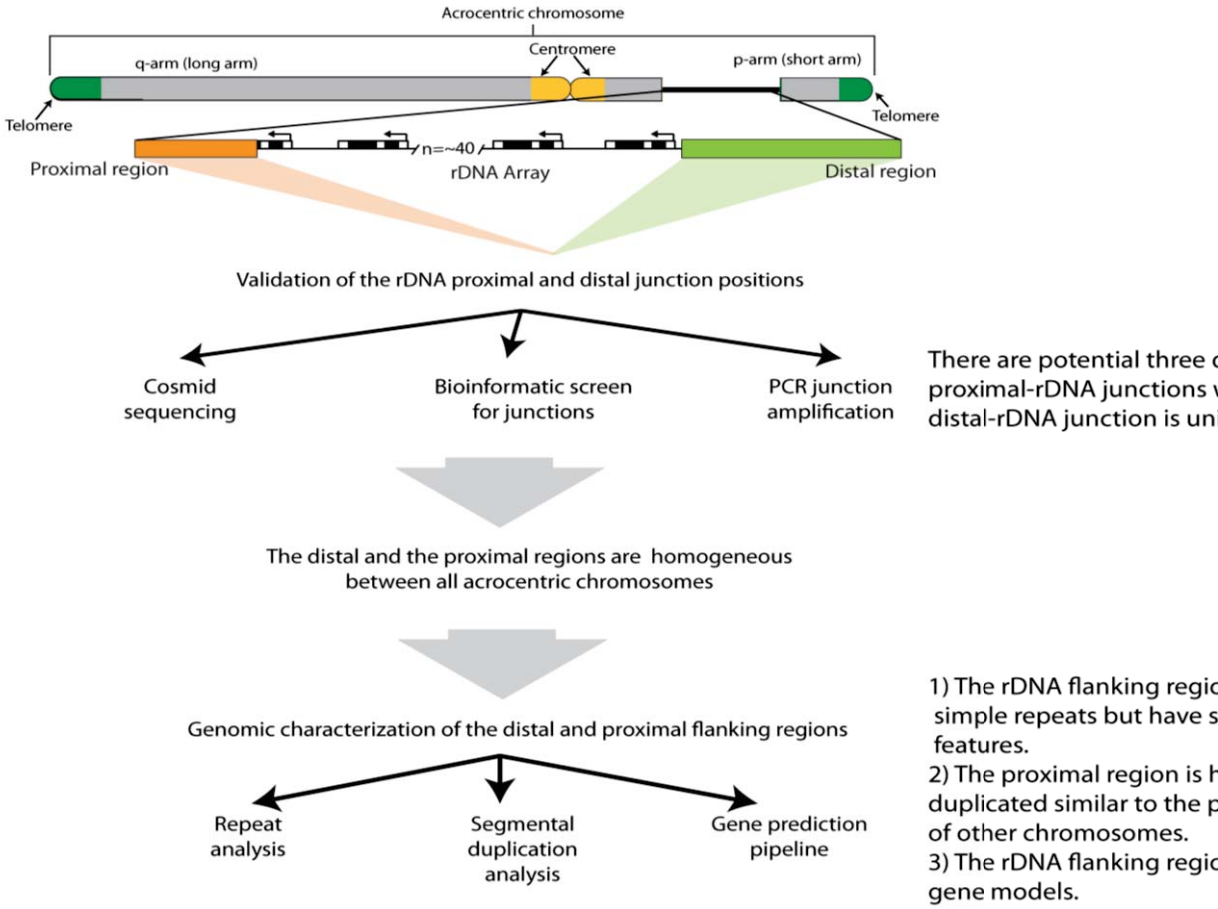


Figure 3.1: Schematic overview of the characterization of the rDNA flanking regions

Flow diagram showing the progression of the project and the different analyses performed regions. The major outcomes of the study are shown to the right of the figure.

3.2. Material and Methods

3.2.1. Sequencing and assembly of distal-rDNA and proximal-rDNA junction cosmids

Indexed libraries were prepared from individual cosmids using a Nextera™ DNA sample prep kit and Nextera™ barcodes (Epicentre NGS). NGS was performed on an Illumina Genome Analyzer IIx, using 54 bp singleton processing (Ambry Genetics, USA). The low quality ends of reads were trimmed with a quality score cutoff of 13 using DynamicTrim and shorter reads were removed with a length cutoff of 25 bp using LengthSort, both a part of the SolexaQA package (Cox *et al.* 2010). Cosmid sequences were assembled using Abyss v1.2.7 (Simpson *et al.* 2009). Velvet v1.101 (Zerbino and Birney 2008) was used to refine the assemblies obtained from Abyss. Different values for parameter k-mer size (-k) and minimum mean k-mer coverage of a unitig (-c) were used for Abyss to obtain the optimum assembly for the different cosmids, as follows:

LA13 165F6: ABYSS -k 35 -c 88 -e 2 -o LA13_165F6_assembly.fasta LA13_165F6 NGS_data.fastq

LA14 101B3: ABYSS -k 35 -c 99 -e 2 -o LA14_101B3_assembly.fasta LA14_101B3 NGS_data.fastq

LA15 25H3: ABYSS -k 35 -c 75 -e 2 -o LA15_25H3_assembly.fasta LA15_25H3 NGS_data.fastq

N 29M24: ABYSS -k 35 -c 114 -e 2 -o N29M24_assembly.fasta N29M24 NGS_data.fastq

LA13 133H12: ABYSS -k 35 -c 100 -e 2 -o LA13_133H12_assemlby.fasta LA13_133H12 NGS_data.fastq

LA15 64C10: ABYSS -k 35 -c 200 -e 2 -o LA15_64C10_assembly.fasta LA15_64C10 NGS_data.fastq

3.2.2. Sequence mapping based screen for proximal-rDNA junctions

A five step mapping method was designed to identify reads in WGS data that are from putative proximal-rDNA junction positions (Figure 3.2):

1) WGS data are mapped to the reference sequence
2) Reads partially mapping to the flanking region are mapped to the rDNA

3) Reads mapping to the rDNA and flanking region are clustered according to their position in the reference sequence

4) Reads from the clusters are filtered

5) Search for the junction in the trace archive database

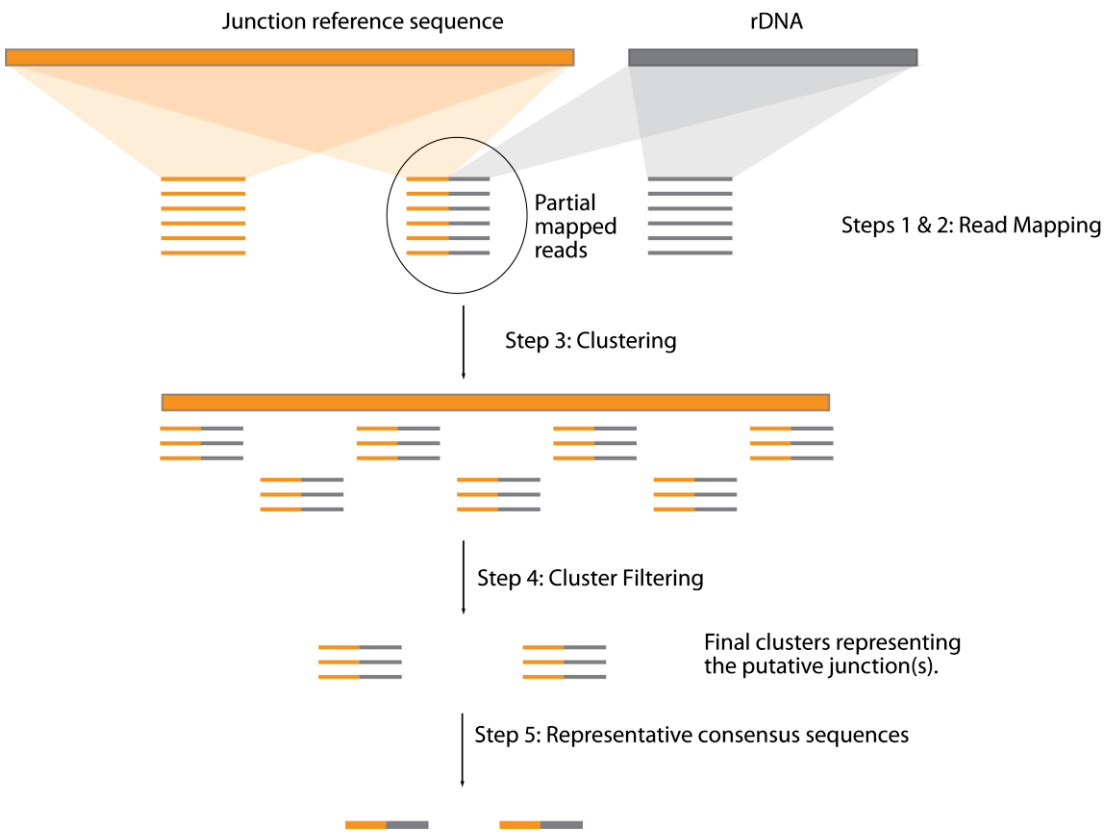


Figure 3.2: Workflow for junction verification mapping pipeline.

Steps 1 & 2: Whole genome shotgun reads (lines) were mapped to the flanking region reference sequence (orange box) and to the rDNA (grey box) with cut-offs of 95% identity and 100 bp minimum alignment length. Step 3: Partially mapped reads that matched both the reference sequence and the rDNA (orange-grey hybrid lines) were clustered according to their position on the reference sequence. Step 4: Clusters were filtered using criteria described in Section 3.2.2.3. Step 5: Cluster sequences were made for each of the resulting clusters.

3.2.2.1. Data Acquisition

Human WGS data from The Center for Applied Genomics (CRA) were downloaded from the Ensemble trace archive (currently available through the NCBI trace archive using the query: species_code="HOMO SAPIENS" AND CENTER_NAME = "CRA" AND STRATEGY = "WGA"). All 27,944,655 reads were used for the analysis.

3.2.2.2. Reference sequence preparation

To search for the proximal-rDNA junctions in all five acrocentric chromosomes, the four proximal region cosmids and the BAC clone CR392039 (with the rDNA sequence trimmed off) were used as reference sequences (Table 3.1). The distal BAC clone AL353644 and GenBank sequence U13369 were used as the reference sequences for the distal region and the human rDNA respectively.

Table 3.1: The part of the flanking region reference used for mapping the WGS reads.

Flanking reference sequence	Reference coordinates used for mapping
LA13 165F6 (13)	1 - 20,193
LA14 101B3 (14)	1 - 27,432
LA15 25H3 (15)	1 - 35,460
N 29M24 (22)	1 - 32,522
CR392039 (21)	1 - 155,929
Distal contig	1 - 379,046

3.2.2.3. Pipeline

The steps described below were repeated for each flanking reference sequence individually:

Step 1: Reads were mapped to the flanking reference sequence using 95% identity and 100 bp minimum alignment length as the cut offs. All reads were considered as single end reads during the mapping. gsMapper v2.3 (454 Roche) was used for mapping. Reads with “Partial” mapping status from the “454ReadStatus.txt” output file were selected from the mapped reads for the next step (the term “Partial” is used in gsMapper to denote reads that only a part maps to the reference sequence).

Step 2: The partially mapped reads obtained from step 1 were then mapped to the human rDNA. Reads that partially mapped to the rDNA with 95% identity and 100 bp minimum alignment length were selected.

Step 3: Reads that partially mapped to both the flanking region reference sequence and the rDNA (in steps 2 and 3) were clustered together according to their mapped position in the flanking region sequence. The flanking region sequence was divided into partially overlapping bins, and the reads were placed into these bins.

Step 4: Reads were removed if they fell in one of the following categories:

a) Same region of the read is mapped to both the flanking and rDNA sequences.

b) Entire read does not match across the potential junction point i.e. only part of the read aligned to the junction regions.

c) Entire read also matches another region of the human genome.

d) There is only one read in the cluster.

Step 5: To check that the potential junction sequences found are not restricted to one sequencing center, the cluster sequences that remained from step 4 were used to search complete human whole genome sequencing data present in the NCBI trace database. This includes all reads submitted by the different sequencing centers, including the CRA dataset. All read hits that were found across the junctions were end-trimmed using quality scores (as BLAST does not take quality score into account) and BLAST (Altschul *et al.* 1990) performed again to the respective reference sequences.

3.2.3. PCR amplification of the proximal-rDNA and distal-rDNA junction positions

3.2.3.1. Junction region amplification

To confirm the junctions identified in sections 3.3.1.1 and 3.3.1.2, primer pairs were designed such that the forward primer was from the flanking region and the reverse primer was from the rDNA. Primers con7_pro/con7_rDNA were used to amplify the proximal-rDNA junction in the 18S, con8_pro_F/con8_rDNA_R were used to amplify the proximal-rDNA junction in ITS-1, and distal_reg/rDNA_reg were used to amplify the distal-rDNA junction in the IGS (Table 3.2). PCR was performed using different PCR protocols and Taq polymerase from different manufacturers (Table 3.3). Human male genomic DNA (Promega; Catalog no. G1471) was used as template.

Table 3.2: Primer pairs used for rDNA junction verification

Primer Name	Sequence	Reference
con7_pro	ACCCTGACTTTATGGCACCTGGG	This study
con7_rDNA	GCGCTCTACCTTACCTACCTGG	This study
con8_pro_F	AGGTCATAGGGAGATAGTGTCG	This study
con8_rDNA_R	AAGAAGGGCGTGTCGTTG	This study
distal_reg	TGCAGGAAAGACGGTGTGCGTG	This study
rDNA_reg	TTGAGGCCTCGAAAGGCGAGAG	This study
M13-F	GTAAAACGACGGCCAG	
M13-R	CAGGAAACAGCTATGAC	

Table 3.3: PCR protocols used for different primer pairs to amplify the junction region.

Primer combination	Denaturation temperature	Annealing Temperature	Extension Temperature	# of cycles	Buffer used
con7_pro/con7_rDNA	95 ^{o}C	63.7 ^{o}C	72 ^{o}C	35 cycles	FastStart Taq polymerase (Roche)
con8_pro_F/con8_rDNA_R	95 ^{o}C	Gradient (50 ^{o}C – 64 ^{o}C)	72 ^{o}C	35-45 cycles	MyTaq™ HS (Bioline), FastStart Taq polymerase (Roche), Taq polymerase (Roche)
Distal_reg/rDNA_reg	95 ^{o}C	62 ^{o}C	72 ^{o}C	35 cycles	MyTaq™ HS (Bioline)

3.2.3.2. Cloning and Transformation

PCR fragments were cloned into the pCR®2.1 plasmid (Invitrogen) and transformed by heat shock into *Escherichia coli* TOP10 electrocompetent cells according to the manufacturer's instructions. To screen for transformed colonies harbouring the correct insert, colony PCR was performed using the M13_F/M13_R primer pair (Table 3.2). Plasmid DNA was extracted from the transformants according to the protocol "Alkaline lysis with SDS method" as described in following steps:

LB medium with ampicillin was inoculated with a transformed bacterial colony and incubated overnight at 37^{o}C in the shaking incubator. 1,500 µl of the overnight culture

transferred into a microcentrifuge tube and was centrifuged at 15,000 rpm for 2 mins at 4^{o}C. The medium was removed and the bacterial pellet resuspended in 200 µl of Alkaline lysis solution I by rigorous vortexing. 400 µl of freshly prepared Alkaline lysis solution II was added to the bacterial suspension and was mixed by inverting the tube 5 times. 300 µl of Alkaline lysis solution III was added to the tube and was mixed by inversion. The microcentrifuge tube was incubated in ice for 5 mins and then centrifuged at 15,000 rpm for 5 mins at 4^{o}C. 600 µl of the supernatant was transferred into a fresh microcentrifuge tube and an equal volume of phenol:chloroform was added. The organic and aqueous phases were mixed by vortexing and the emulsion was centrifuged at 15,000 rpm for 2 mins at 4^{o}C. The aqueous phase was transferred into a fresh microcentrifuge tube and 400 µl of isopropanol was added to precipitate the DNA. This was mixed by gentle vortexing and incubated overnight at -20^{o}C. The DNA was pelleted by centrifuging at 15,000 rpm for 5 mins at room temperature. The supernatant was removed by gentle aspiration and 1 ml of 70% ethanol was added, and this was centrifuged at 15,000 rpm for 5 mins at room temperature. The supernatant was removed and tubes were left at room temperature to evaporate the ethanol. The DNA pellet was then resuspended in autoclaved distilled water.

Alkaline lysis solution I

50 mM glucose

25 mM Tris-Cl (pH 8.0)

10 mM EDTA (pH 8.0)

Alkaline lysis solution II

0.2 N NaOH (freshly diluted from a 10 N stock)

1% (w/v) SDS

Alkaline lysis solution III

5 M potassium acetate, 60.0 ml

Glacial acetic acid, 11.5 ml

H_2O, 28.5 ml

3.2.4. Intra- and inter-chromosomal identity of the rDNA distal and proximal regions

To quantify intra- and inter- chromosomal homogeneity in the rDNA distal region and proximal regions, the level of intra- and inter- chromosomal identity between overlapping regions of the BAC and cosmid sequences was determined using the Stretcher (global alignment tool from EMBOSS; Rice *et al.* 2000); and YASS (Noe and Kucherov 2005).

3.2.5. Repeat content of the rDNA distal and proximal region contigs

The repeat content of the contigs was determined using RepeatMasker. Novel tandem repeats in the contigs were identified using Tandem repeat finder (Benson 1999) and BLAST (Altschul *et al.* 1990). MAFFT (Katoh *et al.* 2009) was used to generate the 138 bp repeat multiple sequence alignment and HMM-Logo (Schuster-Bockler *et al.* 2004) to generate the Logo. mVista was used to draw the sequence similarity plot of the inverted repeat (Frazer *et al.* 2004).

3.2.6. Segmental duplication analysis of the rDNA distal and proximal contigs

Segmental duplicates in the rDNA distal and proximal region contigs were detected using a modified BLAST-based detection scheme called the "whole genome assembly comparison" (Bailey *et al.* 2001a). The human genome assembly hg19 was used for the analysis, and was broken into 400 kb pieces. The repeats in this fragmented human genome and in the contigs were masked using RepeatMasker (http://www.repeatmasker.org/). The contigs were then matched to the fragmented masked genome using BLAST. A cutoff of ≥85% identity over 1 kb was used. Next, the repeats were reinserted into these matched sequences and global alignments were created using Stretcher. All steps were performed using a series of Perl scripts (courtesy J.A. Bailey, University of Massachusetts Medical School, USA). Low identity ends of the sequences were identified from the alignments and trimmed. Where the ends of two human genome fragments match a single region or where a fragment is interrupted by repeats, these fragments were merged together. This step was performed manually. The merged sequences were then aligned again using Stretcher to recalculate the identity.

3.2.7. Gene prediction pipeline for the rDNA distal and proximal contigs

A four-stage pipeline was designed to determine the presence of potential gene coding regions in the rDNA distal and the proximal contigs. The pipeline was first used for the contigs by masking repeat elements in the sequence using RepeatMasker. This sequence is denoted as masked sequence as repeats are masked by Ns. To determine the effect of repeats on gene models, I also used the sequence without masking the repeats. This sequence is denoted as the unmasked sequence.

Step 1: *Ab initio* gene prediction tools Genscan (Burge and Karlin 1997), Fgenesh (Salamov and Solovyev 2000), glimmerHMM (Majoros *et al.* 2004) and GeneMark (Besemer and Borodovsky 2005) were used to predict gene signals.

Step 2: Homology searches for both contigs were performed using BLAST against the following GenBank datasets: non-redundant protein and reference mRNA sequences restricted to primates, and EST sequences restricted to human. Hits were then remapped to the contigs using two spliced alignment tools: Exonerate (Slater and Birney 2005) and PASA (Program to Assemble Spliced Alignments; Salamov and Solovyev 2000). Exonerate was used for protein, reference mRNA, and EST sequences, while PASA was only used for EST sequences. For both tools 90% identity was used as the filter cutoff.

Step 3: All the evidence from Steps 1 and 2 were combined using EVidence Modeller (EVM; Haas *et al.* 2008). To optimize the weightings of the *de novo* gene prediction, protein, and EST evidence streams for merging, EVM simulations were performed using variable weights. Any EVM gene models that did not contain any database evidence or contained one-or-more exons that lacked any database evidence were removed.

Step 4: In the final curation step, EVM gene models, as well as protein sequence matches from Exonerate that were not included by EVM in step 3, were mapped to the contigs to compile the complete set of putative gene models. Apollo v1.11.6 (Lewis *et al.* 2002) was used for visualization and refinement of the gene models.

3.3. Results

3.3.1. Verification of the proximal-rDNA junction

3.3.1.1. Extending linkage into the rDNA from the proximal region junction and searching various junction positions using cosmid sequences

Proximal region: Sakai *et al.* (1995) reported a 493 bp sequence (GenBank Accession # D31961) from the proximal junction with the rDNA that only includes 111 bp of rDNA sequence. Further out of the three proximal region BAC clones identified by McStay Lab, one BAC clone (CR392039 from chr 21) was found to be adjacent to the rDNA array that has junction at position 4,234 bp in 18S of the rDNA (Floutsakou *et al.* 2013). The variation in the position of the proximal rDNA junction in the Sakai *et al.* and BAC clone suggests that more than one junction positions are possible. Since only a small fragment of the rDNA has been reported to be linked to the proximal junction, it is also possible that these sequences are not part of the rDNA array, but instead is from some other region of the human genome where a fragment of the rDNA has been duplicated. Hence, to search if the rDNA sequence next to the proximal-rDNA junction position is likely to be from the rDNA array, I decided to see how far the rDNA sequence that is linked to the proximal junction can be extended by sequencing four cosmids (LA13 165F6 [chr 13], LA14 101B3 [chr 14], LA15 25H3 [chr 15] and N 29M24 [chr 22]) that were identified to cross the proximal junction with the rDNA. These cosmids were identified using 493 bp proximal region sequence identified by Sakai *et al.* (GenBank Accession # D31961) as probe to screen chromosome specific cosmid libraries (see Section 3.1.1 for the details of the cosmid identification). The cosmids were sequenced on the Illumina platform (Section 3.2.1) and obtained NGS data for cosmids were assembled using Abyss (Section 3.2.1; Assembly statistics in Appendix Table 4). The position of the junction in the proximal region cosmids were demarcated by comparing their sequences with the rDNA U13369 using BLAST. The cosmid sequences include between 2.4 kb and 16.5 kb of rDNA sequence, suggesting that the rDNA sequence next to the junction identified by Sakai *et al.* is adjoined to the actual rDNA array (Figure 3.3).

In cosmid sequences, LA13 165F6 (chr 13) and LA15 25H3 (chr 15) the junction position is in ITS-1 at position 6,229 (coordinates according to the rDNA sequence U13369) of the rDNA unit. This junction position in ITS-1 at 6,229 is same as that reported by Sakai *et al.* (1995). In cosmids LA14 101B3 (chr 14) and N 29M24 (chr 22) junction position is in ITS-1

at position 6,339. This other junction is novel. This demonstrates that the proximal region may have three different junction positions: in 6,229 in ITS-1, 6,339 in ITS-1, and 4,234 in 18S.

The extension of the rDNA unit next to the junction position supports that the rDNA unit is not segmentally duplicated. However, it is also possible that the proximal-rDNA junction position region in the cosmids and/or BAC clone is a large segmentally duplicated region and not part of the true rDNA array. To rule out this possibility I compared the sequence of the ETS and 18S rRNA gene from the flanking regions cosmids and BAC clones with that of the human rDNA U13369 using BLAST. The rationale for this comparison is that segmentally duplicated rDNA fragments are expected to have accumulated mutations and thus be diverged compared to the real rDNA units, as the segmentally duplicated units are likely to be non-functional and thus evolving neutrally. The average sequence identity between proximal junction rDNA sequences and the human rDNA is 99.0% (Table 3.4), which is much higher than the segmental duplicates of the ETS/18S rRNA in the reference genome (90.0%) (Table 3.5). This shows that the rDNA adjacent to the proximal region in the cosmids and BAC clones is likely the real rDNA array and not a segmentally duplicated region. However, it cannot be ruled out that these regions are recent long segmental duplicates that have not had enough time to accumulate sequence differences.

Table 3.4: Sequence comparison between the human rDNA and proximal junction rDNA.

Human rDNA region	Proximal junction cosmids and BAC	% identity
human_rDNA_ETS_18S	LA13 165F6 (chr 13)	98.2
human_rDNA_ETS_18S	LA14 101B3 (chr 14)	98.9
human_rDNA_ETS_18S	LA15 25H3 (chr 15)	99.8
human_rDNA_ETS_18S	CR392039 (chr 21)	99.8
human_rDNA_ETS_18S	N 29M24 (chr 22)	98.4
Average % identity		99.0

Table 3.5: Sequence comparison between the human rDNA and segmentally duplicated rDNA fragments.

Human rDNA region	Human reference genome	% identity
human_rDNA_ETS_18S	Homo sapiens chr 16 NC_000016.9	92.0
human_rDNA_ETS_18S	Homo sapiens chr 2 NC_000002.11	90.7
human_rDNA_ETS_18S	Homo sapiens chr Y NC_000024.9	89.2
human_rDNA_ETS_18S	Homo sapiens chr 20 NC_000020.10	88.1
Average % identity		90.0

The proximal region adjacent to the proximal-rDNA junction is slightly variable between the cosmids and BAC clone. In all cosmid sequences that are similar to the proximal region sequence reported by Sakai *et al.* (1995) a 68 bp unique region is present next to the junction followed by block of 147 bp ACRO1 repeats (Figure 3.3). However, the copy number of the ACRO1 repeat block is variable between each cosmid. The proximal region adjacent to the proximal-rDNA junction in BAC clone CR392039 differs from the cosmid sequences as a ~2.7 kb region that includes the 68 bp unique region, the first ACRO1 repeat blocks, adjacent to the proximal junction point, which is present in the cosmids, is deleted from the BAC clone (Figure 3.3)

Distal region: I used a similar approach as used for the proximal region to characterize the distal rDNA junction position in different acrocentric chromosomes. Five BAC clones (AL592188, AL353644, FP671120, CT476837 and FP236383) from the rDNA distal region were identified to contain rDNA sequence at their termini. These BAC clones are from chr 21 and chr 22, and to obtain distal junction position sequences from the remaining acrocentric chromosomes, two cosmids (LA13 133H12 [chr 13] and LA15 64C10 [chr 15]) identified by the McStay lab as coming from the distal region (Section 3.1.1) were sequenced on the Illumina platform (Section 3.2.1). All five BAC clones together with cosmid LA13 133H12 have the junction position at 39,029 in the IGS of the rDNA (coordinate according to the rDNA sequence U13369) (Figure 3.3). LA15 64C10 (and the previously sequenced LA14 138F10) do not contain any rDNA sequence (Figure 3.3). This represent that the distal region is constant at least among chr 13, chr 21 and chr 22 and occurs in the rDNA IGS. Further, unlike the proximal junction region, the distal region sequence adjacent to the junction position is the same in all cases examined to date.

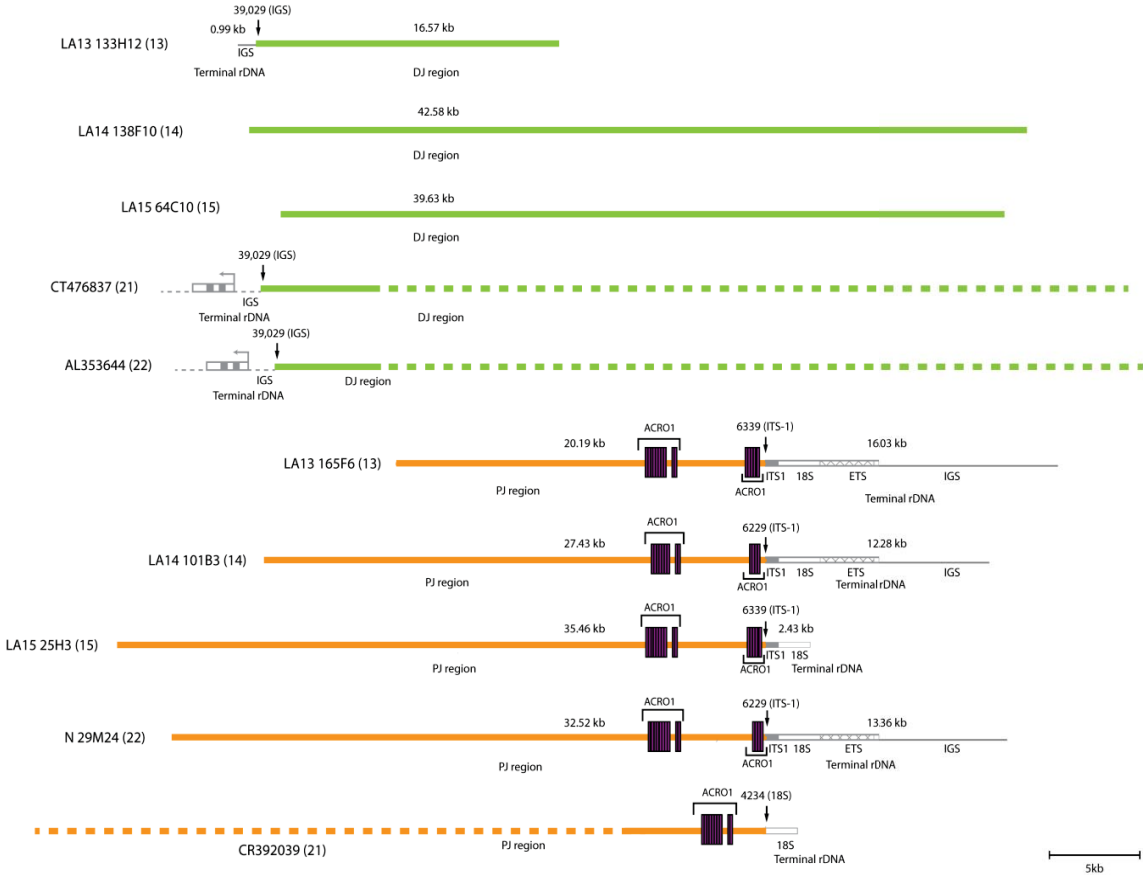


Figure 3.3: Structure of distal region and proximal region clones around the rDNA junction

Three distal region (solid green) and four proximal region (solid orange) cosmid clones from the rDNA flanking regions. The junction position with the rDNA (grey) are shown as arrows with the position (in bp) indicated above the arrow (position is relative to GenBank rDNA sequence U13369). The ACRO1 satellite blocks adjacent to the proximal junction are shown as purple boxes. The length of the flanking regions and the rDNA are indicated immediately above each clone. The cosmid name and chromosome of origin (in parentheses) are indicated to the left of the clone. Since these cosmids do not represent all the acrocentric chromosomes, BAC clones CT476837 and AL353644 (dotted green lines) from the distal region and BAC clone CR392039 (dotted orange line) from the proximal region are included to represent the flanking region from the remaining chromosomes. The scale is shown below.

3.3.1.2. Bioinformatics screen for proximal rDNA junctions

WGS data theoretically contain the information from all regions of the genome and therefore are expected to contain reads that represent the rDNA junction positions where one part of the read covers the flanking region and the other part covers the rDNA next to the junction position. To take an independent approach for searching the proximal junction positions within an individual, I designed a five-step mapping-based pipeline (Section 3.2.2) to identify the reads in WGS data that span the rDNA and the proximal/distal junction regions. The proximal region sequences from all the four cosmids and the BAC clone CR392039 were each used as proximal region reference sequences (Section 3.2.2; Table 3.1). The distal region was included in the analysis, using BAC clone CT476837 (one of five BAC clones that were identified adjacent to the rDNA array) as the distal region reference sequence.

Table 3.6: Results for different steps of the junction mapping pipeline.

Flanking reference sequence	**# reads partially mapped to flanking sequence (Step 1)**	**# reads mapped to the rDNA (Step 2)**	**# of clusters (Step 3)**	**# of clusters after filter (Steps 4 & 5)**	**Junction position of the cluster (# of reads)**
LA13 165F6 (13)	325,570	417	6	1	at position 4,234 in 18S (4)
LA14 101B3 (14)	318,834	1,307	12	2	at position 6,229 in ITS-1 (5) at position 4,234 in 18S (3)
LA15 25H3 (15)	332,822	746	15	1	at position 4,234 in 18S (4)
N 29M24 (22)	448,750	45,556	10	1	at position 6,229 in ITS-1 (5)
CR392039 (21)	828,792	102,938	17	1	at position 4,234 in 18S (4)
Distal contig	554,888	19,023	8	1	at position 39,029 in IGS(24)

WGS data from the CRA sequencing center were mapped to the distal and proximal reference sequences using gsMapper to identify reads that partially mapped to the sequences (column two of the Table 3.6; Section 3.2.2.3). These partially mapped reads were mapped to the human rDNA using gsMapper (column three of Table 3.6; Section 3.2.2.3) to identify the reads that partially map to the proximal region and the rDNA. Such partially mapping reads were clustered according to their position in the flanking region reference sequence

(column four of Table 3.6). Finally, these clusters were filtered using criteria designed to identify the ones that represent putative proximal-rDNA and distal-rDNA junctions (column five of the Table 3.6; Section 3.2.2.3). As expected, all the putative junction reads for the distal-rDNA junction were found to have the junction at position 39,029 in the IGS. This is consistent with the clone screening results above (Section 3.3.1.1), showing that the distal-rDNA junction position is likely to be constant for all five acrocentric chromosomes. This demonstrates that the pipeline is capable of identifying rDNA junction positions correctly. Next, I looked for the proximal junction position in the putative junction reads. Similar to the results from the cosmid sequences, the proximal junction position in these reads is variable, with junction positions at either position 6,229 in ITS-1 or position 4,234 in the 18S rRNA (column six of Table 3.6). However, the junction at 6,339 in ITS-1 was not present in CRA WGS data. These results suggest that the proximal rDNA junction position that was found in cosmids LA14 101B3 and N 29M24 is variable between chromosomes and individuals, with at least three possible junction positions.

3.3.1.3. PCR amplification of the junction regions

To search for additional evidence for the various proximal rDNA junction positions, I decided to amplify these junction regions using human genomic DNA as template. I designed primer sets such that the forward primer binds to rDNA flanking region and the reverse primer binds to the rDNA on the other side of the junction (Figure 3.4). Two primer sets, con7_pro/con7_rDNA and con8_pro_F/con8_rDNA_R, were designed to amplify the proximal junction region in ITS-1 and the 18S rRNA, respectively. A primer set, distal_reg/rDNA_reg, was designed to amplify the distal junction as a positive control.

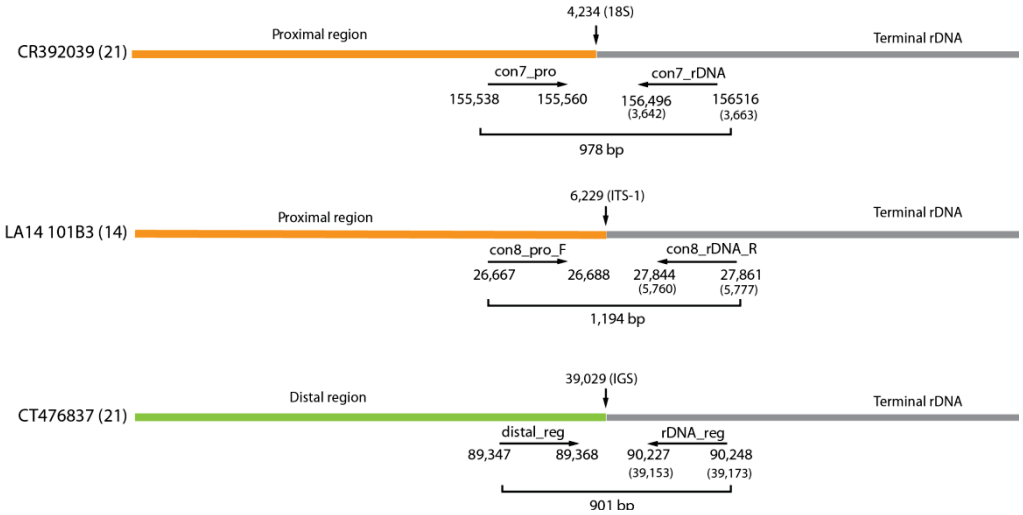


Figure 3.4: The positions of the primer pairs for the amplification of the proximal-rDNA and distal-rDNA junctions

The primer pairs were designed to amplify fragments that span the proximal (orange) and distal (green) rDNA junctions. The junction positions are shown as downward arrows with the positions (in bp) and the regions in the rDNA (grey) indicated above (position is relative to GenBank rDNA U13369). The forward primer (forward arrow) is in the flanking region and the reverse primer (reverse arrow) is in the rDNA. The name of the primer is indicated above the arrow and the positions of the primers are indicated next to the arrows. The expected size of the amplified fragment is indicated in bp below. The names of the cosmid and BAC clones used to design the primers and their chromosomal origins (in parentheses) are indicated to the left. Figure not to scale.

The primer set con7_pro/con7_rDNA amplified a 978 bp fragment that covers the junction at position 4,234 in the 18S rRNA gene (Figure 3.5). This junction is the same as that found in BAC clone CR392039 and is one of the proximal junctions identified from the WGS mapping pipeline. However, I was unable to amplify the proximal junction in ITS-1 using the con8_pro_F/con8_rDNA_R primer set with the same template as above. Different PCR conditions *viz.* temperature gradient PCR, variation in PCR cycle number, different polymerases, and different template concentrations were employed (Section 3.2.3), but either no or nonspecific products were obtained. One possible reason for the inability to amplify the ITS-1 is the presence of the ACRO1 satellite repeat block adjacent to the ITS-1 proximal-rDNA junction (Figure 3.6). ACRO1 satellite repeats are also present in other parts of the genome (Sakai *et al.* 1995) hence it is possible that the proximal region primer (con8_pro_F) from the ITS-1 junction primer pair does not bind uniquely to the junction region and therefore gives nonspecific products. The primer set distal_reg/rDNA_reg

amplified an 901 bp fragment that covers the junction at position 39,029 in IGS (Figure 3.5). This junction is the same as that found in the cosmid screening and the WGS mapping pipeline, confirming that the distal junction position in the IGS is constant.

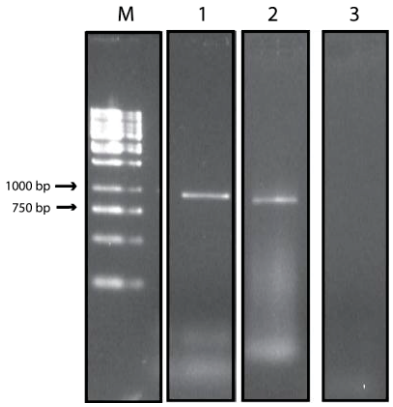


Figure 3.5: Amplification of the flanking region-rDNA junction.

The 18S rRNA proximal-rDNA junction region was amplified using primer pair con7_pro/con7_rDNA to obtain ~1 kb fragment (lane 1) and the distal rDNA junction was amplified using primer pair distal_reg/rDNA_reg (lane 2) to obtain a ~900 bp fragment. Lane 3 is a negative control and lane M is a 1 kb Ladder.

The cosmid sequences, bioinformatics screening of WGS data and amplification of junction regions demonstrate that the proximal region is likely have at least two junction points: one in ITS-1 and the other in the 18S rRNA while the distal region just has a single junction at position 39,029 in IGS.

3.3.2. Inter- and intra-chromosomal sequence conservation of the rDNA flanking regions

After establishing the junction positions for the flanking regions, the next question was how similar the rDNA distal and proximal flanking regions are among the acrocentric chromosomes. Several studies have indicated that the distal and the proximal regions are conserved among the acrocentric chromosomes (Worton *et al.* 1988; Sakai *et al.* 1995; Gonzalez and Sylvester 1997; Gonzalez and Sylvester 2001). These conclusions were based on the comparisons of the length of amplified fragments and restriction map from the flanking regions using hybrid single human chromosome cells. However, because of the lack of availability of sequences from all the acrocentric chromosomes the level of sequence conservation was not known. Therefore, I decided to quantify the level of intra- and inter-

chromosomal sequence conservation in the rDNA flanking regions by comparing the overlapping regions from the cosmids and BAC clones from these regions.

3.3.2.1. Intra-chromosomal sequence conservation of the rDNA distal and proximal regions

First, the intra-chromosomal sequence conservation for the rDNA distal and proximal regions was calculated using the BAC clones from the same chromosomes to get an estimate of the level of intracellular variation in the flanking regions and the background error rate in these sequences. Twelve BAC clones from the distal flanking region of chr 21 were compared to quantify the level of sequence identity between clones from this chromosome (Appendix Table 5). All the BAC clones for the chr 21 are from the BAC library CHORI-507-HSA21 that was constructed using a blood sample from a single individual (http://www.chori.org/). All these BAC clones, except CU633967 and CU633906, are 100% identical to each other. CU633967 and CU633906 are 99.9% identical to each other while 100% identical to the remaining chr 21 BAC clones. The variation between CU633967 and CU633906 is because of a 5 bp gap and a mismatch. These differences are likely to be either sequencing errors or rare inter-cellular variation, as both BAC clones are from the same individual. Similarly, two BAC clones from chr 22 were also compared (Appendix Table 1), with these BAC clones coming from the BAC library RPCI-11.1 which was constructed using a blood sample from a single individual (http://www.chori.org/). These BAC clones (AL353644 and AL592188) have 99% identity, with 117 indels and 524 mismatches. The sequence variants between the clones are interspersed throughout the BAC clones. This variation may represent sequence error and/or intracellular variation within an individual. The level of sequence dissimilarity is surprising given the high similarity between clones from chr 21, and because both these chr 22 BAC clones are from the same individual. Two BAC clones from the proximal flanking region of chr 21 were compared to quantify the level of sequence identity between clones from this chromosome. Both BAC clones (CR392039 and CR381535) for the chr 21 are from the BAC library CHORI-507-HSA21 which was constructed using a blood sample from a single individual (http://www.chori.org/) and are 100% identical to each other.

Overall, as expected both the distal and proximal regions are highly intra-chromosomally conserved with the average intra-chromosomal sequence identity approaches 100%. However, case by case study shows that the level of intra-chromosomal conservation is variable among the chromosomes. The distal region in the chr 21 and chr 22 shows 100% and 99% intra-chromosomal identity respectively. The variation in the level of conservation

between the chromosomes suggests that homologous regions of the flanking regions may slightly variable depending on chromosomes.

3.3.2.2. Inter-chromosomal sequence conservation of the rDNA distal and proximal flanking regions is high

Next, I decided to look at the level of inter-chromosomal sequence conservation in the rDNA flanking regions between acrocentric chromosomes. To determine this, the flanking regions were divided into two parts. The regions up to 43 kb from the rDNA junction position were named the near-distal/near-proximal regions, and the regions away from the rDNA junction position were named the far-distal/far-proximal regions. The cutoff for demarcating the near and far distal/proximal regions was based on cosmid LA14 138F10 from the distal junction region, which is the most distal of the cosmids. Therefore, the near region represents all the region that includes some cosmid sequence and the same cutoff was used for the proximal region for consistency. This separation of the flanking regions was made to observe the change in the level of conservation as we move away from the rDNA array.

Sequence conservation at the near-distal region: I compared BAC clones from chr 21 and chr 22 and cosmids from chr 13, chr 14 and chr 15 to estimate the level of inter-chromosomal sequence conservation near to the distal-rDNA junction position. Since only one cosmid is present for chr 13, chr 14 and chr 15 while several BAC clones are present for chr 21 and chr 22, I selected representative BAC clones for chr 21 and chr 22 to avoid bias in the representation of the chromosomes in estimating the overall level of conservation of acrocentric chromosomes.

Two BAC clones AL592188 and AL353644 from chr 22 and three BAC clones FP671120, CT476837 and FP236383 from chr 21 come from the region adjoining the rDNA array (Section 3.3.1.1). The sequence identity between the three chr 21 BAC clones is 100% therefore, CT476837 was selected as the representative sequence of the chr 21 near distal region. The overall sequence identity between AL592188 and AL353644 is 99% but region adjacent to the distal-rDNA junction position is 100% identical demonstrating that both BAC clones have same sequences near to the junction position and therefore AL353644 was selected as the representative sequence of the chr 22 near distal region. Cosmids LA13 133H12, LA14 138F10 and LA15 64C10 were selected as the representative sequences of the chr 13, chr 14 and chr 15 near-distal regions, respectively. The sequences of selected clones were compared using YASS. The sequence identity between the clones varies from 99.0% to 99.7% (Table 3.7) with an average 99.1%, demonstrating that the distal flanking region is highly conserved among the acrocentric chromosomes.

Table 3.7: Sequence similarity matrix for the near-distal region cosmid and BAC clones

Clones (Chromosome no.)	LA13 133H12 (13)	LA14 138F10 (14)	LA15 64C10 (15)	CT476837 (21)	AL353644 (22)
LA13 133H12 (13)	X	Identity=99.7 % Length=12,814	Identity=99.6 % Length=13,544	Identity= 99.2% Length= 16,580	Identity=99.6 % Length= 16,452
LA14 138F10 (14)	Identity=99.7% Length=12,814	X	Identity=99.6 % Length=38,939	Identity=99.5 % Length=42,641	Identity=99.0 % Length=42,635
LA15 64C10 (15)	Identity=99.6% Length=13,544	Identity=99.6 % Length=38,939	X	Identity= 99.6% Length= 39,656	Identity= 99.0% Length= 39,669
CT476837 (21)	Identity= 99.2% Length= 16,580	Identity=99.5 % Length=42,641	Identity= 99.6% Length= 39,656	X	Identity=99.1 % Length=90,214
AL353644 (22)	Identity=99.6% Length= 16,452	Identity=99.0 % Length=42,635	Identity= 99.0% Length= 39,669	Identity=99.1 % Length=90,214	X

[a] Length = pairwise comparison alignment length.

Sequence conservation at the far-distal region: To determine if this distal flanking region conservation is restricted to the junction with the rDNA or extends away from the junction point, I compared BAC clone AC011841 with BAC clones from chr 21 using YASS. All the BAC clones from the far-distal region except AC011841 are from chr 21, and therefore AC011841 was compared with BAC clones from chr 21 to calculate the sequence conservation in the far distal region. The sequence comparison between AC011841 and the chr 21 BAC clones shows that the sequence identity in the far distal region varies from 98.3% to 98.8%, with an average sequence identity of 98.5% (Table 3.8).

Table 3.8: Sequence similarity matrix for the far distal region BAC clones.

Clones (Chromosome no.)	CT476834 (21)	CU633904 (21)	CU633906 (21)	CU633967 (21)	CU634019 (21)	CU638689 (21)
AC011841 (17?)	Identity = 98.3% Length = 55,002	Identity = 98.5% Length = 35,589	Identity = 98.9% Length = 92,734	Identity = 98.4% Length = 37,422	Identity = 98.3% Length = 47,490	Identity = 98.7% Length = 74,515

[a] Length = pairwise comparison alignment length.

The high sequence conservation in the far distal region demonstrates that inter chromosomal distal region conservation is not restricted to around the junction region but is extended towards the telomere. However, higher sequence conservation in the near distal region compare to far distal region demonstrate that the region near to the junction position is more constrained than the region away from the junction.

Sequence conservation at the near proximal region: To estimate the level of inter-chromosomal conservation in the region near to the proximal rDNA junction position, I selected four cosmids LA13 165F6, LA14 101B3, LA15 25H3 and N 29M24, and one BAC clone CR392039 as representatives of chr 13, chr 14, chr 15, chr 22 and chr 21 respectively. The selected clones were compared to each other using YASS. The sequence identity between the clones varies from 86.7% to 99.9%, with an average of 93.3% (Table 3.9). The sequences vary between the acrocentric chromosomes in this near-proximal region mainly because of variation in repeat element composition (Figure 3.6). Overall, the pairwise comparisons between the clones of all the five acrocentric chromosomes from the near proximal region demonstrate that the region is more variable among the acrocentric chromosomes compare to the near distal region. However, the variation is mainly restricted to the repeat elements while remaining regions are more or less conserved.

Table 3.9: Sequence similarity matrix for the near-proximal region cosmids and BAC clones.

Clones (Chromosome no.)	LA13 165F6 (13)	LA14 101B3 (14)	LA15 25H3 (15)	N 29M24 (22)	CR392039 (21)
LA13 165F6 (13)	x	Identity=94.2 % (18,989) Gap= 1% (397) Length= 20,159	Identity= 99.9% (20,182) Gap= 0% (16) Length= 20,204	Identity=93.4 % (19,225) Gap= 3% (732) Length= 20,573	Identity= 96.4% (17,245) Gap= 2% (421) Length= 17,897
LA14 101B3 (14)	Identity=94.2 %(18,989) Gap= 1% (397) Length= 20,159	X	Identity= 94.6% (24,525) Gap= 1% (485) Length= 25,913	Identity= 94.5% (26,505) Gap= 3% (857) Length= 28,050	Identity= 94.0% (23,979) Gap= 2% (757) Length= 25,500
LA15 25H3 (15)	Identity= 99.9% (20,182) Gap= 0% (16) Length= 20,204	Identity= 94.6% (24,525) Gap= 1% (485) Length= 25,913	X	Identity= 86.7% (28,092) Gap= 10% (3,533) Length= 32,415	Identity= 87.8% (30,236) Gap= 9% (3,367) Length= 34,455
N 29M24 (22)	Identity=93.4 % (19,225) Gap= 3% (732) Length= 20,573	Identity= 94.5% (26,505) Gap= 3% (857) Length= 28,050	Identity= 86.7% (28,092) Gap= 10% (3,533) Length= 32,415	x	Identity=96.6 % (28,984) Gap= 1% (457) Length= 29,999
CR392039 (21)	Identity= 96.4% (17,245) Gap= 2% (421) Length= 17,897	Identity= 94.0% (23,979) Gap= 2% (757) Length= 25,500	Identity= 87.8% (30,236) Gap= 9% (3,367) Length= 34,455	Identity=96.6 % (28,984) Gap= 1% (457) Length= 29,999	X

[a] Length = pairwise comparison alignment length. [b] gap = number of gaps in the alignment. These are highlighted for the inter-chromosomal comparisons of the proximal regions as they make a significant contribution to the sequence variation

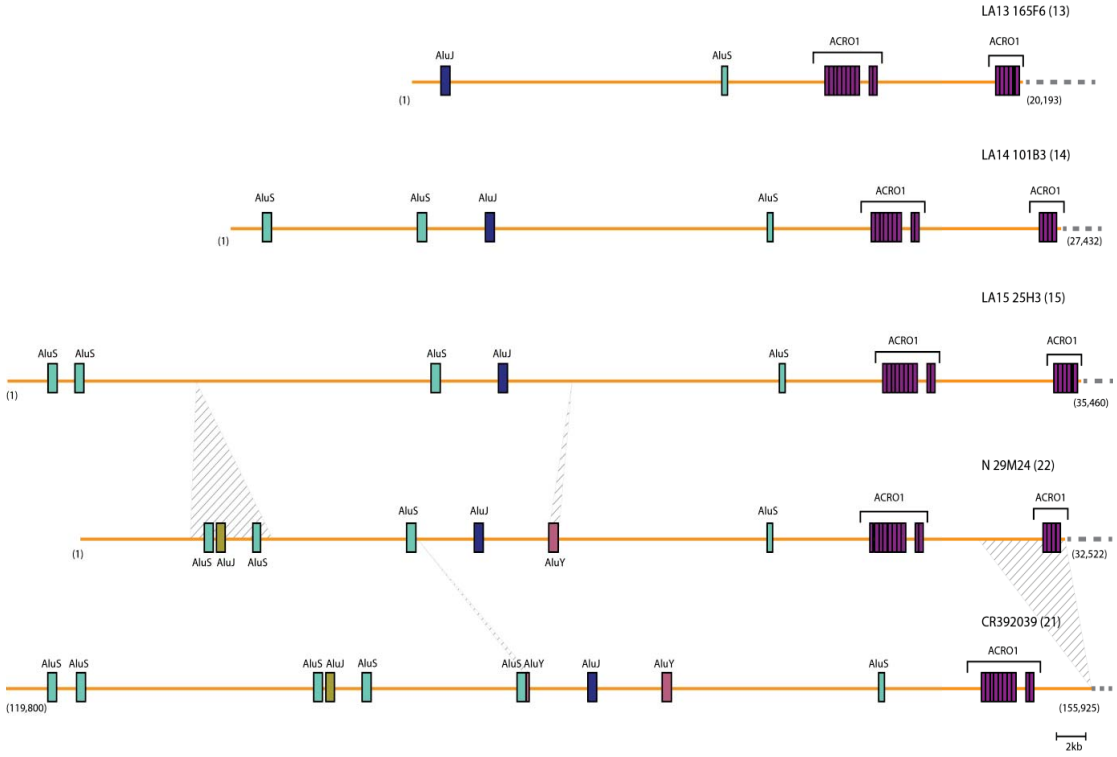


Figure 3.6: Inter-chromosomal variation in the near proximal region due to Alu elements and 147 bp ACRO1 repeat.

The proximal region sequence (orange line) adjacent to the rDNA (dotted grey line) junction is variable among the acrocentric chromosomes because of Alu element (coloured boxes except purple) and ACRO1 repeat (purple boxes) indels and copy number variation. The indels are represented by grey shaded areas. The coordinates of proximal region in the cosmid and BAC sequences is depicted in this figure are in parentheses next to the sequence. The clone names and chromosome origins (in parentheses) are indicated above the sequences. Scale is shown at the bottom.

Sequence conservation at the far proximal flanking region: To check if this level of conservation is maintained at the far end of the proximal flanking region, I compared BAC clones CR381535 (chr 21) and AC145212 (unplaced) using YASS. The sequence identity between these two BAC clones is 97.9%, indicating that the proximal flanking region is also conserved further away from the rDNA.

3.3.2.3. Overall conservation of the distal region and the proximal region

The quantification of intra- and inter- chromosomal sequence identity shows that the regions around the rDNA are highly conserved across all the acrocentric chromosomes (Figure 3.7). The high intra-chromosomal conservation in the distal and proximal regions was expected. For the proximal region, the inter-chromosomal sequence conservation is lower than the distal region. This greater level of variation mostly results from repeat element variation near to the junction of the proximal region with the rDNA. Therefore, similar to the distal region, the proximal region also seems to be highly conserved among acrocentric chromosomes.

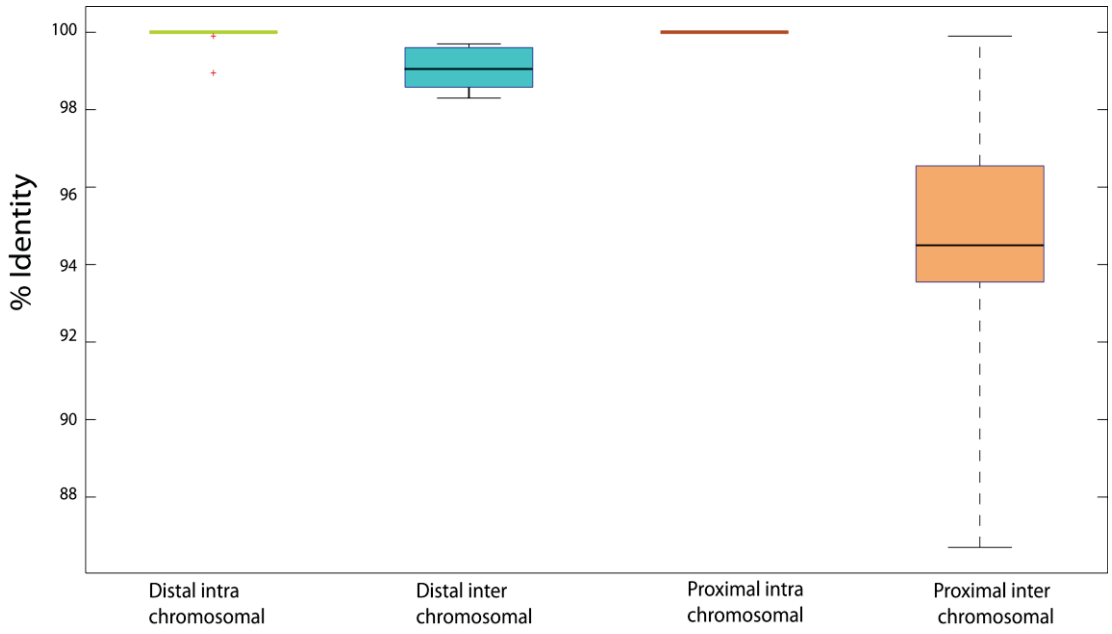


Figure 3.7: Average intra- and inter-chromosomal identities between distal and proximal flanking region clones

The horizontal black line inside the each box represents the mean of the identities. The upper and lower whiskers represent the upper and lower limits of the identities respectively. The region represent by each box is indicated below them on the horizontal axis. The vertical axis represents the level of sequence conservation.

3.3.3. Distal and Proximal contig construction

After determining the level of sequence conservation of the distal and proximal rDNA flanking regions, I next decided to characterize these regions. To search for features in the rDNA distal and proximal regions, I constructed representative sequences for each of them by merging BAC clones. The BAC clones were selected to span the largest stretches of the regions, and where possible BAC clones derived from the same chromosome were used to reduce inter chromosomal variation. Four BAC clones (CT476837, CT476834, CU633906 and AC011841) representing the minimum subset of overlapping clones were selected to construct the distal flanking region representative sequence. The overlapping regions between BACs CT476837 (chr 21) and CT476834 (chr 21), and CT476834 and CU633906 (chr 21) are 100% identical, but the identity decreases to ~98% between CU633906 and AC011841 (chr 17?) (Figure 3.8a). To construct the sequence, the BAC clones were merged in the order CT476837, CT476834, CU633906 and AC011841, such that the region that overlaps with the previous BAC clone was truncated before merging. This resulted in a 379,046 bp distal flanking region sequence (Figure 3.8a) that was named the "distal contig". Two BACs (CR392039 and CR381535) were selected to construct the proximal region representative sequence, and the overlap between these two clones shows 100% sequence identity. The overlapping region in BAC clone CR392039 was truncated before merging it with BAC clone CR381535 to obtain a 207,338 bp proximal flanking region sequence (Figure 3.8b) that was named the "proximal contig". To identify the positions of the cosmids and BAC clones in these flanking region contigs, all the cosmids and BAC clones sequences were mapped to the appropriate contig (Figure 3.9 and Figure 3.10).

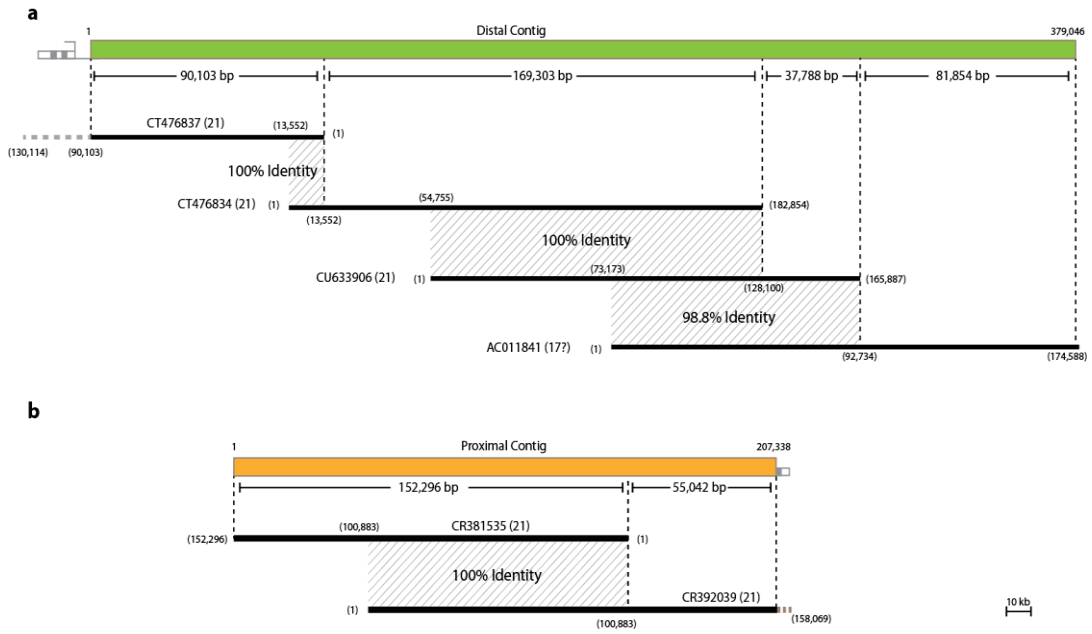


Figure 3.8: Scheme to construct the distal and proximal contigs.

a) *The distal contig (green) was constructed by merging four BAC clones (solid black lines) from two different chromosomes to obtain 379 kb of the distal flanking region. The clone names and chromosomes of origin (in parentheses) are indicated.* ***b)*** *The proximal contig (orange) was constructed by merging two BAC clones (solid black lines) from the same chromosome to obtain 207 kb of the proximal flanking region. The clone names and chromosome of origin (in parentheses) are indicated. The percent identities of the overlapping regions (shaded areas) between BAC clones are also shown. The positions (dotted black lines) and lengths (immediately below the schematic contig) of BAC clone fragments used for constructing the contigs are shown. Where BAC clones overlap, the overlapping region in the lower BAC clone was trimmed for the merger.*

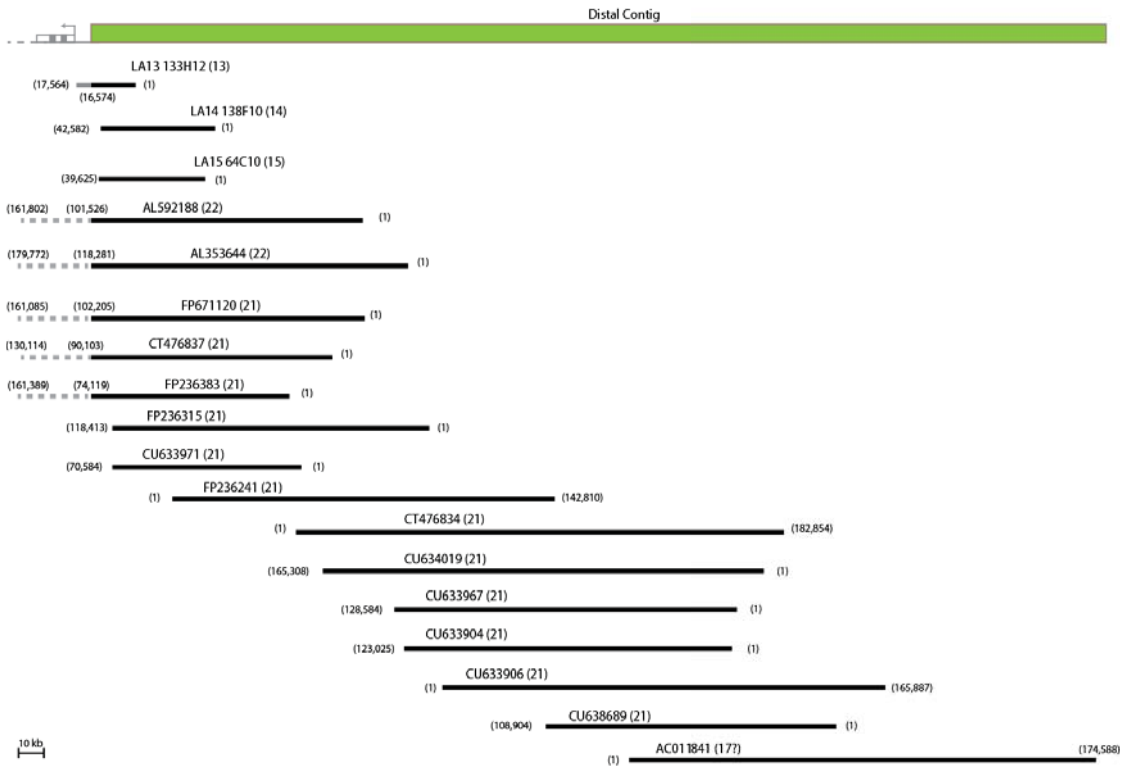


Figure 3.9: Locations of the distal region clones in the distal contig.

The cosmid and BAC sequences (black lines) from the distal region were mapped to the distal contig (green) to show their locations. The clone names with chromosome of origin in parentheses are indicated above the sequences. The rDNA unit next to the distal junction position is represented by grey dotted lines. The sequence coordinates, including the distal junction positions of the cosmids and BAC clones, are indicated in parentheses next to the sequences.

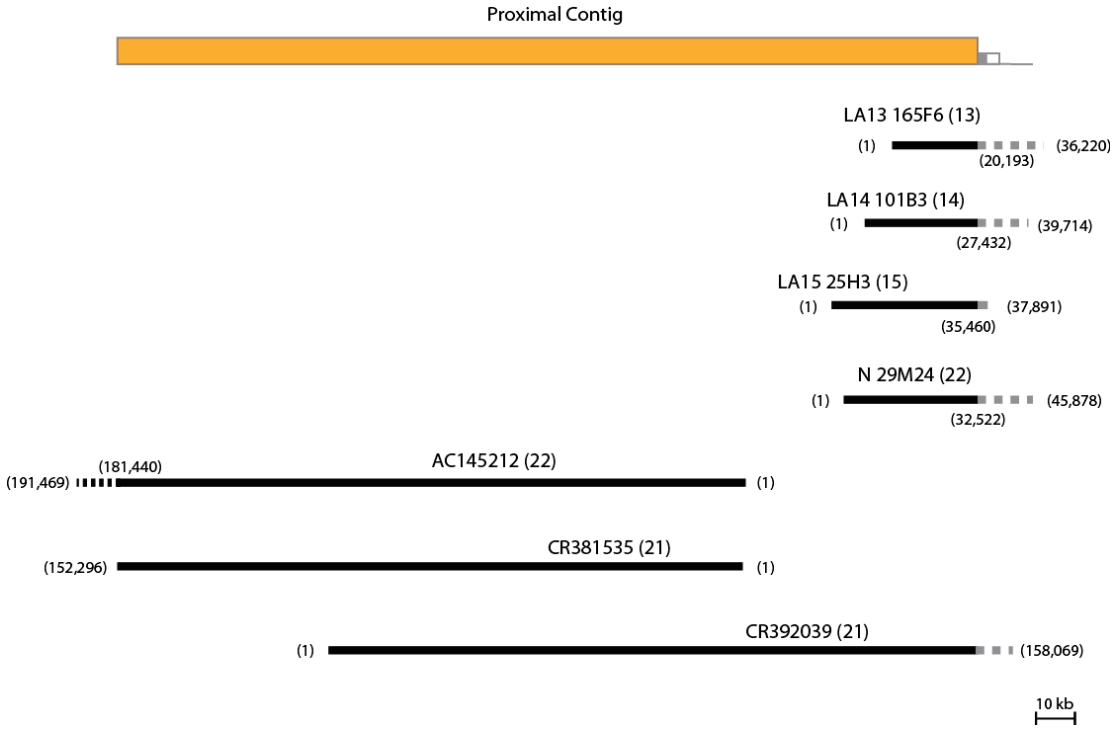


Figure 3.10: Locations of the proximal region clones in the proximal contig.

The cosmid and BAC sequences (black lines) from the proximal region were mapped to the proximal contig (orange) to show their locations. The clone names with chromosome of origin in parentheses are indicated above the sequences. The rDNA unit next to the proximal junction position is represented by grey dotted lines. The sequence coordinates, including the proximal junction positions of the cosmids and BAC clones, are indicated in parentheses next to the sequences.

3.3.4. Characterization of the distal and proximal contigs

3.3.4.1. Repeat content of the distal and proximal contigs

The rDNA flanking regions have been assumed to be heterochromatic and mainly composed of repeat elements (Lander *et al.* 2001). To address this, I decided to determine the repeat composition of the distal and proximal contigs. The tool RepeatMasker, which identifies repeats in a given sequence by comparing it with an annotated database of repeat sequences, was used to search for repeats in the distal and proximal contigs (Section 3.2.5). The total repeat content for the distal and proximal contigs is 64.4% and 54.4% respectively (Figure 3.11; Appendix Table 6-Appendix Table 7) which is slightly higher than the total average repeat content of the entire human genome assembly hg19 (50.6%). The higher repeat content of the distal contig compared to the proximal contig is primarily due to the presence of a large 38,596 bp block made of ~800 copies of 48 bp satellite repeats known as CER

satellite at the far end of the distal contig (Figure 3.12). The repeat content of the distal contig drops to 60.5% if this satellite region is removed.

Excluding the large satellite region in the distal contig, the repeat content in the proximal and distal contigs is broadly comparable with that of the whole human genome (Figure 3.11). The percentage of retrotransposon elements i.e. LINEs, SINEs and LTRs in the distal and proximal contigs are variable compare to the whole genome average. The percentage of LINE elements in the distal contig is higher than the rest of the genome, and the distal contig LINE elements are mainly LINE-1 (L1) elements. Both the distal and the proximal contigs have fewer SINE elements, or more specifically Alu elements, than the rest of the genome, while the percentage of LTR elements in both contigs is higher than the rest of the genome. Interestingly, unlike these retrotransposon elements, DNA transposons are absent in distal contig while the proximal contig resembles the rest of the genome. The distal contig has a comparatively higher proportion of satellites than the genome as a whole because of CER satellite block at the end of the contig. However, the satellite percentage of the distal contig is almost same as whole genome if CER satellite blocks are ignored (Figure 3.11).

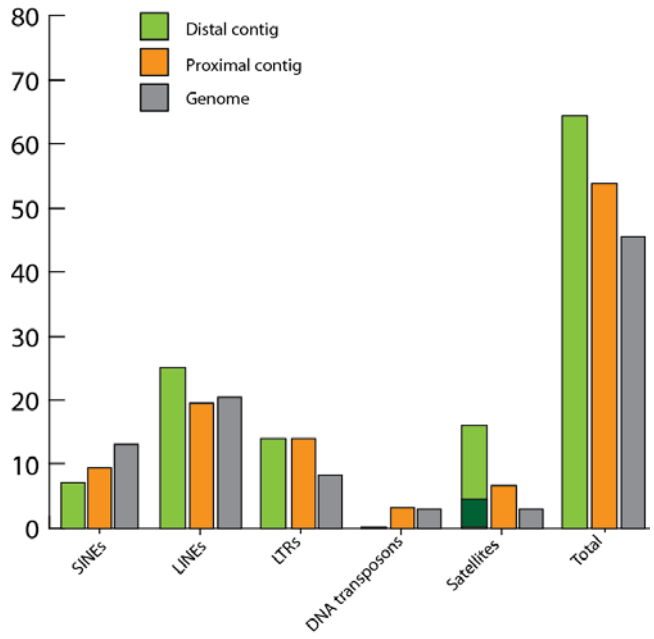


Figure 3.11: Repeat composition of the distal and proximal contigs.

The percentage of the distal contig, proximal contig and total human genome occupied by different classes of repeat element, as well as total repeat content, is plotted. The overlapping dark green bar in the "Satellites" plot represents the satellite content of the distal contig after removing the CER satellite block.

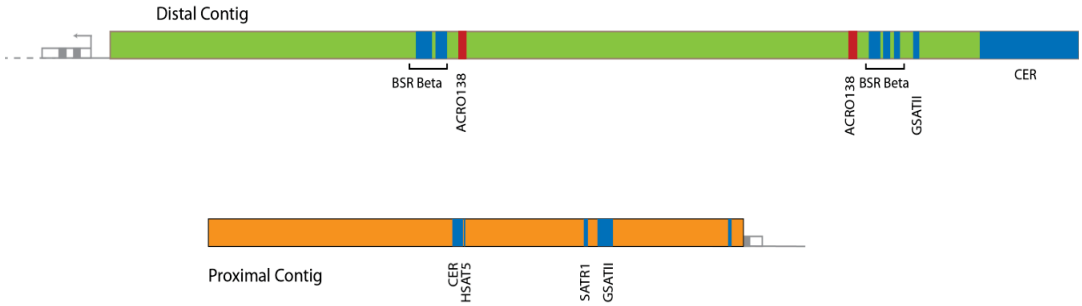


Figure 3.12: Locations of novel and satellite repeats in the distal and proximal contigs.

RepeatMasker identified several satellite repeats (blue boxes) in the distal contig (green) and the proximal contig (orange). A novel 138 bp repeat, ACRO138 (red block), was identified in the distal contig.

Novel ACRO138 repeat: A novel 138 bp repeat was identified in the distal contig by the tandem repeat search (Section 3.2.5) using TRF and BLAST (Figure 3.13). Based on its position on the acrocentric chromosome and the repeat unit length the repeat, I named the repeat ACRO138 (Figure 3.12). The distal contig has two blocks of ACRO138 repeats (Figure 3.12). The first block consists of 18 repeat units and is located at position 136,616-139,228, while the second block consists of 19 repeat units and is located at position 289,297-292,203. To check if ACRO138 is restricted to the distal contig or is present in the other regions of the human genome I searched the current human genome assembly with the ACRO138 consensus sequence using BLAST. The search shows matches for ACRO138 in chromosomes 3, 4, 7, 10 and 19, indicating it is not restricted to acrocentric chromosomes. In chromosome 3, 4, and 10 the repeat is present at the 5' end of the *FRG*2 gene paralogs, while in chr 19 it is present at both 5' and 3' ends of zinc finger proteins. The function of the FRG2 protein is not known but it is found to be highly expressed in muscle fibres with Facioscapulohumeral muscular dystrophy (Gabellini *et al.* 2006).

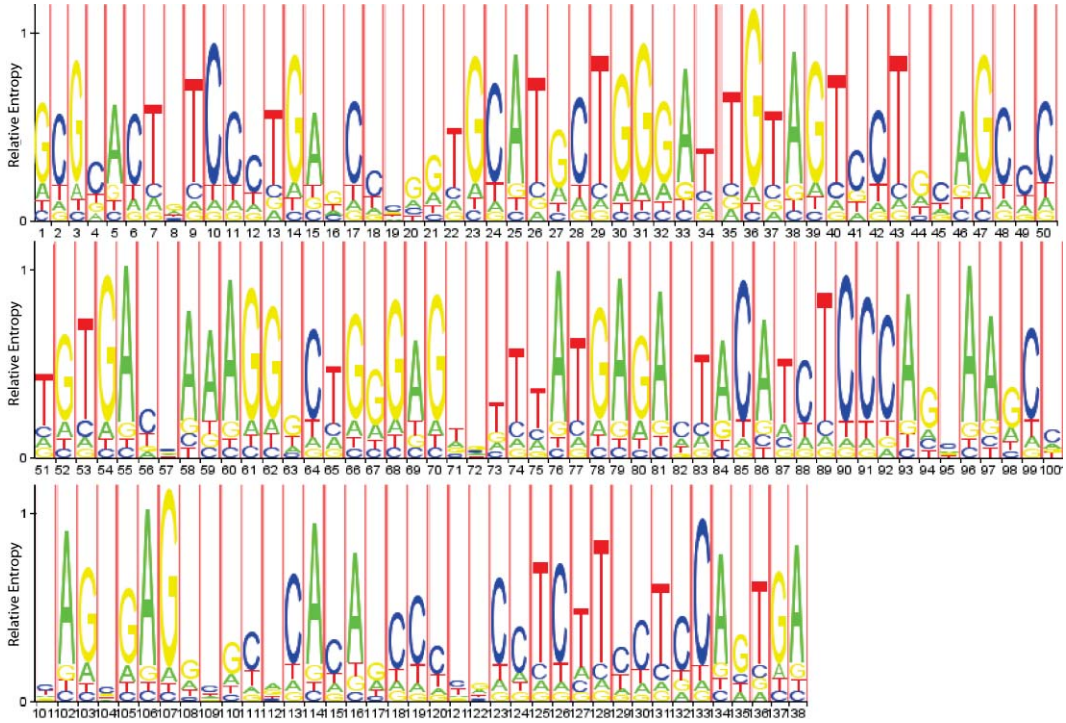


Figure 3.13: HMM logo for the 138 bp ACRO138 repeats.

The consensus sequence of the ACRO138 repeats was obtained using a multiple sequence alignment of the 37 repeat units present in the distal contig. An HMM logo (Schuster-Bockler et al. 2004) was constructed from this alignment. The height of each base represents the probability of the presence of that base at that specific position. The thickness of dark red line represent the probability of insertion of one base and thickness of the dark and light red vertical lines together on the right side of each base shows the number of bases expected to be inserted at that position.

3.3.4.2. A large inverted repeat in the distal contig

A large inverted repeat was found in the distal contig (Figure 3.14). It is 227,965 bp in length, with each arm being ~109 kb and the intervening region between the arms being 5,762 bp (Figure 3.14.a). The arms show regions of high sequence identity separated by more diverged regions. The average sequence identity between the two arms is 79.5% (Figure 3.14.b). The presence of the large inverted repeat on the distal contig accounts for the duplication of ACRO138 repeats and a 136 bp satellite named BSR/beta in the distal contig (Figure 3.12).

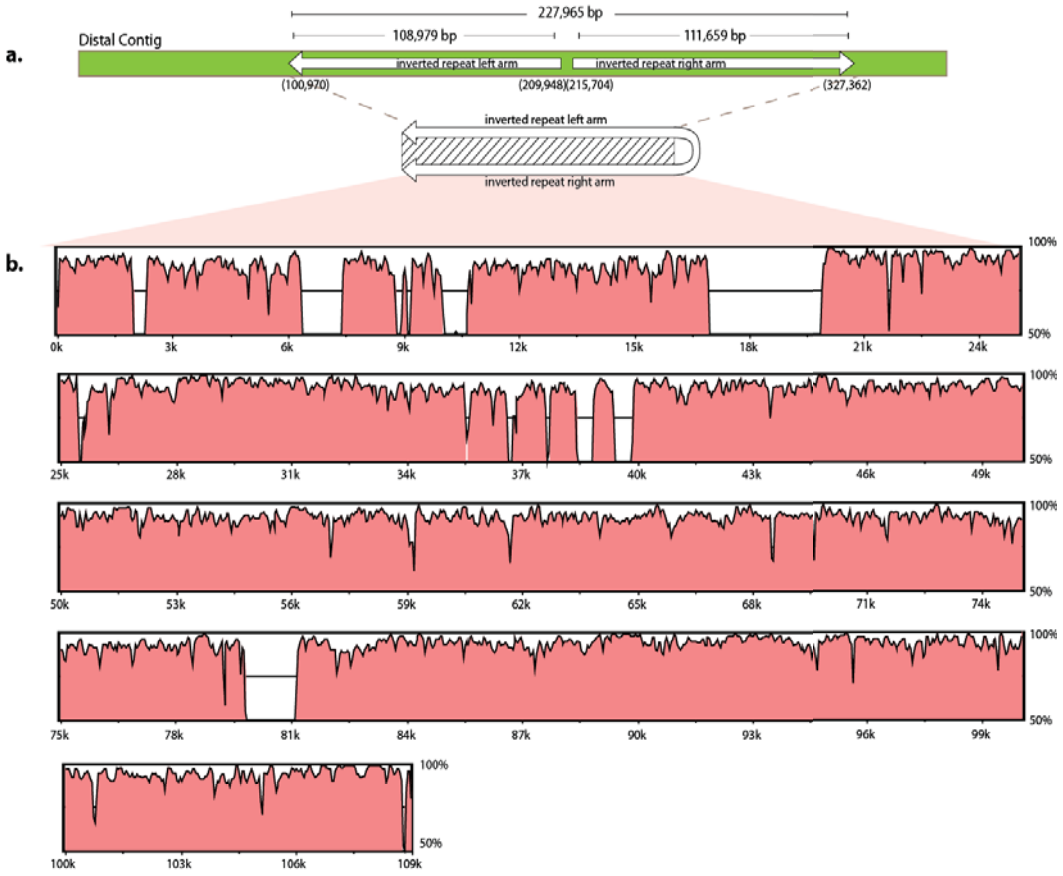


Figure 3.14: Sequence similarity between the arms of the large inverted repeat in the distal contig.

a) A large inverted repeat (solid white arrows) is present in the distal contig (green). The length of each arm is indicated just above the arms and the position in the distal contig is given in parentheses below the arrow. The arms of the inverted repeat are separated by ~5 kb. b) A similarity plot showing the level of sequence identity between the inverted repeat arms was obtained using a 100 bp sliding window. The horizontal axis represents the alignment position and the vertical axis represents the percent identity between the arms. Regions with >50% similarity between the arms are shaded in pink.

3.3.4.3. The level of segmental duplication in the proximal and distal contigs

Segmental duplicates are large fragments of DNA that are present in more than one region of the genome. Segmental duplication is a prominent feature of the human genome and ~5% of the total genome is known to be segmentally duplicated (Bailey *et al.* 2002). To search for segmental duplicates from the rDNA distal and proximal contigs, I used a modified version of the BLAST-based technique called the "Whole Genome Assembly Comparison" (Section

3.2.6). A number of segmental duplicates (>1 kb in length and with >85% sequence identity) were found from both contigs (Figure 3.15; Appendix Table 3-4). Interestingly, however, the distal and proximal contigs show contrasting patterns of segmental duplication (Table 3.10). Proximal segmental duplicates are more frequent, longer, and have greater sequence identity than distal segmental duplicates (Figure 3.15). Furthermore, the majority of proximal segmental duplicates are found in pericentromeric regions of the genome as previously observed (Piccini et al. 2001; Lyle et al. 2007), while the majority of distal segmental duplicates are found in subtelomeric regions (Figure 3.15). Most strikingly, the level of segmental duplicated DNA is vastly different, with 7.3% of the distal contig being segmentally duplicated while 92.4% of the proximal contig is segmentally duplicated (Table 3.10). Therefore, only 7.6% of the proximal contig consists of sequence that is not duplicated elsewhere in the genome. These results demonstrate that the rDNA distal and proximal contigs have different genomic characteristics in humans: the segmental duplication profile of the proximal resembles pericentromeric regions, while that of the distal region resembles subtelomeric regions.

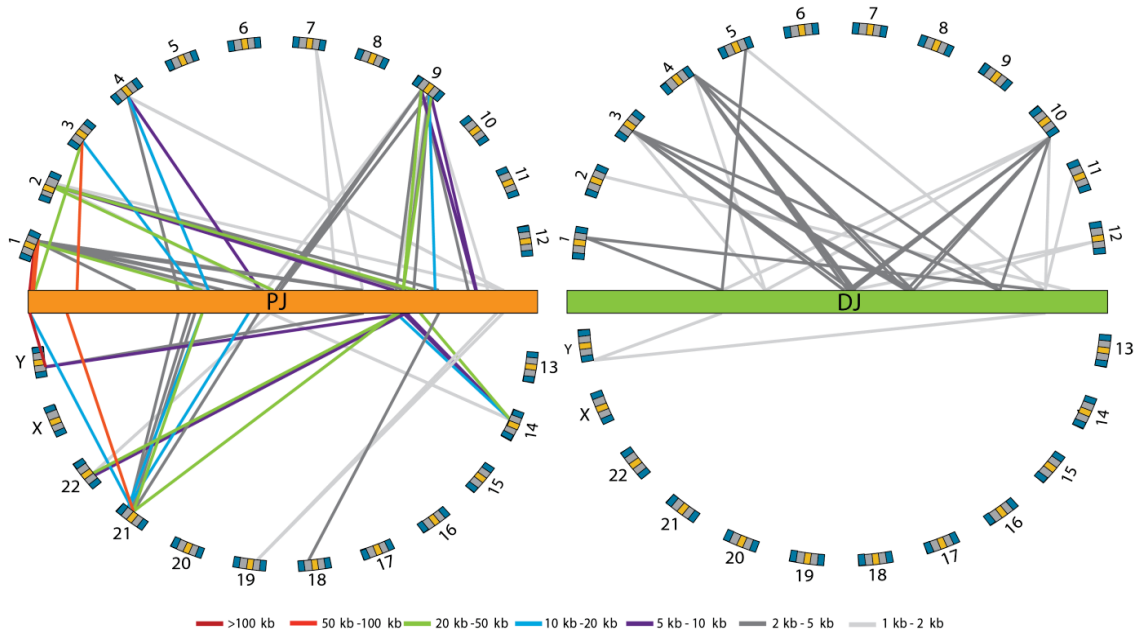


Figure 3.15: Segmental duplication in the proximal and distal contigs.

The proximal contig (orange) is highly segmentally duplicated compared to the distal contig (green). The proximal contig has segmental duplicates (coloured lines) with 12 chromosomes while the distal contig has segmental duplicates with 9 chromosomes. The chromosomes are represented as grey boxes arranged at the periphery around the contigs with centromeres as yellow boxes and telomeres as blue boxes. The segmental duplicates from the proximal contig are mainly to pericentromeric regions of chromosomes, while those from the distal contig are mainly to subtelomeric regions. The positions of the segmental duplicate lines on the contigs represent the start positions of the segmental duplicates, while the colour of segmental duplicate lines represents the length of the segmental duplicate, as indicated below the figure.

Table 3.10: Segmental duplication comparison between the distal and proximal contigs.

Segmental duplication feature	Distal contig	Proximal contig
Number of segmentally duplicated regions	31	98
Average segmental duplicate length (kb)	2.3	11.8
Average percent identity between segmental duplicates	88.2%	93.1%
Position of duplicate region	subtelomeric	pericentromeric
Percent of contig duplicated	7.3%	92.4%

In the proximal contig, there is variation in the level of duplication within the contig. The near proximal region (closer to the rDNA proximal junction position) harbours less segmental duplicates compare to the far proximal region (closer to the centromere). Seventy-

two segmental duplicates are from the far-proximal region, while 19 segmental duplicates are from the near-proximal region. Seven segmental duplicates span these two regions. In some cases, the same region of the proximal contig is segmentally duplicated more than once to the same chromosome. These duplicated segmental duplicates have different sequence identities. For example, the region of the proximal contig covering coordinates 1-73,523 is segmentally duplicated thrice to chr 1 with sequence identities of 96.5%, 94.6% and 92.6%. This variability in sequence identity suggests that the sequence exchanges that result in these segmental duplication events did not occur in one burst, but are result of multiple events. In the distal contig, the segmental duplicated have palindromic symmetry because of the presence of the large inverted repeat. Twelve segmental duplicates present in the left arm of the inverted repeat also have an inverted paired segmental duplicate in the right arm of the inverted repeat.

3.3.4.4. Putative gene models in the distal and proximal contigs

To determine whether the rDNA distal and proximal contigs potentially contain genes, I undertook a gene search. I used a four-step gene prediction pipeline that integrates *de novo* gene prediction tools with EST, mRNA and protein evidence to predict putative genes in the contigs (Section 3.2.7). First, putative genes were predicted using repeat masked sequences of the distal and proximal contigs. For the distal contig, the gene prediction pipeline predicted five putative genes (Figure 3.16). All these putative genes except one are single exon genes (Table 3.11). For the proximal contig, four putative genes were predicted (Figure 3.16). Two of these gene models are multi-exonic and the other two have only a single exon (Table 3.12). Recent studies have shown that repeat elements can also be parts of genes and can alter gene splicing (Huh and Hynes 1994; Rohlfs *et al.* 2000; Sorek *et al.* 2002). To check if the repeat elements in these contigs may also form a part of the genes, genes were predicted from the distal and proximal contigs using unmasked sequences. For the distal contig eight putative genes were predicted, and for the proximal contig six putative genes were predicted (Figure 3.16). Five of the putative genes in the unmasked distal contig are the same as the five from the masked distal contig. The remaining three distal contig putative genes from the unmasked analysis are single exon. Similar to the distal contig, all the putative genes from the masked proximal contigs were a subset of the unmasked proximal contig putative genes. The remaining two putative proximal genes specific to the unmasked analysis are located in intronic regions of another putative (pg1.1) gene as alternative transcripts. These two putative genes are single exon.

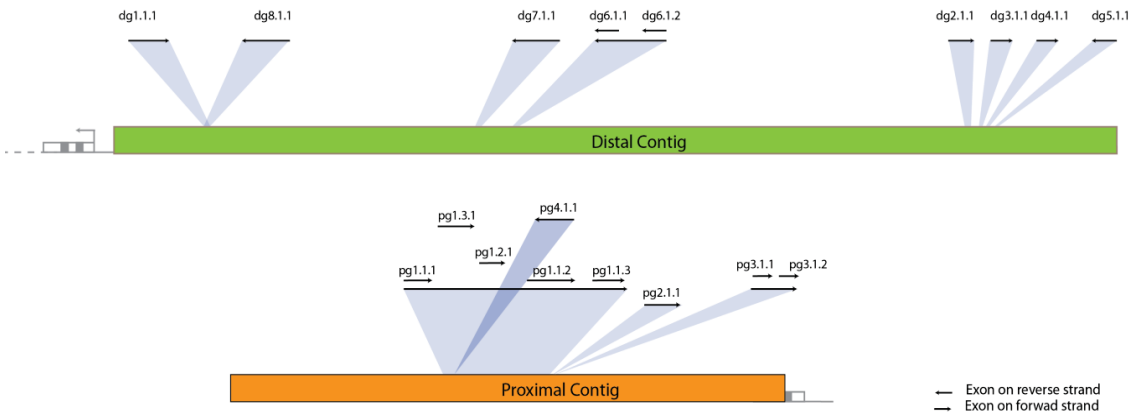


Figure 3.16: Gene models in the distal and proximal contigs

Gene models (black arrows) in the distal contig (green) and proximal contig (orange) were identified using a four-step gene prediction pipeline, and blue shading shows their positions. The positions of exons in the multi exon and alternative transcript gene models are indicated by small arrows above the main gene model arrow. The gene models in the distal contig were labelled as "dg" and in the proximal contig as "pg" followed by gene, transcript and exon numbers. The different proximal region gene models pg1.1, pg1.2 and pg1.3 in the pg1 gene region are represented as tiers of arrows.

Table 3.11: Putative gene models from the distal contig

Gene model number	Start coordinate	End coordinate	Length	Strand
dg1.1.1	33,924	34,145	221	+
dg2.1.1*	321,828	322,277	449	+
dg3.1.1*	327,412	327,705	293	+
dg4.1.1	330,230	330,427	197	+
dg5.1.1	333,297	332,554	743	-
dg6.1.1*	150,885	150,689	196	-
dg6.1.2*	155,079	154,917	162	-
dg7.1.1*	137,251	136,850	401	-
dg8.1.1	33.974	33,600	374	-

* predicted in both masked and unmasked sequences

Table 3.12: Putative gene models from the proximal contig

Gene model number	Start coordinate	End coordinate	Length	Strand
pg1.1.1*	68,543	68,599	57	+
pg1.1.2*	125,644	125,874	231	+
pg1.1.3*	126,489	126,590	102	+
pg1.2.1	112,302	113,033	732	+
pg1.3.1	119,489	119,629	141	+
pg2.1.1*	127,300	127,674	375	+
pg3.1.1*	159,642	159,861	220	+
pg3.1.2*	159,946	160,145	200	+
pg4.1.1*	107,900	107,694	205	-

* predicted in both masked and unmasked sequences

3.4. Discussion

In this work, I have characterized genomic regions that represent a major gap in the human genome assembly, the regions flanking the rDNA arrays. This analysis identifies various features associated with these distal and proximal flanking regions, as well as the contrasting nature of the two regions. The distal region has a single rDNA junction position, a large inverted repeat, a novel repeat element (ACRO138), a large block of CER satellite repeats at one end, and a small percentage of sequence that is segmentally duplicated. In contrast, the rDNA junction point in the proximal region appears to vary between the acrocentric chromosomes and possibly individuals, is very highly segmentally duplicated, and does not have any specific repeat element pattern. The gene prediction pipeline shows that both flanking regions have gene coding potential, contrary to previous assumptions that the short arms of the acrocentric chromosomes are devoid of genes other than the rDNA and are mainly composed of repeat elements (Dunham *et al.* 1999; Hattori *et al.* 2000; Lander *et al.* 2001; Heilig *et al.* 2003; Dunham *et al.* 2004; Zody *et al.* 2006). The quantification of sequence identity between clones from these flanking regions reveals that both the proximal and the distal flanking regions are highly homogeneous among acrocentric chromosomes, with the distal region showing higher homogeneity than the proximal. Overall, this study demonstrates that the rDNA flanking regions are similar to the rest of the genome. The specific genome organizations exhibited by these regions suggest that they might play functional roles in the nucleolus.

3.4.1. High conservation of the flanking regions across the acrocentric chromosomes

The quantification of the sequence conservation in the distal and proximal flanking regions reveals that homogenisation in the short arms of the acrocentric chromosomes is not restricted to the rDNA array but is also extended further towards the centromere and the telomere in the flanking regions. The rDNA arrays are known to undergo mitotic recombination among the acrocentric chromosomes (Krystal *et al.* 1981). The high level of inter-chromosomal homogeneity in the rDNA flanking regions demonstrates that not only the rDNA array but also the region surrounding it exchange sequence during mitotic division. The higher level of inter-chromosomal sequence identity and low level of segmental duplication found in the distal region compared to the proximal region suggests that most of the recombination based exchanges with the distal regions are occurring among the acrocentric chromosomes. The subtelomeric regions are known to have a higher propensity to undergo the inter-chromosomal translocations (Linardopoulou *et al.* 2005), and this may explain the homogenization of the distal region. In addition to the high level of inter-chromosomal homogeneity, the proximal region also has a high level of segmental duplication. This suggests that the proximal region actively exchanges sequences with other chromosomes in addition to the acrocentric chromosomes.

3.4.2. The proximal region is a segmental duplication hub

A prominent feature of the rDNA proximal region is the high level of inter-chromosomal segmental duplication. The majority of the proximal segmental duplicates are with chr 1, chr 2, chr 3, chr 4, chr 9 and chr Y. Previous studies have shown that chr 1, chr 4, chr 9 and chr Y are colocalized near to the nucleolus and form part of the outer heterochromatic shell of the nucleolus (Manuelidis and Borden 1988; Léger *et al.* 1994). Therefore, it is possible that the close proximity of the proximal region to these chromosomes facilitates recombination with them, resulting in the high level of segmental duplication with these chromosomes.

The proximal region also shows intra-chromosomal segmental duplications with acrocentric chromosomes 14, 21 and 22. A ~200 kb sequence from the pericentromeric region of the short arm of the chr 21 has been included in the reference genome (Hattori *et al.* 2000). However, for the remaining acrocentric chromosomes the short arms are completely absent from the reference human genome sequence. The intra-chromosomal segmental duplicates in chr 14 and chr 22 are with the long arm while for chr 21 they are with the short arm. The variation in the location of segmental duplicates among the acrocentric chromosomes from the short arms is likely to be because of the difference in the status of completeness of short

arm sequence among the acrocentric chromosomes in the reference genome. Therefore, it is possible that intra-chromosomal proximal region segmental duplicates with the short arms of the other acrocentric chromosomes also exist but are not detectable because of this lack of sequence information. However, contrary to the short arms, the long arms of the acrocentric chromosomes are more or less complete. The variations in the presence of segmental duplicates among long arms of the acrocentric chromosomes shows that the arms of the acrocentric chromosomes interact differently depending on the chromosomes. The reason behind the variation in interaction is not clear.

Intra-chromosomal segmental duplication is known to play role in pericentric inversion (reversal of chromosomal segment that involves break in the both arms of chromosomes and the inverted segment includes the centromere) by promoting intra-chromosomal recombination (Goidts *et al.* 2004; Kehrer-Sawatzki *et al.* 2005). Monosomy of chr 22 caused by pericentric inversion between the short and long arms of the chromosome is known to result in the familial monosomy 22 syndrome characterized by mental and physical retardation (Watt *et al.* 1985). The mechanism behind this pericentric inversion is still not known. One possibility is that intra-chromosomal segmental duplication of the proximal region with the long arm of the chr 22 promotes pericentromeric inversion.

3.4.3. Putative genes in the masked and unmasked distal and proximal regions

A number of putative genes were predicted by the *ab initio* gene prediction pipeline, indicating that the distal and proximal regions have gene coding potential. All these predicted putative genes have homology with putative proteins identified by other sources based on cDNA isolated from cell extracts (Ota *et al.* 2004). Lyle *et al.* (2007) reported several putative genes in the proximal region that are potentially transcribed. The high level of sequence conservation of the flanking regions among the acrocentric chromosomes suggests that these putative genes are probably present in all the acrocentric chromosomes. The gene prediction analysis of the flanking regions by including (unmasked) and excluding (masked) the repeats has shown variation in the gene content of the distal and the proximal regions because of repeat elements. This suggests that certain repeat elements likely to be part of the genes. Previous studies have reported the presence of repeat elements in the exons of some genes (Shapiro and von Sternberg 2005). All the genes present in the masked sequences are also present in the unmasked sequences, demonstrating that the presence of these putative genes is not affected by the repeat elements. The gene dg5.1.1 (332,554-333,297) is part of a complete LINE L1 ORF2 (position 328,444-334,470). The LINE ORF2 contains genes for reverse transcriptase and the endonuclease that facilitate LINE

duplication. Clearly, the existence and function of these putative proteins require further experimental validation.

3.4.4. Significance of inverted repeat in the distal region

The presence of a large inverted repeat in the distal contig is another striking property of this region. Initially large palindromes were thought to be a characteristic feature of chr X and chr Y (Skaletsky *et al.* 2003). Later a genome wide search demonstrated that large inverted repeats are also present on chr 1, chr 2, chr 6, chr 7, chr 11, chr 13, chr 15, chr 17 and chr 22 (Warburton *et al.* 2004), showing that large inverted repeats are spread genome wide. The role of these large inverted repeats is still not clear. Unlike small inverted repeats (with arms 5-200 bp) that are known to play role in several biological processes, such as origins of replication, regulation of gene expression, nucleosome structure, and recombination (Pearson *et al.* 1996), the importance of large inverted repeats in the genome is still not explored in detail.

Large inverted repeats have been proposed to be involved in the elimination of deleterious mutations in the genomic regions that do not undergo meiotic recombination by facilitating gene conversion (Betran *et al.* 2012). The role of inverted repeats in sequence maintenance is most widely studied in human chr Y, where several large inverted repeats are present (Skaletsky *et al.* 2003). The absence of a homologous chromosome in males means that most of chr Y does not undergo meiotic recombination except a small pseudoautosomal region that is homologous to chr X. Theoretically, in the absence of an efficient mechanism to purge deleterious mutations, chr Y should degenerate very rapidly. However, since the split of chimpanzee and human, no genes have been lost from human chr Y (Hughes *et al.* 2005; Hughes *et al.* 2010). This chr Y sequence integrity has been proposed as result of non-allelic homologous recombination (NAHR) (recombination between two DNA segments of high identity that are not at the same position on the chromosome) between the arms of the large inverted repeats present in the Y chromosome (Rozen *et al.* 2003; Skaletsky *et al.* 2003; Marais *et al.* 2010; Scott *et al.* 2010; Betran *et al.* 2012). The repeat arms of inverted repeats >99% sequence identity. Therefore, it is proposed that in the event of occurrence of a deleterious mutation, it is repaired by recombination between the arms of the inverted repeats. Supporting this, two recent simulation studies have shown that gene conversion due to the presence of inverted repeats is sufficient to preserve the genomic integrity of chr Y (Connallon and Clark 2010; Marais *et al.* 2010). Meiotic recombination is known to be suppressed in the rDNA (Petes and Botstein 1977; Høgset and Øyen 1984), and it is possible that this suppression extends into the flanking region. If so, the large inverted repeat in the

distal region may be playing a similar role in allowing recombination events that remove deleterious mutations from the region and maintain its integrity.

NAHR also causes deletion of large genomic segments and plays a role in the development of several genetic disorders (Barbouti *et al.* 2004; Carvalho and Lupski 2008; Scott *et al.* 2010). For example, partial monosomy of the short arm of chr X results from deletion of a 5 Mb region in Xp11.1-p11.22, and this is promoted by inverted repeats in the region (Scott *et al.* 2010). The human genetic disease, Turner syndrome, can result from this partial monosomy (Scott *et al.* 2010). Therefore, it is possible that the inverted repeat in the rDNA distal region will play a role in the partial monosomy of the acrocentric chromosomes. Partial monosomy of the short arms of the acrocentric chromosomes has been associated with human diseases e.g. Down syndrome and cancer (Muleris *et al.* 1984; Tschentscher *et al.* 2001; Lyle *et al.* 2008). Therefore, the role of this distal region inverted repeat needs to be studied further.

The large inverted repeat in the distal region causes duplication of the RNA transcribing regions in the distal region. A ChIP-seq mapping analysis using data from the ENCODE project has shown that the histone modifications H3K4me2/3, H3K9ac, and H3K27ac, and the transcription factors CTCF and Pol II that are associated with active promoters, have a mirror symmetry as a result of the inverted repeat (Floutsakou *et al.* 2013). These promoters were found to be associated with active transcription using RT-PCR and by mapping transcripts from RNA-seq data obtained from the ENCODE project (Floutsakou *et al.* 2013).

3.4.5. The flanking regions have a repeat content similar to the rest of the genome

Heterochromatic regions usually consist of a high percentage of repeat elements (Lander *et al.* 2001). The rDNA flanking regions were thought to be heterochromatic, repeat-rich regions (Lander *et al.* 2001). However here I show that the rDNA distal and proximal regions have repeat contents that are in the same range (±5%) as the whole genome average. LINE elements have been found to be enriched in gene-poor regions of the genome, while SINE elements are enriched in gene-rich regions (Versteeg *et al.* 2003). The distal region has a higher LINE and a lower SINE content compared to the genome-wide average, and is similar to heterochromatic regions of the genome. In the proximal region, a lower SINE content but an approximately equal LINE content compared to the genome-wide average suggests that this region is similar to the euchromatic regions of the genome. Interestingly, this is opposite to the segmental duplication results, where the proximal region shows a greater level of segmental duplication with heterochromatic regions of the genome than the

distal region, suggesting that the proximal region is associating with other heterochromatic regions. However, the role of repeat elements varies depending on their location (Shapiro and von Sternberg 2005), therefore further work is required to correlate the repeat element the flanking regions to explore their role.

A prominent feature of the distal region is the presence of a large block of 48 bp CER satellite repeats. The presence of the CER satellites in human on the short arm of the acrocentric chromosomes has been reported previously (Metzdorf *et al.* 1988). Further, these CER satellites are also present in the apes (chimpanzee, gorilla, orangutan and gibbon) but not in the monkeys (Metzdorf *et al.* 1988; Pusch *et al.* 2002). The CER satellites in chimpanzee and gorilla are present on all the rDNA containing chromosomes (Metzdorf *et al.* 1988). The role of this satellite element in the distal region is not known. The CER satellites in the centromeric location are present near to alpha-satellites (Metzdorf *et al.* 1988; Pusch *et al.* 2002). The alpha satellite at the centromeres is known to be associated with the centromeric protein–G (CENP-G). Initially it was thought that CENP-G associated with the alpha-satellite repeats (He *et al.* 1998), but it has been shown to also be present on the chr Y centromere that lacks alpha-satellite, suggesting that CENP-G associated to other repeat elements together with alpha satellites (Gimelli *et al.* 2000). A gel mobility shift assay has shown that the CER satellite has the ability to bind with a protein of molecular mass 95 kDa (Pusch *et al.* 2002). The CENP-G has the same molecular mass, therefore it can be speculated that it may be associated to the CER satellite. The location of the CER satellites near to alpha-satellite repeats in centromeric regions makes them a potential candidate for CENP-G binding. CENP-G is known to bind at neocentromeres and inactive centromeres (Gimelli *et al.* 2000) and is thought to play role in the formation and stabilization of kinetochore (Gimelli *et al.* 2000). The distal flanking region is present on the periphery of the nucleolus (Floutsakou *et al.* 2013). It is possible that if CENP-G is associated with CER satellite, the satellite block at the end of the distal region may have a role anchoring acrocentric chromosomes to the periphery of nucleoli, similar to the role of the kinetochore in anchoring chromosomes to the metaphase plate (Rieder and Salmon 1998). Further, experiments will be required to establish the binding of CENP-G to CER satellite repeat.

3.4.6. The flanking regions boundary

The BAC clones identified from the rDNA flanking regions (Section 3.1.2) revealed the sequence composition of these regions. However, to precisely establish the position of the NORs one question still needs to be answered: what are the boundaries of the rDNA flanking regions on both the proximal and the distal ends of the rDNA array that can still be called NORs. Comparing the definition of the NOR proposed in this thesis based on its genomic

content with the definition given by Barbara McClintock based on the function associated with the NOR, the boundary of the distal and the proximal rDNA flanking regions is at the extremity of both regions that are part of the nucleolus. One possible way to identify the rDNA flanking region boundaries is by isolating intact nucleoli from the cells and then analysing the sequence content of the chromosomal region that forms this nucleolar structure. However, considering the limitations associated with isolating intact nucleoli, it will be difficult to demarcate the rDNA flanking regions using this method.

3.4.7. The sequence of the short arm of acrocentric chromosomes beyond the identified rDNA flanking region sequences

The short arms of the five acrocentric chromosomes represent one of the largest gaps in the human genome assembly. This study has characterized 536.3 kb (207.3 kb of the proximal flanking region and 379.0 kb of the distal flanking region) of the short arms of the acrocentric chromosomes that is not part of the current assembly. Since the length of the short arms of the acrocentric chromosomes is not known, it is not possible to estimate what percent of the region still needs to be identified, or in other words, how large the gaps between the identified regions and the centromeric/telomeric ends of the short arms are. Further, it is difficult to predict the sequence content of the remaining gaps in the short arms. Considering the sequence composition of the known regions, the gaps may have unique, highly segmentally duplicated, or highly repetitive sequences. Reports of repeats in the short arms of the acrocentric chromosomes may provide a means for these regions to be further extended. For example, Lyle *et al.* (Lyle *et al.* 1995) have reported blocks of D4Z4 repeat in the distal flanking region. However, this repeat is not present in the distal contig identified here therefore, it is likely to be present in the region further towards the telomere.

3.4.8. Computational challenges to study the rDNA flanking regions.

To characterize the rDNA flanking region several standard computational tools and published pipelines were employed. The biggest challenge in characterizing the rDNA flanking regions was the lack of tools to perform the segmental duplication analysis, despite whole genome segmental duplication analysis being a part of several whole genome projects. Instead, most published work on the identification of segmental duplicates has been done using in-house scripts (Bailey *et al.* 2001b; Scally *et al.* 2012). A part of the segmental duplication analysis of the rDNA flanking regions was performed by scripts provided by Jeff Bailey. However, the script was designed to perform whole genome segmental duplication analyses and therefore I had to adapt it for use on small genomic segments. Many of the modules and parameters used in the script were optimal for very large genomic segments and

therefore several steps were performed manually. It was feasible to perform parts of analysis manually for the rDNA flanking regions as this only encompassed two genomic regions, but the availability of a tool for performing segmental duplication analysis of smaller genomic segments would have reduced the experimental time considerably. Such a tool would also be of use for performing whole genome segmental duplication analysis of genomic sequences that fill gaps in previously available genome assemblies, and the general availability of a segmental duplication tool would aid in the analysis of novel genome assemblies.

This study has shown that, in contrast to the previous perception, the rDNA flanking regions are not composed of simple repeat sequences but instead have a very complex genomic structure. The study has revealed that the rDNA flanking regions have several genomic features that can potentially play roles in different biological processes. The short arms of the acrocentric chromosomes represent one of the most biologically critical gaps in the human genome assembly. The characterization of the flanking regions presented here are a step closer towards filling these gaps.

[Blank Page]

Chapter 4

Conclusions and Future Directions

[Blank Page]

4.1. Conclusions

This study characterizes the two different components of the human nucleolar organizer regions (NORs): the intergenic spacer (IGS) and the rDNA flanking regions. I have combined the results of different computational approaches to identify properties and regions of interest. The two major outcomes of this work are:

1) I have identified a number of conserved regions in the human IGS that may be functional, and have explored their potential involvement in transcription, transcription regulation, cell cycle regulation and replication.
2) The human rDNA flanking regions has a complex genomic structure and contain various genomic features that are likely to play role in different biological processes.

4.1.1. The functional regions in the human IGS

Phylogenetic footprinting is a powerful tool to identify the functional regions that are conserved during evolutionary time. Using this technique, I have identified a number of potential functional regions in the human IGS. I have shown that these potential functional regions are not restricted to the promoter region but are dispersed throughout the IGS. In this study, I have demonstrated despite of high repeat content in the human IGS it contains several potential functional elements that correspond to both unique regions and the repeat elements. However, the potential functional regions identified and characterized in this study are those that are conserved, and therefore these are restricted to elements that have potential function across the primate species used here. This means there may be additional functional elements in the human IGS that are specific to humans or have evolved more recently in the IGS.

Several identified conserved regions in this study correspond to known functional elements in the IGS, which includes promoter, first three terminator, IGS transcripts and potential regulatory protein binding sites. On the contrary, the last two terminators that are known to be non-functional are not conserved. This strongly exhibits that the identified conserved regions does represents the functional regions.

The RNA-seq analysis using long poly(A)+ and poly(A)-, and small poly(A)+ from six cell types (three noncancerous cell types: GM12878, H1-hESC, and HUVEC; and three cancerous cell types: K562, HeLa-S3 and HepG2) from the ENCODE project revealed several transcripts from the IGS. Several transcripts identified in the analysis intersect conserved regions. These transcripts can be categorized into three different types of novel

transcripts: a cancer cell specific transcript (antisense to the pRNA); a tissue-specific transcript (small poly(A)+ transcript specifically found in the embryonic cell type H1-hESC); and transcripts present in multiple tissue types (transcript from the cdc27 pseudogene and other long polyA(-) transcripts). The biological significance of these transcripts is unknown. Based on their position in the IGS, I have proposed potential roles of three of the transcripts identified here. First, I proposed that the transcript antisense to the pRNA has a role in rDNA unit activation in cancer cells by pairing with pRNA and therefore preventing pRNA to silence the rDNA transcription (Section 2.4.1). Second, I have suggested that the transcript from the cdc27 pseudogene may have a role in regulating the *CDC27* gene that encodes a protein involved in cell cycle regulation (Section 2.4.2). Third, I have speculated that a small poly(A)+ transcript from a region around ~28.5-30.5 kb in the IGS that is transcribed only in the embryonic cell line, H1-hESC, of the cell lines used in this study may have some tissue specific function, perhaps related to the regulation of active rDNA unit number (Section 2.4.1). Overall, the RNA-seq analysis has demonstrated that a large portion of the human IGS is transcriptionally active, which mirrors the human genome as a whole (Dunham *et al.* 2012). The presence of transcripts in multiple cell types suggests that these transcripts are probably real, and their overlap with conserved regions suggests they may be functional. However, the hypotheses I have developed for their possible functions are based on the location of the noncoding transcripts in the IGS, and further experiments are required to verify these hypotheses. ChIP-seq analysis revealed that Pol II and Pol III machineries are associated with the IGS at specific positions (Section 2.4.3). The noncoding transcriptional activity of the human IGS has been assumed to be restricted to Pol I. However, association of Pol II and Pol III with the human IGS provide evidence that Pol II and possibly Pol III IGS may have a role in IGS transcription. A further investigation will be required to establish the role of Pol II in the human IGS transcription.

The chromatin profiling revealed three regions enriched in active histone modifications in the IGS corresponding to a region near the promoter (Cluster-1) and a second region in the middle of the IGS (Cluster-2), together with a region showing cell type specific enrichment in H1-hESC (Cluster-3). The different levels of enrichment of the active histone modifications in the regions suggest that they potentially act as promoter or enhancer. All three clusters coincide with the conserved regions suggesting that they may be functional. However, similar to RNA-seq results further experiments are required to verify the regulatory role of the regions. The ChIP-seq analysis using data for ORC from HeLa-S3 has identified three potential sites for the origin of replication. All three sites correspond to the conserved regions supporting them as potential site of origin of replication. Further, two sites of ORC enrichment correspond to previously reported potential sites of origin of replication

(Gencheva and Russev 1996) further supporting that the identified regions are probably functional.

All together, this study demonstrates that the human IGS consists of several potential functional regions that are conserved among the primates. These conserved regions potentially function as transcription regulators (promoter/enhancer), transcribed region and origin of replication. Further investigation will be required to establish the roles of the conserved regions. 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.

4.1.2. Characterization of the rDNA flanking regions

I have characterized a major gap in the human genome assembly that corresponds to the regions surrounding the rDNA array that I call the rDNA flanking regions. The paradigm that the rDNA flanking regions are made of only repeat elements and therefore are heterochromatic has been widely accepted for a long time. However, I have shown that the rDNA flanking regions do not just consist of repetitive elements but have complex genomic structures that are similar to euchromatic regions but also have unique characteristics. The study has revealed the contrasting nature of the distal and the proximal flanking regions. The proximal region is highly segmentally duplicated and shares sequences with a number of different chromosomes suggesting that this region is likely to be recombinationally active. The presence of high segmental duplication in the proximal region demonstrates that the sequence of the region is not unique to the proximal flanking region. The lack of sequence specific to the rDNA flanking region makes it unlikely that the proximal region has a nucleolar-specific role. Furthermore, the proximal region has different junction points in the rDNA array, suggesting that the proximal rDNA junction position tolerates variation. The distal region is highly conserved among the short arms of the acrocentric chromosomes and only a small percentage of this region is segmentally duplicated. The distal region has a large inverted repeat. The large inverted repeat is characteristic of the genomic regions that similar do not undergo meiotic recombination (Betran *et al.* 2012). The meiotic recombination is known to be repressed in the rDNA array and probably same nature is extended to the distal region. Another feature of the rDNA distal region is the presence of the CER satellite block. CER satellites are known to be present in the pericentromeric region of chromosomes and therefore its presence in the distal region was surprising. FISH experiments using the BAC clone from the distal flanking region have shown that the distal region is located at the

periphery of the nucleolus. I hypothesized that the CER satellite may have an association with the CENP-G protein and therefore, it is possible that the CER satellite forms a kinetochore-like structure that may act as anchor for the acrocentric chromosomes. However, this hypothesis is based on the similar molecular weight of CENP-G and a protein known to be associated with the CER satellite. Together these results suggest that the distal region has sequence features specific to the short arms of the acrocentric chromosomes and that these features may contain elements that function in nucleolar biology.

4.1.3. The broader significance of the study

This thesis has addressed the most fundamental challenge to study the nucleolar organizer region of the primates i.e. the absence of the sequences for the region. The complete primate rDNA sequences has provided an insight of the geneomic content of the region that represent one of the largest gap in the primate genome. The avalaitbility of primate rDNA sequences will be helpful to address several unanswered questions realted to the rDNA like evolution of the rDNA in primates, the varation/similarty of the rDNA region compare to the rest of the primate genome, functional roles of the rDNA other rRNA synthesis and so on. Further, the characterization of the human rDNA flanking regions has established that the role of the small arm of the acrocentric chrosomes is not just ribosome biogenesis but this genomic regionpotentiall have several other biological functions as well. Further, this work has also established protocol for various computational analyses like sequence assembly for one of the most challenging regions of genome i.e. repetitive regions. Overall, this work has provided an asset to the scientific community to explore the genomic region that has been neglected for very long time because of its sequence complexity.

4.2. Future Directions

4.2.1. Verification of the IGS transcripts identified using publically available datasets

The transcriptome analysis of the IGS using ENCODE RNA-seq data has identified several potential long poly-A (-) transcripts and a small RNA transcript from the region. The RNA-seq data used for the analysis are from a small subset of primary and immortalized cell types. These cell types may not represent what is happening *in vivo*, therefore it is necessary to further investigate whether the identified transcripts are expressed in human cells or are artefacts of these cell types. To do this, the identified transcripts can be searched for in transcriptome datasets that have been derived from primary human tissues.

A complementary approach is to use cap analysis gene expression (CAGE), which provides data that represents 25-30 bp of the 5' end of transcripts that are captured using the 5' (7-methylguanylate) cap (Shiraki *et al.* 2003). The process of capping is thought to be restricted to the mRNA, however studies have shown that certain small and long RNAs that are transcribed by Pol II also show similar posttranscriptional processing and their promoters coincides with the CAGE tags (Katayama *et al.* 2005b; Fejes-Toth *et al.* 2009). The IGS transcripts identified in this study are likely to noncoding, and given the association of Pol II with the human IGS (Section 2.3.9), it is probable that some of these are transcribed by Pol II. Therefore FANTOM project CAGE data from various cell types and tissue samples (Forrest *et al.* 2014) could be employed to identify the transcription start sites of the IGS transcripts, adding further support for their existence.

RNA-seq datasets for various cancerous tissues from several published studies (Kalyana-Sundaram *et al.* 2012) are publically available. To verify the potential cancer specific transcripts, transcriptome analysis can be performed using these datasets. Further, to verify if the identified cancer specific transcripts are specific to the cancerous tissue, transcriptome data from cancer/noncancerous matched tissues can be analysed. Such datasets can also be employed to address the level of differential expression of any IGS transcripts that are present in both cancerous and noncancerous tissues. Such data is accessible through The Cancer Genome Atlas project that provides total RNA-seq data for cancer/noncancerous matched tissues.

4.2.2. Exploring the role of the transcripts from the IGS

Transcript antisense to the pRNA transcript: The antisense transcript identified in this study was found in all three cancerous cell types included in this study but not in the noncancerous cell types, suggesting it may be a characteristic of cancerous cells. Further validation will be required to establish the presence and cancer-specific expression of this transcript. I proposed that this transcript antisense to the pRNA transcript pairs with the pRNA and hinders its rDNA silencing activity. Cancerous cells are thought to have an increased number of active rDNA units compared to non-cancerous cells (Miller *et al.* 1979; Murao *et al.* 1982), therefore my proposed mechanism of this novel transcript in cancer cells is consistent with the increase in active rDNA units. A knockdown study that targets this specific antisense transcript in a variety of different cancer types will be useful to identify whether it does play a role in cancer development.

Small poly(A)+ transcript from the region ~28.5-30.5 kb: The biological role of the poly-(A)+ transcript from the IGS region represented by cluster 2 (~28.5-30.5 kb) in embryonic

H1-hESC cells is unknown. Embryonic cells are known to have a higher number of active rDNA units than differentiated cells. Since these transcripts are found in only H1-hESC cells, I have suggested that this transcript has some embryonic cell-specific function. One possible role for this transcript could be regulating rDNA methylation in embryonic cells. It is known from mouse that embryonic cells have lower levels of rDNA methylation than differentiated cells. It is possible that this transcript somehow promotes the demethylation of the rDNA in embryonic cells. Experimental approaches, including inducing this transcript in non-embryonic cells, knocking down the transcript in embryonic cells, and then comparing the level of rDNA methylation, will be helpful to test this hypothesis.

Long poly(A)- transcripts that includes the cdc27 pseudogene: The pseudogenes are known to regulate the expression of their parent gene. Therefore, the transcript identified from cdc27 pseudogene may have a similar role in the regulation of expression of its parent CDC27 gene. Experimental approach to identify the interaction of cdc27 pseudogene transcript and Cdc27 mRNA will be helpful to identify its role in the regulation of cdc27 and the cell cycle process.

Remaining long poly(A)- transcripts: RNA-seq has identified several other long poly(A)-transcripts in the IGS, their role is unknown. However, similar to IGS_{28}RNA (the transcript that plays a role in the sequestration of the VHL protein), these transcripts may have a role in the sequestration of nucleolar proteins. Biotinylated capture primers with magnetic beads targeting these transcripts in combination with mass spectroscopy could be used to identify the proteins that bind to these transcripts.

4.2.3. Verification of the identified origin of replication

This study has identified three potential origin of replication in the human IGS using ChIP-seq data obtained from Hela-S3 for ORCs. The association of the ORCs demarcates the potential origin of replication site but it not essential that the identified site activate during the cell division. Therefore, further experiments are required to verify the potential origin of replication. The Repli-Seq data is obtained by sequencing the labelled nascent DNA strand that is isolated from the replication bubble is used to demarcate the sites of origin of replication (Hansen *et al.* 2010). Such data representing the different stages of the cell cycle for various cell types is publically available through ENCODE project and can be employed to verify the origin of replication in the rDNA. Dimitrova (Dimitrova 2011) has reported that different regions of the rDNA act as origin of replication depending of the cell cycle stage. It is probable that the identified potential origin of replication activated during different stages

of cell cycles. Since the available data that represent different cell cycle stages, can be employed to compare the variation in activation of the potential origin of replication sites.

4.2.4. Role of the identified conserved regions

Although RNA-seq and ChIP-seq analyses provide some evidence for the potential roles of several of the conserved regions in the human IGS I found in this study, there are still many conserved regions that are not associated with any histone modifications or RNA transcripts. Further studies are required to establish the functions of these regions. One approach to characterize the remaining conserved regions could be to map ChIP-seq data for factors involved in processes that are thought to be regulated or governed by the rDNA, for example cellular proliferation (NANOG and Jun) and cell cycle regulation (p53). Further, it is possible that these regions are transcribed or enriched for histone modifications and transcription factors in other tissues not included in this study. Therefore, a detailed RNA-seq and ChIP-seq analysis using additional cell types may be able to shed light on the possible roles of some of these uncharacterized conserved IGS regions.

4.2.5. Comparative analysis of human IGS transcripts.

The phylogenetic footprinting analysis has revealed the presence of several highly conserved regions in the primate IGS. However, due to the absence of functional annotation of the primate IGS it is not possible to predict if the observed sequence conservation translates into functional conservation. One strategy to assay for functional conservation of the IGS is by performing comparative analysis of the transcripts from the region. The Nonhuman Primate Reference Transcriptome Resource (NHPRTR) is a repository that has RNA-seq data for 13 species from 21 different tissue samples (Pipes *et al.* 2013). These data can be mapped to the reference primate rDNA sequences to identify transcripts from the respective primate IGS. These transcripts can then be compared to those from the human IGS to determine the level of transcript conservation. Such a study would not only determine the conservation of transcripts from the IGS but would also shed light on the evolution of long poly(A)-transcripts, which is far from understood (Katinakis *et al.* 1980; Yang *et al.* 2011). A recent comparative study on long poly(A)+ noncoding transcripts has demonstrated that several lncRNAs are conserved amongst the primates and may play roles in processes like spermatogenesis, synaptic transmission and placenta development (Necsulea *et al.* 2014). Investigating the conservation of the human IGS transcripts identified here among the primates will also help test whether these transcripts result from transcriptional noise rather than some biological function.

4.2.6. Role of Pol II in IGS transcription

All the known transcripts from the human IGS i.e. pRNA, IGS_{16}RNA and IGS_{28}RNA are reported to be transcribed by RNA polymerase I machinery (Audas *et al.* 2012; Jacob *et al.* 2012). The ChIP-seq analysis has shown that Pol II and its cofactors are associated with the human IGS, which suggests that Pol II may have a role in IGS transcription. To test this hypothesis one possible experiment can be to confirm if RNA polymerase II transcribes any of the idetifed transcripts from the RNA-seq analysis. Small poly(A)+ transcript in embryonic cell type H1-hESC can be one of the potential candidate for the experiment. The ChIP-seq and RNA-seq analysis of the human IGS have demonstrated that in H1-hESC the peaks for Pol II and its cofactor TBP coincides with the signal for the small poly(A)+ transcript (Figure 2.35). It is possible that the colocalization of the three signals is because the identified transcript is transcribed by Pol II. An experiment involving Pol I (e.g. CX-5461) and Pol II (e.g. α-Amanitin) inhibitors paired with RT-PCR, to detrermine the effect of inhibition of a specific polymerase on the expression of transcript, can be employed to find out the role of Pol II in the IGS transcription.

4.2.7. Phylogenetic footprinting of the rDNA flanking regions

The identification and characterization of the rDNA flanking regions sequences have provided a platform to identify the potential functional elements present in these regions. The ENCODE analysis and the gene prediction pipeline has revealed the presence of several potential transcripts and transcriptional regulators in the distal flanking region (Floutsakou *et al.* 2013). However, it cannot be ruled out that several other functional elements in addition to these transcriptional regulators may also be present in the distal region. A comprehensive study is required to identify these functional elements. One possible strategy to achieve this is a phylogenetic footprinting analysis of the rDNA flanking regions similar to that performed for the human IGS in Chapter 2. Further, it is difficult to target the proximal region using molecular biology techniques because of the high level of segmental duplication. Therefore, a computational method like phylogenetic footprinting may be useful to study this region, too. However, similar to the human IGS phylogenetic footprinting here, the biggest bottleneck is the absence of the sequences that flank the rDNA for other primate species. Therefore, the first step to perform this analysis must be the identification of the sequence of the rDNA flanking regions in the other primates. A strategy similar to the one performed to identify the human rDNA flanking regions (section 3.1.2) involving chromosome specific cosmid library screening, BAC walking and fluorescence in situ hybridization could be employed for target primate species to identify the flanking sequences.

A comparative study involving the rDNA flanking regions of different yeast species has shown despite different chromosomal locations and arrangements, the flanking regions are generally conserved (Proux-Wera *et al.* 2013). This demonstrates that during the event of translocation the flanking regions and the rDNA array moves together to another chromosomes. In the primates, the NORs are distributed on multiple chromosomes. In chimpanzee *(Pan troglodytes*) NORs are present on chr 14, 15, 17, 22 and 23, in gorilla (*Gorilla gorilla*) on chr 22 and 23, in orangutan (*Pongi pygmaues abelii*) chr 11, 12, 13, 14, 15, 16, 17, 22 and 23, in gibbon (*Hylobates lar*) on chr 15 and in macaque (*Macaque mulatta*) on chr 13. The chimpanzee chr 14, 15, 17, 22 and 23 are homologues to human chr 13, 14, 18, 21 and 22, gorilla chr 22 and 23 are homologues to human chr 21 and 22 while in orangutan chr 11, 12, 13, 14, 15, 16, 17, 22 and 23 are homologues to human chr 2p, 2q, 9, 13, 14, 15, 18, 21 and 22 respectively. The chromosome 15 in gibbon and chromosome 13 in rhesus macaque are the marker chromosomes and do not have homologs in human. Considering the multi and nonhomologous chromosomal distribution of the NORs in the primates, it is difficult to say if similar to the rDNA, where a high selective evolutionary constrain is applicable because of associated housekeeping function, the flanking regions are also conserved among the primates. However, considering the case in yeast it is probable that the flanking regions in some of the chromosomes are conserved among the primates.

4.2.8. Role of the rDNA flanking regions in nucleolar formation/fusion

Cancerous cells are marked by nucleolar enlargement that results from the fusion of smaller nucleoli, a process known as nucleolar fusion (Maggi and Weber 2005; Montanaro *et al.* 2008). However, the mechanism that drives this nucleolar fusion is still not known. Our collaborators have found that the rDNA flanking regions are localised to the periphery of the nucleoli (Floutsakou *et al.* 2013), suggesting that these flanking regions may contain elements that play roles in nucleolar formation/fusion. One possibility is that mutational changes in these flanking region elements may prompt nucleolar fusion, leading to the cell becoming cancerous. To explore this, a genome wide association study using the flanking region sequences assembled here will be helpful to identify variants in cancerous and noncancerous cells that may direct the process of nucleolar fusion.

[Blank Page]

Appendix I

Tables and Figures

[Blank Page]

Appendix Table 1: Assembly statistics for the primate whole genome assemblies.

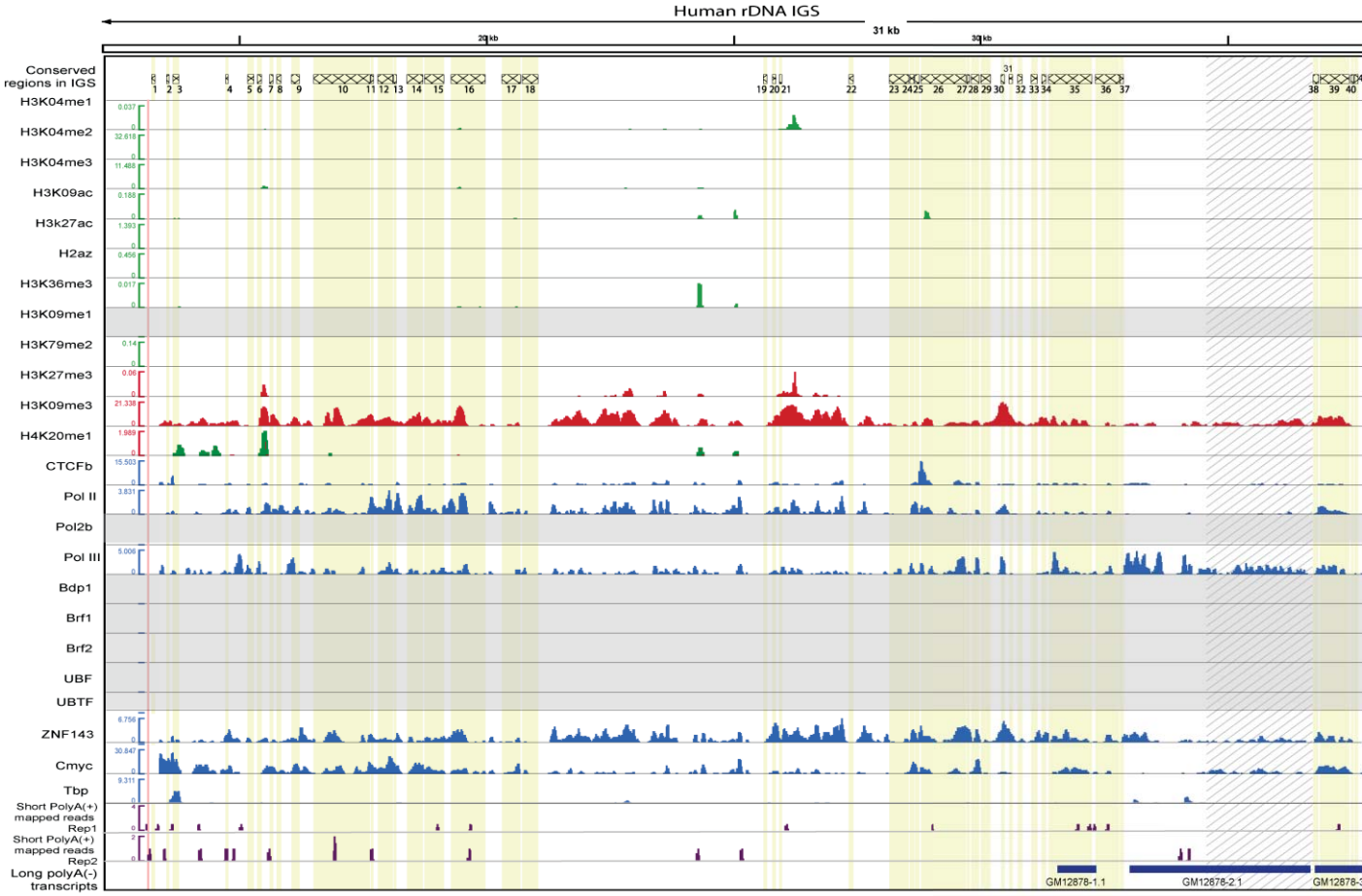
Organism	# of contigs	Max contig length (bp)	Mean (bp)	Length weighted mean	N50	Total size of contigs (bp)
Pan troglodytes	67,048	188,775	7,325	10,138	8,028	491,159,851
Gorilla gorilla	40,813	158,554	7,235	10,253	7,707	295,320,131
Pongo abelii	59,433	158,420	6,305	7,987	6,392	374,781,986
Nomascus leucogenys	120,247	186,034	7,796	17,777	13,019	831,131,126
Macaca mulatta	78,153	197,050	4,772	5,542	4,800	372,969,636
Callithrix jacchus	133,543	207,318	6,813	9,145	7,641	909,956,131

Appendix Table 2: Coordinates for the conserved regions in the human IGS.

Conserved region name	Start coordinate	End coordinate
conR_1	13335	13406
conR_2	13639	13694
conR_3	13767	13885
conR_4	14828	14894
conR_5	15282	15407
conR_6	15470	15565
conR_7	15725	15791
conR_8	15866	15950
conR_9	16167	16322
conR_10	16608	17747
conR_11	17762	17822
conR_12	17907	18206
conR_13	18230	18280
conR_14	18488	18813
conR_15	18836	19229
conR_16	19372	20066
conR_17	20397	20777
conR_18	20807	21139
conR_19	25675	25751
conR_20	25843	25929
conR_21	25988	26052
conR_22	27385	27484
conR_23	28214	28600
conR_24	28619	28711
conR_25	28721	28806
conR_26	28854	29764

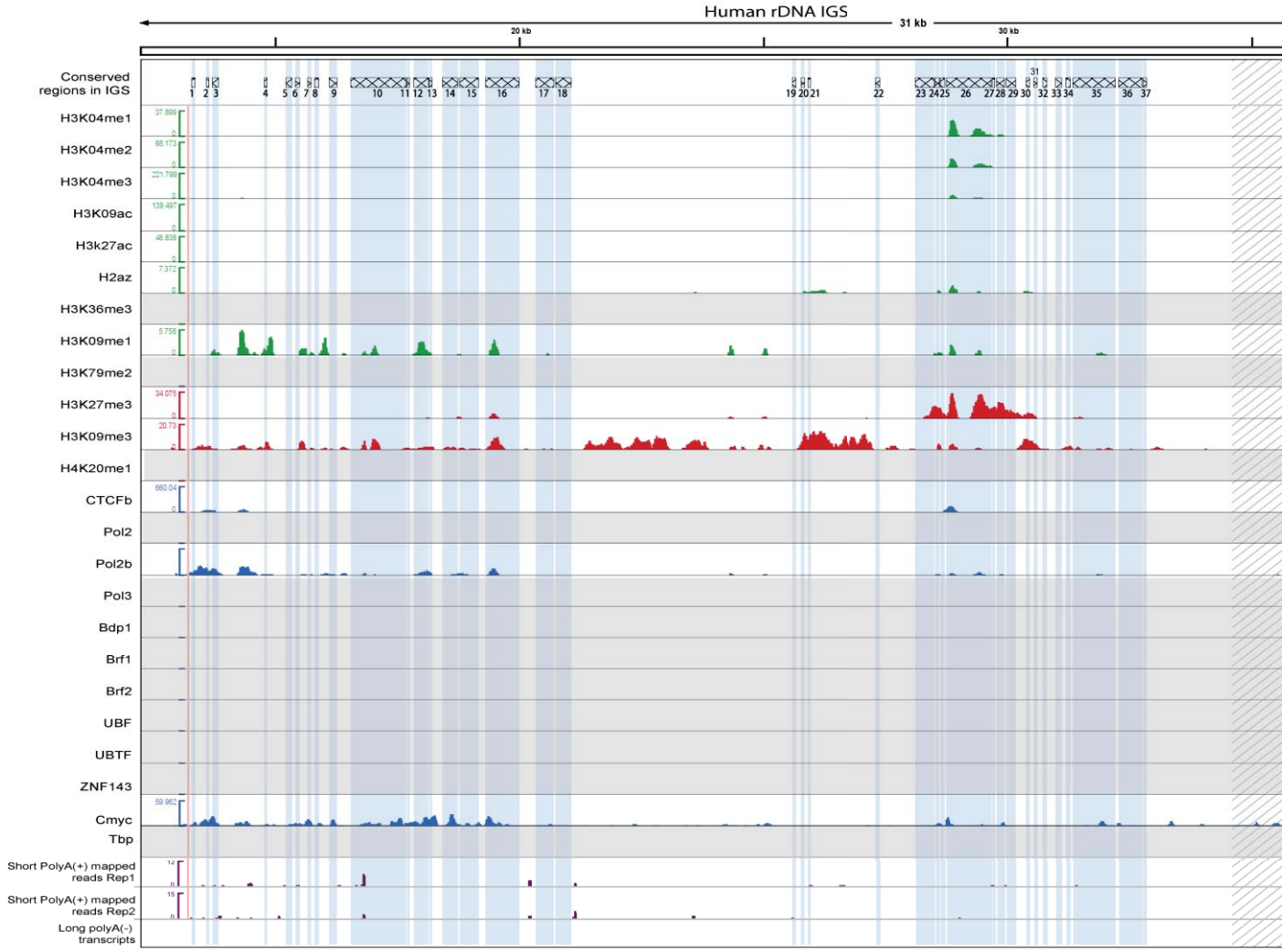
Appendix Table 2: Coordinates for the conserved regions in the human IGS.

Conserved region name	Start coordinate	End coordinate
conR_27	29794	29845
conR_28	29870	30045
conR_29	30074	30268
conR_30	30475	30555
conR_31	30635	30706
conR_32	30817	30908
conR_33	31089	31205
conR_34	31302	31378
conR_35	31443	32312
conR_36	32382	32854
conR_37	32876	32952
conR_38	36769	36870
conR_39	36894	37480
conR_40	37525	37574
conR_41	37592	37669
conR_42	38377	38468
conR_43	38500	38914
conR_44	38950	39228
conR_45	39392	39751
conR_46	39775	40107
conR_47	40210	40296
conR_48	40329	40575
conR_49	41633	41699
conR_50	41923	42010
conR_51	42297	42407
conR_52	42711	42811
conR_53	42831	43972



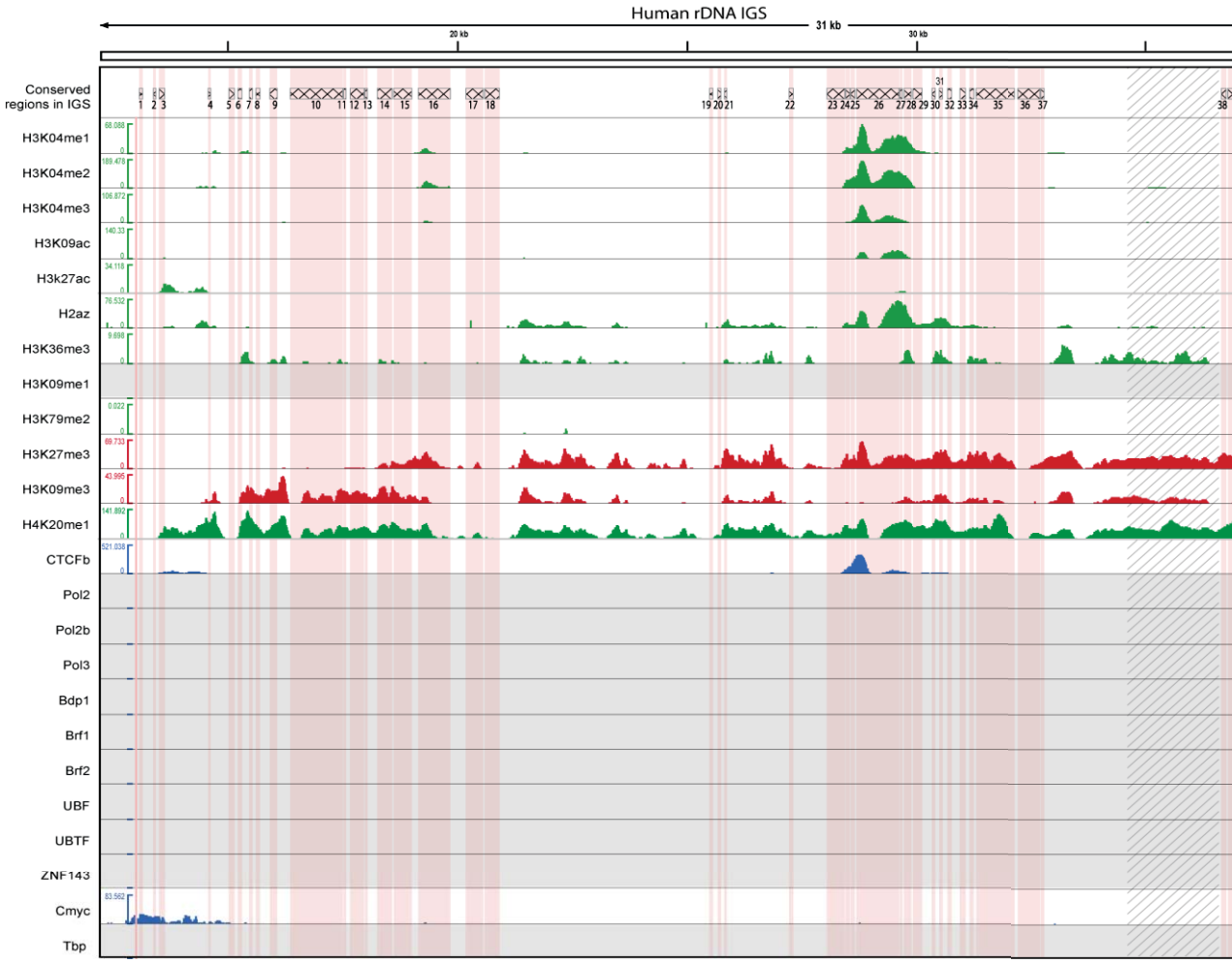
Appendix Figure 1: Chromatin, transcription factor and transcript landscape of the IGS in lymphoblas

The hatched boxes with the blue shaded regions below represent the conserved regions with their name indicated below. The shown as diagonally shaded region. Each row represents the enrichment for an active histone modifications (green peaks), rep TF (blue peaks. The last three rows represent small poly(A)+ signal (purple peaks) or long ploy(A)-(blue boxes). The name of t data is indicated to the left. The scale on left of each row represents the level of enrichment. The scale above shows the posit demarcated by the pink vertical line. The grey rows represents the absence of the data for the histone modification or TF in the



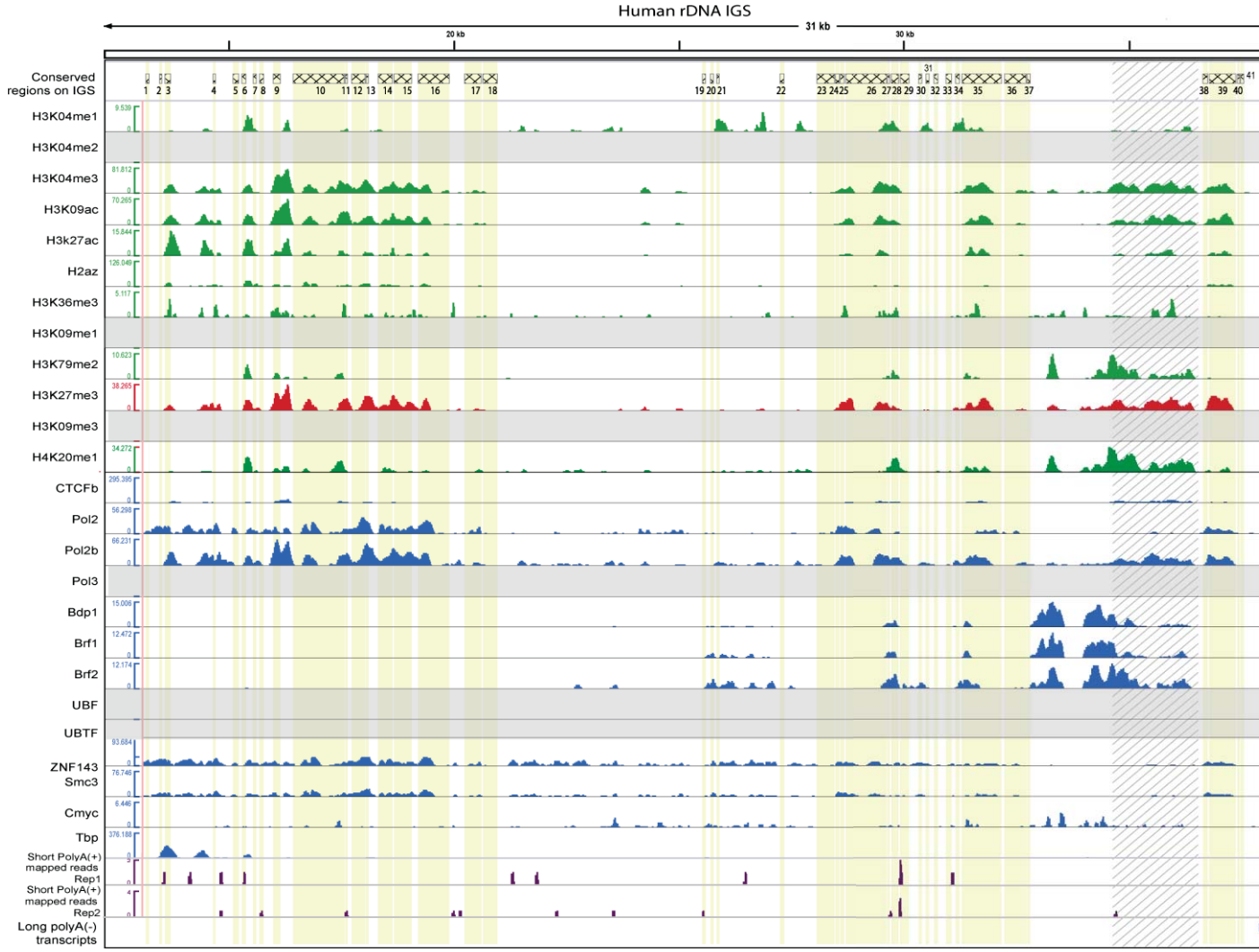
Appendix Figure 2: Chromatin, transcription factor and transcript landscape of the IGS in umbilical v

Notations are same as in Appendix Figure 1.



Appendix Figure 3: Chromatin, transcription factor and transcript landscape of the IGS

Notations are same as in Appendix Figure 1.



Appendix Figure 4: Chromatin, transcription factor and transcript landscape of the IGS in cerv

Notations are same as in Appendix Figure 1

Appendix Figure 5: Chromatin, transcription factor and transcript landscape of the IGS in leukaemia

Notations are same as in Appendix Figure 1.

[Blank Page]

Appendix Table 3: Sequencing statics of the distal and proximal cosmids.

Cosmid name	Yield (Mbp)	% PF[a]	# Reads	% of >= Q30 Bases (PF)	Mean Quality Score (PF)	Mean length of reads (bp)
LA13 165F6	233	91.62	4,621,240	94.67	37.59	55
LA14 101B3	203	91.23	4,036,040	94.18	37.25	55
LA15 25H3	217	92.92	4,237,550	95.5	37.88	55
N 29M24	281	92.39	5,530,692	94.99	37.67	55
LA13 133H12	438	88.40	9,013,761	93.28	37.12	55
LA15 64C10	407	91.45	8,105,879	95.13	37.77	55

Appendix Table 4: Assembly statistics for the distal and proximal cosmid assemblies

Cosmid name	Average Coverage	# of contigs	# of contigs > 100	# of contigs > N50	Max contig length (bp)	Mean (bp)	N50	Total size of contigs (bp)
LA13 165F6	4,172	83	24	4	12224	2067	4351	49619
LA14 101B3	2,932	68	25	4	12163	2138	5585	53460
LA15 25H3	4,363	30	9	2	21892	4979	6521	44817
N 29M24	4,681	609	91	13	3286	438	812	39879
LA13 133H12	10,736	265	52	10	4427	814	1182	42337
LA15_64C10	7,554	304	66	14	2573	719	1114	47490

Appendix Table 5: Sequence similarity matrix for the distal region BAC clones.

	AL592188 (22)	AL353644 (22)	CT476834 (21)	CT476837 (21)	CU633904 (21)	CU633906 (21)	CU633967 (21)	CU633971 (21)	CU634019 (21)	CU638689 (21)	F 4
AL592188 (22)	X	Iden = 99.0% Len = 101,647	Iden = 99.1% Len = 24,990	Iden = 99.2% Len = 88,539	No overlap	No overlap	No overlap	Iden = 99.1% Len = 70,584	Iden = 98.8% Len = 14,940	No overlap	I 9 L 7
AL353644 (22)	Iden = 99.0% Len = 101,647	X	Iden = 98.9% Len = 41,921	Iden = 99.2% Len = 88,535	No overlap	No overlap	Iden = 99.0% Len = 5,127	Iden = 99.1% Len = 70,584	Iden = 98.8% Len = 31,863	No overlap	I 9 L 8
CT476834 (21)	Iden = 99.1% Len = 24,990	Iden = 98.9% Len = 41,921	x	Iden = 100% Len = 13,552	Iden = 100% Len = 123,026	Iden = 100% Len = 128,101	Iden = 100% Len = 128,588	Iden = 100% Len = 2000	Iden = 100% Len = 165,310	Iden = 100% Len = 89,432	I 1 L 9
CT476837 (21)	Iden = 99.2% Len = 88,539	Iden = 99.2% Len = 8,535	Iden = 100% Len = 13,552	X	No overlap	No overlap	No overlap	Iden = 100% Len = 70,585	Iden = 100% Len = 3,503	No overlap	I 1 L 5
CU633904 (21)	No overlap	No overlap	Iden = 100% Len = 123,026	No overlap	X	Iden = 100% Len = 108,738	Iden = 100% Len = 123,026	No overlap	Iden = 100% Len = 123,026	Iden = 100% Len = 70,069	I 1 L 5
CU633906 (21)	No overlap	No overlap	Iden = 100% Len = 128,101	No overlap	Iden = 100% Len = 108,738	X	Iden = 99.9% Len = 110,590	No overlap	Iden = 100% Len = 120,606	Iden = 100% Len = 108,905	I 1 L 4
CU633967 (21)	No overlap	Iden = 99.0% Len = 5,127	Iden = 100% Len = 128,588	No overlap	Iden = 100% Len = 123,026	Iden = 99.9% Len = 110,590	X	No overlap	Iden = 100% Len = 128,588	Iden = 100% Len = 71,921	I 1 L 5
CU633971 (21)	Iden = 99.1% Len = 70,584	Iden = 99.1% Len = 70,584	Iden = 100% Len = 2,000	Iden = 100% Len = 70,585	No overlap	No overlap	No overlap	X	No overlap	No overlap	I 1 L 4

Appendix Table 5: Sequence similarity matrix for the distal region BAC Clones

	AL592188 (22)	AL353644 (22)	CT476834 (21)	CT476837 (21)	CU633904 (21)	CU633906 (21)	CU633967 (21)	CU633971 (21)	CU634019 (21)	CU638689 (21)	F 4
CU634019 (21)	Iden = 98.8% Len = 14,940	Iden = 98.8% Len = 31,863	Iden = 100% Len = 165,310	Iden = 100% Len = 3,503	Iden = 100% Len = 123,026	Iden = 100% Len = 120,606	Iden = 100% Len = 128,588	No overlap	X	Iden = 100% Len = 81,937	Id 1 L 8
CU638689 (21)	No overlap	No overlap	Iden = 100% Len = 89,432	No overlap	Iden = 100% Len = 70,069	Iden = 100% Len = 108,905	Iden = 100% Len = 71,921	No overlap	Iden = 100% Len = 81,937	X	Id 1 L 3
FP236241 (21)	Iden = 99.1% Len = 71,235	Iden = 98.7% Len = 88,164	Iden = 100% Len = 96,652	Iden = 100% Len = 59,714	Iden = 100% Len = 56,186	Iden = 100% Len = 41,898	Iden = 100% Len = 59,986	Iden = 100% Len = 48,162	Iden = 100% Len = 86,603	Iden = 100% Len = 3,229	X
FP236315 (21)	Iden = 99.1% Len = 93,698	Iden = 99.0% Len = 110,608	Iden = 100% Len = 49, 829	Iden = 100% Len = 82,137	Iden = 100% Len = 9,363	No overlap	Iden = 100% Len = 13,073	Iden = 100% Len = 70,584	Iden = 100% Len = 39,780	No overlap	Id 1 L 9
FP236383 (21)	Iden = 99.2% Len = 74,223	Iden = 99.1% Len = 74,227	No overlap	Iden = 100% Len = 74,119	No overlap	No overlap	No overlap	Iden = 100% Len = 66,153	No overlap	No overlap	Id 1 L 4
FP671120 (21)	Iden = 99.1% Len = 101,648	Iden = 99.1% Len = 102,335	Iden = 100% Len = 25,655	Iden = 100% Len = 90,103	No overlap	No overlap	No overlap	Iden = 100% Len = 70,585	Iden = 100% Len = 15,606	No overlap	Id 1 L 7
AC011841 (17?)	No overlap	No overlap	Iden = 98.31% Len = 55,002	No overlap	Iden = 98.46% Len = 35,589	Iden = 98.85% Len = 92,734	Iden = 98.42% Len = 37,422	No overlap	Iden = 98.3% Len = 47,490	Iden = 98.7% Len = 74,515	N o

Appendix Table 6: Repeat statistics of the distal contig.

```
==================================================

Total length: 379046 bp (379046 bp excl N/X-runs)
GC level: 41.48 %
bases masked: 244630 bp ( 64.54 %)
==================================================
 number of length percentage
 elements* occupied of sequence
--------------------------------------------------
SINEs: 117 27250 bp 7.19 %
 ALUs 108 26260 bp 6.93 %
 MIRs 9 990 bp 0.26 %

LINEs: 78 95143 bp 25.10 %
 LINE1 75 94873 bp 25.03 %
 LINE2 3 270 bp 0.07 %
 L3/CR1 0 0 bp 0.00 %

LTR elements: 80 52738 bp 13.91 %
 ERVL 21 18975 bp 5.01 %
 ERVL-MaLRs 23 6926 bp 1.83 %
 ERV_classI 36 26837 bp 7.08 %
 ERV_classII 0 0 bp 0.00 %

DNA elements: 4 598 bp 0.16 %
 hAT-Charlie 4 598 bp 0.16 %
 TcMar-Tigger 0 0 bp 0.00 %

Unclassified: 1 2003 bp 0.53 %

Total interspersed repeats: 177732 bp 46.89 %


Small RNA: 5 897 bp 0.24 %

Satellites: 18 61181 bp 16.14 %
Simple repeats: 62 3512 bp 0.93 %
Low complexity: 11 1308 bp 0.35 %
==================================================
```

Appendix Table 7: Repeat statistics of the proximal contig.

```
==================================================
Total length: 209483 bp (209483 bp excl N/X-runs)
GC level: 43.32 %
bases masked: 112232 bp ( 53.58 %)
==================================================
 number of length percentage
 elements* occupied of sequence
--------------------------------------------------
SINEs: 73 19309 bp 9.22 %
 ALUs 66 18302 bp 8.74 %
 MIRs 7 1007 bp 0.48 %

LINEs: 48 40600 bp 19.38 %
 LINE1 41 39077 bp 18.65 %
 LINE2 4 983 bp 0.47 %
 L3/CR1 3 540 bp 0.26 %

LTR elements: 41 29025 bp 13.86 %
 ERVL 6 1817 bp 0.87 %
 ERVL-MaLRs 15 6375 bp 3.04 %
 ERV_classI 16 14078 bp 6.72 %
 ERV_classII 2 5877 bp 2.81 %

DNA elements: 18 6494 bp 3.10 %
 hAT-Charlie 6 980 bp 0.47 %
 TcMar-Tigger 10 4830 bp 2.31 %

Unclassified: 0 0 bp 0.00 %

Total interspersed repeats: 95428 bp 45.55 %


Small RNA: 1 577 bp 0.28 %

Satellites: 12 13688 bp 6.53 %
Simple repeats: 39 2350 bp 1.12 %
Low complexity: 3 189 bp 0.09 %
==================================================
```

Appendix Table 8: Segmentally duplicated regions from the proximal contig.

Contig	Start	End	Length	Duplicate chromosome position	Start	End	% identity
PJ	1	105260	105260	Chromosome 1	143533694	143428925	96.5
PJ	1	73800	73800	Chromosome 1	142553006	142568640	94.6
PJ	1	73523	73523	Chromosome 1	142957003	142882933	92.6
PJ	74595	105050	30456	Chromosome 1	142632870	142662577	92
PJ	83766	89181	5416	Chromosome 1	142867130	142861830	93.6
PJ	45647	50040	4394	Chromosome 1	826486	830889	92.6
PJ	100926	105105	4180	Chromosome 1	143290746	143286566	94.1
PJ	142793	146966	4174	Chromosome 1	143236907	143241077	96.5
PJ	142793	146963	4171	Chromosome 1	143351481	143347325	96.4
PJ	100926	105050	4125	Chromosome 1	143154903	143159025	95.5
PJ	142793	145163	2371	Chromosome 1	142797349	142799737	96
PJ	100942	103157	2216	Chromosome 1	142853069	142850868	94
PJ	77250	79225	1976	Chromosome 1	143134409	143136392	95.1
PJ	105310	143700	38391	Chromosome 2	95520400	95559252	90.4
PJ	166722	191289	24568	Chromosome 2	132522403	132547076	90.9
PJ	157565	166067	8503	Chromosome 2	132513740	132522358	93.2
PJ	176920	180292	3373	Chromosome 2	132577642	132580992	92.7
PJ	162850	166067	3218	Chromosome 2	132555152	132558361	94.5
PJ	103245	105238	1994	Chromosome 2	132391991	132393988	84.9
PJ	103700	105238	1539	Chromosome 2	131424391	131425925	87.6
PJ	205180	206424	1245	Chromosome 2	162137144	162138410	89.1
PJ	103245	104369	1125	Chromosome 2	132034672	132035800	85.2
PJ	103245	104369	1125	Chromosome 2	131209987	131211114	84.8
PJ	103245	104369	1125	Chromosome 2	131207537	131208665	85.6
PJ	103245	104369	1125	Chromosome 2	131428006	131429134	85.8
PJ	103254	104365	1112	Chromosome 2	131427378	131428485	85.1
PJ	103254	104365	1112	Chromosome 2	131208186	131209296	85.7
PJ	103254	104364	1111	Chromosome 2	132034042	132035151	85.4
PJ	103245	104298	1054	Chromosome 2	130819347	130820398	85.1
PJ	103350	104369	1020	Chromosome 2	130822611	130823632	85.2
PJ	103245	104260	1016	Chromosome 2	130823142	130824161	86.7
PJ	20453	71388	50936	Chromosome 3	75829445	75880859	92

Contig	Start	End	Length	Duplicate chromosome position	Start	End	% identity
PJ	360	20452	20093	Chromosome 3	75887065	75906840	91.1
PJ	71560	88466	16907	Chromosome 3	75809054	75826395	90.1
PJ	77250	94391	17142	Chromosome 4	49180639	49198010	93.5
PJ	77250	88799	11550	Chromosome 4	49608353	49620270	91.1
PJ	99701	105105	5405	Chromosome 4	49202173	49207588	94.5
PJ	99701	105105	5405	Chromosome 4	49588922	49594333	94.4
PJ	64214	67839	3626	Chromosome 4	49161181	49164805	93
PJ	64214	66367	2154	Chromosome 4	49628635	49630784	92.9
PJ	204592	206423	1832	Chromosome 4	49282737	49284629	86.1
PJ	204592	206423	1832	Chromosome 4	49294922	49296797	87.7
PJ	204592	206423	1832	Chromosome 4	49301077	49302933	87.9
PJ	204592	206423	1832	Chromosome 4	49307357	49309231	87.9
PJ	204592	206423	1832	Chromosome 4	49514283	49516172	87.2
PJ	204592	205855	1264	Chromosome 4	49311042	49312353	87.7
PJ	143988	145443	1456	Chromosome 7	97499612	97501087	88.5
PJ	132568	133633	1066	Chromosome 7	97488744	97489797	88.2
PJ	161526	199611	38086	Chromosome 9	42259637	42297833	92
PJ	161526	192870	31345	Chromosome 9	70655139	70686615	93.3
PJ	161526	192869	31344	Chromosome 9	45400618	45431950	93.6
PJ	175341	191110	15770	Chromosome 9	68322537	68338220	94.1
PJ	192869	199611	6743	Chromosome 9	45394186	45400585	90
PJ	193195	199610	6416	Chromosome 9	70686619	70693040	94.9
PJ	193198	199611	6414	Chromosome 9	43203986	43210444	94.4
PJ	161526	166068	4543	Chromosome 9	68308254	68312765	94.7
PJ	106413	109844	3432	Chromosome 9	44117027	44120374	87.7
PJ	108137	111348	3212	Chromosome 9	42364995	42368241	90.6
PJ	189660	192870	3211	Chromosome 9	43210454	43213696	94.1
PJ	158413	161573	3161	Chromosome 9	45431993	45435142	95.7
PJ	108232	111348	3117	Chromosome 9	69378612	69381771	90.9
PJ	108232	111348	3117	Chromosome 9	43133754	43136914	91
PJ	108235	111348	3114	Chromosome 9	67923394	67926548	90.9
PJ	166072	168898	2827	Chromosome 9	68312893	68315739	94.9
PJ	158413	160331	1919	Chromosome 9	70653271	70655188	96.1

Contig	Start	End	Length	Duplicate chromosome position	Start	End	% identity
PJ	158413	160331	1919	Chromosome 9	42257769	42259686	95.8
PJ	158413	160260	1848	Chromosome 9	68306368	68308214	96
PJ	106413	107996	1584	Chromosome 9	43132024	43133529	88.2
PJ	198109	199611	1503	Chromosome 9	68338785	68340315	94.5
PJ	102087	103220	1134	Chromosome 9	38567007	38568129	86.5
PJ	168902	191112	22211	Chromosome 14	19758417	19780416	91.7
PJ	158413	168513	10101	Chromosome 14	19817857	19828180	93
PJ	161526	168901	7376	Chromosome 14	19750845	19758348	93.2
PJ	162863	168513	5651	Chromosome 14	19361516	19367278	92.5
PJ	103982	105265	1284	Chromosome 14	19600050	19601322	85.1
PJ	103982	105265	1284	Chromosome 14	19974090	19975362	85.1
PJ	176920	180678	3759	Chromosome 18	14989355	14993073	93
PJ	201937	203471	1535	Chromosome 19	44916370	44917954	84.9
PJ	204592	206073	1482	Chromosome 19	44962091	44963591	89.8
PJ	202036	203471	1436	Chromosome 19	44959179	44960652	87
PJ	204672	206073	1402	Chromosome 19	44913397	44914821	89.2
PJ	16075	71345	55271	Chromosome 21	10113638	10168755	94.6
PJ	158413	199611	41199	Chromosome 21	10604062	10645045	92.8
PJ	74600	95600	21001	Chromosome 21	10178710	10199160	92.6
PJ	74595	94392	19798	Chromosome 21	9662432	9682082	94.4
PJ	1	15750	15750	Chromosome 21	10097985	10113638	94.1
PJ	94695	105103	10409	Chromosome 21	9682085	9692595	90.4
PJ	64214	69025	4812	Chromosome 21	9645775	9650595	93.1
PJ	99060	103157	4098	Chromosome 21	10199865	10203948	95.1
PJ	71425	73760	2336	Chromosome 21	9653425	9655770	92
PJ	69195	71350	2156	Chromosome 21	9650595	9652751	93.6
PJ	159895	191112	31218	Chromosome 22	16062555	16093911	92
PJ	162863	168513	5651	Chromosome 22	16460105	16465867	92.5
PJ	103982	105265	1284	Chromosome 22	16241735	16243007	85.1
PJ	103982	105265	1284	Chromosome 22	16246649	16247921	85.9
PJ	1	141920	141920	Chromosome Y	13295249	13436504	95.6
PJ	164404	171133	6730	Chromosome Y	13239281	13246083	96.1
PJ	144448	146947	2500	Chromosome Y	13292733	13295248	96.45

Appendix Table 9: Segmentally duplicated regions from the proximal contig.

Contig	Start	End	Length	Duplicate chromosome location	Start	End	% identity
DJ	104572	107241	2670	Chromosome 1	68626275	68628991	85.9
DJ	321433	324062	2630	Chromosome 1	68626302	68628991	86.8
DJ	339231	340355	1125	Chromosome 2	133118377	133119492	86.9
DJ	133743	134999	1257	Chromosome 3	75692439	75693702	90.5
DJ	187100	192124	5025	Chromosome 3	75683443	75688394	86.5
DJ	192335	195685	3351	Chromosome 3	75679409	75682745	88.5
DJ	229508	233248	3741	Chromosome 3	75679017	75682746	88.7
DJ	233455	235987	2533	Chromosome 3	75683438	75685973	89.4
DJ	291986	295386	3401	Chromosome 3	75690329	75693702	88.8
DJ	133743	134999	1257	Chromosome 4	190967606	190968868	90.2
DJ	189397	192119	2723	Chromosome 4	190975243	190978018	87.3
DJ	192335	194410	2076	Chromosome 4	190978791	190980900	89.1
DJ	194452	196632	2181	Chromosome 4	190980897	190983106	89.3
DJ	228930	233248	4319	Chromosome 4	190978790	190983106	88.5
DJ	233455	235987	2533	Chromosome 4	190975469	190978029	88.4
DJ	291986	294160	2175	Chromosome 4	190968819	190970979	87.7
DJ	104624	106629	2006	Chromosome 5	97912035	97914081	87.2
DJ	322240	324062	1823	Chromosome 5	97912219	97914060	85.7
DJ	104572	105914	1343	Chromosome 10	126676918	126678260	90.4
DJ	322759	324062	1304	Chromosome 10	126676945	126678269	91.5
DJ	133743	134999	1257	Chromosome 10	135459443	135460705	89.9
DJ	189397	192124	2728	Chromosome 10	135466940	135469711	87.7
DJ	192335	196632	4298	Chromosome 10	135470341	135474655	88.8
DJ	228930	233248	4319	Chromosome 10	135470340	135474655	89.1
DJ	233455	235987	2533	Chromosome 10	135467165	135469716	88.9
DJ	291986	295386	3401	Chromosome 10	135459443	135462818	87.9
DJ	322516	323669	1154	Chromosome 11	43544396	43545549	85.6
DJ	191180	192208	1029	Chromosome 12	38482027	38483048	85.3
DJ	233375	234434	1060	Chromosome 12	38482027	38483075	85.2
DJ	104453	105987	1535	Chromosome Y	59001390	59002913	91.1
DJ	322768	324223	1456	Chromosome Y	59001463	59002943	88.6

Appendix Table 10: Multiple sequence alignment for the ACRO138 repeat.

```
DJ_136616_136731 --------------------------CATGCTGGGAT------TGTAGTCC-TGTAGCCC
DJ_136738_136867 ---------TGTCCCTGACCCCAGTGCATAATGGGAT------GGTAGTCC-TGCAGCCC
DJ_136868_137009 AATGCGCACTCTCCCTGAGCCGGGTGCATGCTGGGAT------TGTAGTCC-TGCAGCCC
DJ_137010_137147 ---GCGCACATTCCCTGCGCCCGGTCCATGCTAGGAT------TGTAGCGC-TGCAGCCC
DJ_137151_137289 ---GCGCACTCTCCCTGATCCCGATGTATGCTGGGAT------TGTAGTGA-TGCAGCCC
DJ_137290_137427 ---GCGCGTTGTCCCTGAGCAGGATGCATGCTGGGAT------TGTAGTCC-TGAAGCCC
DJ_137429_137567 ---GCGGACTGCCCCGGCGCCCCGTGCATGCTGGAAT------TGTAGTCC-TACAGCGA
DJ_137572_137706 -----GCACTCTCCCTGAGCCAGAGATATGCTGAAAT------TGTACTGC-TGCAGCCC
DJ_137711_137834 ------CACTGTCCATGAACCCGCTGCATGCTGGCAT------TGTAGTCC-TGCAGCCC
DJ_137857_137974 ------------------------TGCATCCTGGGAT------TGTAGTCC-TGTAGCCC
DJ_137974_138108 -ATGTGGACTCTCCCTGAGCCCCGTGCATGCTGGGAT------TGTAGTC-TTATAGCAC
DJ_138117_138253 ---GCGCACTGTCACTGAGCTGGGTGCATGCTAGGAT------TGTAGTCC-TGCAGCCT
DJ_138257_138396 ---GCGCACTATCCCTGAGCTGGGTGCATGTTGGGAT------TGCAGTCC-TGGATCTC
DJ_138397_138541 ---GCGCGCTGTCCCTGAGCCCAGTGCATGCTGGGGCTGGAAGTGTAGTCC-TTCAGGCC
DJ_138607_138743 -----GCACTGTCTCTTAGCCAGGTGCATGCTGGGAT------TGTAGTC-TTCCCGCCC
DJ_138744_138871 ---GCGCACTATCCTTGATCTTGGTGCATACTGGGAT------TGTAGTCC-TGCTGCCC
DJ_138885_139020 -----GCACTATCCCTGATCCCGGTGCATGATGGGAA------TGTAGTCC-TGCAGCCC
DJ_139111_139228 ---GCGCACTGTCTCTGATTCCGGTGTATGCTGGAAT------TGGGGTGC-TGCAGCCC
R_DJ_289297_289418 ---GCGCACTGTCCCTGATTCCGTTGTATGCTGGGAT------TGTGGTGC-TGCAGCCC
R_DJ_289508_289645 ----CGCACTATCCCTGATCCCGGTGCATGATGGAAA------TGTAGTCC-TTCAGCCC
R_DJ_289651_289787 AACGCGCACTATCCTTGATCTTGGTGCATACTGGGAT------TGTAGTTC-TGCGGCCC
R_DJ_289785_289923 ---GCGCAATTTCCTTGAGCCCGGTGCTGGCTGGGAG------TGTAGTCC-TGCAGCCC
R_DJ_289924_290063 --------CTGTCCCTGAGCCCAGTGCATGCTGGGGCTGGAAGTGTAGTCC-TTCAGCCC
R_DJ_290069_290209 ---GCGCACTGTCCCTGAGCTGGGTGCATGCTGGGAT------TGCAGTCCTTGGCGATC
R_DJ_290212_290349 ---GCGCACTGTCACTGAGCTGGGTGCATACTAGGAA------TGTATTCC-TGCAGCCC
R_DJ_290358_290486 -------ACTGTCCGTGAGCCCCAAGCATGCTGGGAT------TGTAGTC-TTATAGCAC
R_DJ_290492_290609 ------------------------TGCATTCTGGGAT------TGGAGTCC-TGCAGTCC
R_DJ_290631_290763 ---------TGTCCGTGAGCCCCAAGCATGCTGGGAT------TGTAGTC-TTATAGCAC
R_DJ_290771_290886 --------------------------CATACTGGGAT------TGTAGTCC-TGCAGCCC
R_DJ_290910_291045 ------CACAGTCCCTGAACTCGCTGCATGCTGGGAT------TGTAGTCC-TGCAGCCC
R_DJ_291093_291231 ---GCGAACTGTCCCGGTGCCCCGTGCATGCTGGAAT------TGTAGTCC-TACAGCTA
R_DJ_291233_291370 ---GCGCGCTCTCCCTGAGCACGGTGCATGCTGGGAT------TGTAGTCC-TGAAGCCC
R_DJ_291530_291663 -------ACTTTCCCTGAGCCCGGTGCATGCTAGGAT------TGTAGCGC-TGCAGCCC
R_DJ_291668_291808 AATGCGCACTGTCCCTGAGCTGGGTGCATGCTGGGAT------TGTAGTCC-TGCAGCCC
R_DJ_291809_291944 ---GCGGACTGTCCCTGACCCCAGTGCATACTGGGAT------GGTAGTCC-TGCATCCC
R_DJ_291945_292085 -ATACGCATTGTCCCTGAGACAGGTGCATGCTAAGAT------TGTAGTCC-TGAAGCTC
R_DJ_292085_292203 ----------------------GGCGCATGCTGGGAT------TGTAGTCC-TGTAGCCA
DJ_136616_136731   TTTGACCAAAGGGCTGGGAGTGTTTATAAGAATACATCTCCCAGCAAG-CCGAGGGAGAC
DJ_136738_136867   TGTGACACAAGTTCTGGTAGTCTTTATGAAACTACATCTCCCAGCAAG-CAGAAGGAGGC
DJ_136868_137009   GGTGATGAGAGGTCTGGGAGTGTTTATGAGACTGCAACTCCCACCAAG-CCCAGAGAGGC
DJ_137010_137147   AGTGACCAAAGGGCTGGGAGTGTTTATGAGACTGCATCTCCCAGCAAG-ACCAGCGAGGT
DJ_137151_137289   AGTGACCAAAGGGCTGGGAGTGTTTACGAGAATACGTATCCCAAAAAG-CATAGCGAGAA
DJ_137290_137427   TGTGACCAAAGGGCTGGGAGAAATAAAGAGACAACATCTCCCAGAAAG-CCCAGCAAGGC
DJ_137429_137567   TGTGATGAAAGGGCTGGTAGTGTTTATGAGACTACCTCTCCCAGCAAG-CCCAGAGAGGT
DJ_137572_137706   TGCGACCAAACGACTGGG-GTAGTTATGAGACTGCATCTCCCTGCAAG-CCCAGCGAGGC
DJ_137711_137834   TGTGACCAAAGGGCCAA----------GAGACCACATCTCCCAGAAGA-CCTAGGGAGAC
DJ_137857_137974   TGTGACAAAAGGTCTGAGAGTCTTTATGAAACAACATCTCCCAGCAAA-CGCAGCGAGGT
DJ_137974_138108   TGTGACCATAGGGCAGGGAGAGGCCATGGGACTACATCTCCCAGGAAG-CCCAGCAAGGC
```

```
DJ_138117_138253   TATGACCAAAGGGATGGGAGTGTTTATGAGAATACATCTCCCAGTACG-CCCAG-GAGGT
DJ_138257_138396   TGTGACCAAAGGGCTGGGAGCGTTAATGAGACTACATCTCCCAAAAAATCACAGCTAGAA
DJ_138397_138541   TGTGATGAAAGGGCTGGGAGGTTTTATGAGAATACAACTCCCAGCAAG-CCTGGCGAGTA
DJ_138607_138743   TATGACCAAAGGGTTGGGTATGTTTATGAGAATACATATCCCACCAAG-TCCAGCGAGGC
DJ_138744_138871   TGTAATGAAAAGTCTGGGTGTCTTTATGAAACTACATCTCCCAGGAAG-CCAAAGGAGGC
DJ_138885_139020   TGTGACCAAAGGGCTGGGAGTGTTTATGAGACAGCATCTCTCAGCAAG-CAAAGCAAGGC
DJ_139111_139228   TGTGACCAAAGGGCTGGGAGTCTTTATAAGACTACATCTCCCAGCAAG-CACAA-GAGGT
R_DJ_289297_289418 TGTGACAAAGGGGCTGGGAGTCTTTATAAGACTACATCTCCCAGCAAG-CCCAA-GAGGC
R_DJ_289508_289645 TGTGACCAAAGGGCTGGGAGTGTTTATGAGACCGCATCTCTCAGCAAC-TAAAGCAAGGC
R_DJ_289651_289787 TGTAATGAAAGGTCTG-GTGACTTTATGAAACTACATCTCCCAGCAAG-CCAAAGGAGGC
R_DJ_289785_289923 TGTGACCAAAGGTTTGGGAGTATTTATGAGAATACATATCCCACCAAG-CCCAGCGAGAC
R_DJ_289924_290063 TGTGATGATAGGGCTGCGAGGATTTATGAGAATACATCTCCCAGCAAG-CCCAGCGAGTA
R_DJ_290069_290209 TATGACCAAAGGGCTAGGAGTGTTAATGAGACTACATCTCCCAAAAAAGCAGAGTGAGAA
R_DJ_290212_290349 TGTGAGCAAAGAGCTGGGAGTGTTTATGAGAATACATCTCCCAGTACT-CCCAG-GAGGT
R_DJ_290358_290486 CGTGACCAAAGGACAGGGAGAGGCCATGAGACTACAACTCTCAGGAAA-CCCAGCAAGGC
R_DJ_290492_290609 CGTGACAAAAGGTCTCAGAGTCTTTATGAAACTACATTTCCTAGCAAG-TGCAGCGAGGT
R_DJ_290631_290763 TGTGAGCCAAGGGTAGGGAGAGGACACGAGACTACATCTCCGAGAAAA-CCTAGGGAGAC
R_DJ_290771_290886 TATGACAAAAGGTCTGAGAGGCTTTATGAAACTACATTTCCCAAGAAG-CGCAGCGAGGT
R_DJ_290910_291045 TGTGACCAAAGGGCCAAGAGTGTTTATGAGACTACATCTCCCAGAAAA-CGTAGGGAGAG
R_DJ_291093_291231 TGTGATGAAAGGGCTGGGATTGGTTATCAGAATGCATCTCCCAGCAAG-CCCAGCGAGGC
R_DJ_291233_291370 TGTGACAAAAGGGCTGGGAGAAATAATGAGACTACATCTCCCAGAAAG-CCCAGCGAGAC
R_DJ_291530_291663 TGTGACAAAAGGGCTGTGAGTGTTTATGAGACTGCATCTTCCACCAAG-CCCAGAGAGGC
R_DJ_291668_291808 GGTGATGAAAGGTCTGGGAGTGTTTATGAGACTACATCTCCTACCAAG-CCCAG-GAGGT
R_DJ_291809_291944 TGTGATGAAAGTTCTGGGAGTCTCTATGAAACTACCTCTCCCAGGAAG-CAGAAGGAGGG
R_DJ_291945_292085 TGCGACCAAAGGGCTGGGAGTGTTTATGAGCATATATCTCCCAGCAAG-CCTAGGAAGAC
R_DJ_292085_292203 TTTGACCAAAGGGCTGGGAGTGTTTATGAGAATACAACTCCCAGCAAT-CCTAGGGAGGA
```

Appendix Table 11: Consensus sequence of ACRO138 repeats.

```
> consensus_seq_of_ACRO138
GCGCACTGTCCCTGAGCCCGGTGCATGCTGGGATTGTAGTCCTGCAGCCCTGTGACCAAAGG
GCTGGGAGTGTTTATGAGACTACATCTCCCAGCAAGCCCAGCGAGGCGCNCACAGCCCCGCC
TCTTCCTCCAGTGA
```

[Blank Page]

Appendix II

Statement of contributions

[Blank Page]

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Saumya Agrawal

Name/Title of Principal Supervisor: Dr. Austen Ganley

Name of Published Research Output and full reference:
The shared genomic architecture of human nucleolar organizer regions.
Floutsakou, I., Agrawal, S., Nguyen, T. T., Seoighe, C., Ganley, A. R. D., & McStay, B. (2013). The shared genomic architecture of human nucleolar organizer regions. Genome research. doi: 10.1101/gr.157941.113.

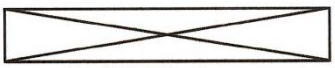
In which Chapter is the Published Work: Chapter 3

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

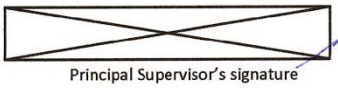
  and / or

- Describe the contribution that the candidate has made to the Published Work:
  Saumya's work accounts for approximately 30% of this paper. In particular, she has performed most of the bioinformatics characterizations of these flanking regions, aside from the ChIP-seq and RNA-seq mapping components, and these are the parts described in her thesis.


Candidate's Signature

Saumya Agrawal

20 - Dec - 2013
Date


Principal Supervisor's signature

20 December 2013
Date

[Blank Page]

Appendix III

Publication arising from this work

[Blank Page]

The shared genomic architecture of human nucleolar organizer regions

Ioanna Floutsakou, [1,4] Saumya Agrawal, [2,4] Thong T. Nguyen, [3,4] Cathal Seoighe, [3] Austen R.D. Ganley, [2] and Brian McStay [1,5]

[1]Centre for Chromosome Biology, School of Natural Sciences, National University of Ireland, Galway, Galway, Ireland; [2]Institute of Natural and Mathematical Sciences, Massey University, Auckland 0632, New Zealand; [3]School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Galway, Ireland

The short arms of the five acrocentric human chromosomes harbor sequences that direct the assembly and function of the nucleolus, one of the key functional domains of the nucleus, yet they are absent from the current human genome assembly. Here we describe the genomic architecture of these human nucleolar organizers. Sequences distal and proximal to ribosomal gene arrays are conserved among the acrocentric chromosomes, suggesting they are sites off requent recombination. Although previously believed to be heterochromatic, characterization of these two flanking regions reveals that they share a complex genomic architecture similar to other euchromatic regions of the genome, but they have distinct genomic characteristics. Proximal sequences are almost entirely segmentally duplicated, similar to the regions bordering centromeres. In contrast, the distal sequence is predominantly unique to the acrocentric short arms and is dominated by a very large inverted repeat. We show that the distal element is localized to the periphery of the nucleolus, where it appears to anchor the ribosomal gene repeats. This, combined with its complex chromatin structure and transcriptional activity, suggests that this region is involved in nucleolar organization. Our results provide a platform for investigating the role of NORs in nucleolar formation and function, and open the door for determining the role of these regions in the well-known empirical association of nucleoli with pathology.

[Supplemental material is available for this article.]

A detailed description of how the genome is organized within the nuclei of human cells to facilitate proper cellular functions is one of the major unsolved problems in biology. While genome-wide technologies are beginning to have an impact (Lieberman-Aiden et al. 2009; The ENCODE Project Consortium 2012), the picture is incomplete since critical regions of the human genome remain unidentified. Prominent among these missing regions is the chromosomal context around which nucleoli form, termed the nucleolar organizer region (NOR) (McClintock 1934). The nucleolus is the largest functional domain within the nucleus and is the site of ribosome biogenesis. It has a distinct structure and houses ribosomal RNA gene (rDNA) transcription, preribosomal RNA (pre-rRNA) processing, and preribosome assembly (Olson 2011). The rDNA repeats encode the major rRNA species and are organized into large head-to-tail tandem arrays located at the NORs (McStay and Grummt 2008). Extensive binding by the nucleolar DNA binding protein, UBF (encoded by *UBTF*), across the rDNA is responsible for their distinctive appearance on metaphase chromosomes, in which they form secondary constrictions (Mais et al. 2005; McStay and Grummt 2008). Despite ribosome biogenesis being central to cellular biology, many aspects of nucleolar formation, organization, and function remain to be elucidated, and the genomic architecture of NORs has not been described for any vertebrate to date.

In humans, the approximately 300 rDNA repeats are distributed among five NORs on the short arms of the acrocentric chromosomes (HSA13-15, HSA21, and HSA22) (Henderson et al. 1972; Schmickel 1973; Stults et al. 2008). In most human cells, a majority of NORs are active and coalesce to form between one and three nucleoli (Savino et al. 2001). While discrete, nucleoli are not encapsulated and instead seem to be self-organizing, highly dynamic structures (Dundr et al. 2002; Andersen et al. 2005; Sirri et al. 2008) that are spatially isolated from the rest of the nucleoplasm by a shell of heterochromatin (Nemeth and Langst 2011). However, the apparent heterochromatic nature of the regions flanking the rDNA has made them a low priority for genomic analysis (International Human Genome Sequencing Consortium 2004), and thus, the five acrocentric chromosome short arms are missing from the current human genome assembly.

The enormous demand for ribosomes by actively growing cells puts nucleoli at the forefront of cellular growth regulation. Links between nucleoli and growth pathologies date back over a century to observations of abnormal nucleoli in tumor cells (Pianese 1896). Recently, molecular studies have begun to clarify this relationship, with evidence supporting roles for tumor suppressor genes and oncogenes in rDNA transcriptional regulation (Budde and Grummt 1999; Hannan et al. 2000; Grummt 2003; Grandori et al. 2005), and a direct role for increased rDNA transcription in the development of malignancy (Bywater et al. 2012). Surprisingly, recent studies have also revealed that the nucleolus plays roles in many other biological processes, ranging from aging and cell cycle progression to X-chromosome inactivation and viral replication (Visintin et al. 1999; Boisvert et al. 2007; Zhang et al. 2007; Ganley et al. 2009). The central role of the nucleolus in growth regulation, coupled with the potential for the regions adjacent to the rDNA to contribute to NOR

[4]These authors contributed equally to this work.
[5]Corresponding author
E-mail brian.mcstay@nuigalway.ie
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.157941.113.
Freely available online through the Genome Research Open Access option.

function, prompted us to investigate the genomic context in which the rDNA resides. Here we present the identification and characterization of >550 kb of sequence flanking both sides of the rDNA array.

Results

Identification of rDNA flanking regions

To obtain the sequences flanking the rDNA arrays, we made use of preexisting sequences adjacent to the rDNA on the proximal (centromeric; 493 bp) (Sakai et al. 1995) and distal (telomeric; 8.3 kb) sides of the rDNA (Gonzalez and Sylvester 1997). Probes derived from these sequences were used to screen single-chromosome cosmid libraries, and several positive clones were sequenced. Searches of GenBank using the resulting sequences, coupled with BAC walking, ultimately resulted in identification of 15 BAC clones from the distal junction (DJ), five of which include some rDNA sequence (Supplemental Fig. 1). Similar searches identified three BAC clones from the proximal junction (PJ), one of which includes rDNA (Supplemental Fig. 2). These cover 379 kb of DJ sequence and 207 kb of PJ sequence flanking the rDNA (Fig. 1A).

We sought evidence that these putative junction sequences adjoin the rDNA. Hybridization of DJ BAC clones to metaphase chromosome spreads places these regions distal to the rDNA on the acrocentric chromosome short arms (Fig. 1B). Further, FISH on combed DNA molecules (Bensimon et al. 1994) clearly shows the DJ region is adjacent to the rDNA (Fig. 1C). Finally, a bioinformatic screen confirmed that the DJ adjoins the rDNA (Supplemental Methods).

Similar evidence was more difficult to obtain for the PJ, particularly with hybridization-based approaches. Metaphase FISH shows the PJ to be centromere proximal to the rDNA, as expected, but additional signals are observed distal to the rDNA and on other chromosomes (Fig. 1B; Supplemental Fig. 3), and clear signals were not observed from DNA combing. These difficulties stem from the high level of PJ segmental duplication (see below). Therefore we sought additional sequence-based evidence for the PJ adjoining the rDNA. First, we sequenced PJ-containing cosmids to show that the PJ is linked to at least 16 kb of rDNA (Supplemental Fig. 4). The PJ contig has previously been identified as a chr21 short-arm pericentromeric region (Lyle et al. 2007). To rule out the possibility that the PJ is adjacent to a large piece of segmentally duplicated rDNA, we compared rDNA external transcribed spacer (ETS) and 18S rRNA coding regions adjoining the DJ (a true junction) and the PJ. If the PJ is adjacent to duplicated, nonfunctional rDNA, it should have accumulated mutations, yet we found these PJ rDNA sequences to be 98.8% identical to the DJ rDNA. This is much higher identity than known 18S/ETS segmental duplicates (maximum of 93.2% identity), suggesting that the PJ-linked rDNA is intact. Finally, we implemented a bioinformatic strategy to look for sequence reads that span the PJ/rDNA junction (Supplemental Fig. 5). The positions of the PJ/rDNA junction in the cosmids and BACs are slightly different, and we found both these PJ junction positions with no evidence for additional PJ junctions (Supplemental Methods). These results suggest that the PJ adjoins the rDNA, although the precise junction position was found to vary slightly among the acrocentric chromosomes (Supplemental Fig. 6).

Interchromosomal conservation of rDNA flanking regions

Previous reports suggested that the sequences distal to the rDNA are conserved across all acrocentric chromosomes (Worton et al. 1988; Gonzalez and Sylvester 1997). Our DJ metaphase FISH results are consistent with this, showing hybridization signals for all the acrocentric chromosomes (Fig. 1B). The FISH also suggests that the PJ is conserved across the acrocentric chromosomes (Fig. 1B). To confirm this and to further validate the integrity of the flanking region sequences, we screened genomic DNA from a panel of mouse somatic cell hybrids, each containing a single human acrocentric chromosome for flanking region sequence. PCR at five intervals across the DJ each gave the expected product for all five acrocentric chromosomes (Fig. 2A). We could not screen across the PJ because only one unique region is present (see below). Nevertheless, this region is also shared among all five acrocentric chromosomes (Fig. 2A). Furthermore, screening genomic DNA from lines containing a chr21 translocation originating within the rDNA confirmed the orientation of the DJ and PJ relative to the rDNA (Fig. 2A) and that the unique PJ region is located exclusively on the proximal side. These results suggest that both the DJ and the PJ are conserved across all five acrocentric chromosomes.

To quantify how similar the DJ and PJ regions are between the different acrocentric chromosomes, we determined

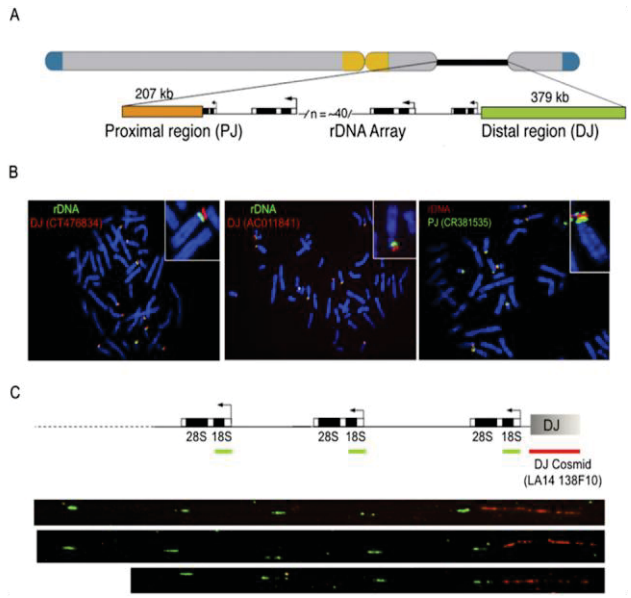


Figure 1. Human rDNA flanking regions. (*A*) Schematic human acrocentric chromosome showing telomeric (blue) and centromeric (yellow) regions, and the NOR (black line), expanded *below* into rDNA, PJ (orange), and DJ (green) regions. Not to scale. (*B*) DJ and PJ localize distally and proximally to rDNA, respectively, on all acrocentric chromosomes. FISH was performed on normal human metaphase spreads with DJ BAC (red) and rDNA (green) probes (*left* panels), and PJ BAC (green) and rDNA (red) probes (*right* panel). Chromosomes are DAPI-stained. (*C*) DNA combing of HeLa cell nucleolar DNA shows DJ (red) is physically linked to 18S rDNA (green). Three representative images are shown *below* the hybridization scheme.

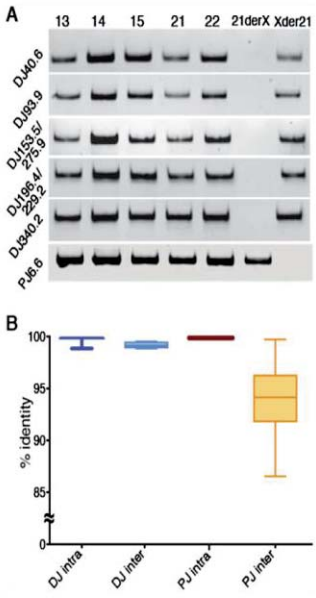


Figure 2. DJ and PJ acrocentric chromosome conservation. (*A*) PCR performed at increasing distances (*left*) into the DJ from mouse somatic cell hybrids carrying a single human acrocentric chromosome (indicated *above*). The *right-hand* lanes show PCR performed on the reciprocal products (Xder21 and 21derX) of a chr21 translocation that originates in the rDNA, confirming the DJ is located distally to the rDNA. *Bottom* panel is the same, but uses primers to the single unique PJ region. (*B*) Average intrachromosomal and interchromosomal DJ and PJ sequence identities from pairwise comparisons of representative BAC and cosmid clones are plotted.

BAC and cosmid sequence identities from intra- and interchromosomal pairwise comparisons for PJ and DJ clones (Fig. 2B; Supplemental Data 1). Intrachromosomal sequence identities approach 100% for both DJ and PJ sequences, as expected, with any differences likely resulting from sequence error. DJ interchromosomal identity averages 99.1%, suggesting there is a very active homogenization mechanism that maintains DJ sequence identity between the acrocentric chromosomes. PJ interchromosomal sequence identity is lower, averaging 93.3%, with the variation predominantly arising from interchromosomal polymorphisms in the rDNA junction position and Alu elements (Supplemental Fig. 6). This suggests there is also an active homogenization mechanism that maintains PJ sequences across acrocentric chromosomes (Fig. 2B).

Characterization of DJ and PJ sequences

The DJ and PJ are uncharacterized regions of the human genome, so we next employed a series of bioinformatic approaches to determine whether they harbor genomic features of interest. Consensus DJ and PJ contig sequences (Fig. 3A) were generated from a minimum set of overlapping BACs, using BACs from the same chromosome where possible (Supplemental Fig. 7).

To determine whether these regions primarily consist of repetitive elements characteristic of constitutive heterochromatin (International Human Genome Sequencing Consortium 2004), we characterized the repeat composition of the DJ and PJ. The transposable repeat element content of both regions is similar to that of the human genome as a whole (Supplemental Fig. 8), except for a lack of known DNA transposons in the DJ. Both regions also contain blocks of satellite repeats, most notably a large (38.6 kb) block of 48-bp satellite repeats at the distal end of the DJ (Fig. 3A). These repeats, initially classed as chr22 pericentromeric repeats (Metzdorf et al. 1988), are now referred to as CER satellites (Jurka et al. 2005). CER blocks are found distal to the rDNA on all acrocentric chromosomes, with additional pericentromeric blocks on chr14 and chr22 (Supplemental Fig. 9). Finally, we looked for novel repeats. Two blocks of a novel 138-bp tandem repeat that we call ACRO138 are present within the DJ (Fig. 3A; Supplemental Fig. 10). Most strikingly, we discovered a large inverted repeat that dominates the DJ (Fig. 3A). Each arm of the inverted repeat is ~109 kb, and the two arms share an average sequence identity of 80%. Alignment of the two arms reveals that the underlying sequence identity is higher than this but is interrupted by a number of indels (Supplemental Fig. 11), implying either that the inverted repeat is young or that there are mechanisms to maintain sequence identity between the arms.

Although these rDNA flanking regions were thought to be heterochromatic (International Human Genome Sequencing Consortium 2004), we wondered whether they contain any genes. To address this, we designed a gene predication pipeline that integrates ORF prediction, mRNA, EST, and protein data to identify potential gene-coding regions. Eight putative DJ genes and four PJ genes were predicted. These gene models all had support from multiple data sources, but the majority are single exon (Supplemental Fig. 12; Supplemental Tables 1, 2). Experimental validation will be required to determine which of these putative genes are real.

Segmental duplications (duplications of DNA to elsewhere in the genome) are a prominent feature of the human genome and are commonly enriched at centromere boundaries (She et al. 2004; Bailey and Eichler 2006). Given the proximity of the DJ and PJ to centromeres we undertook a segmental duplication analysis. A number of segmental duplications (>1 kb in length and with >85% sequence identity) were found from both regions (Fig. 3B). Interestingly, the DJ and PJ show different segmental duplication patterns. PJ segmental duplicates are more frequent and longer and have greater sequence identity than DJ segmental duplicates (Fig. 3B; Supplemental Tables 3, 4). Furthermore, the majority of PJ segmental duplicates are found in centromeric/pericentromeric regions of the genome as previously observed (Piccini et al. 2001; Lyle et al. 2007), while the majority of DJ segmental duplicates are found in euchromatic/telomeric regions (Fig. 3B). Most strikingly, we found that the level of segmental duplicated DNA is vastly different, with 7.3% of the DJ being segmentally duplicated versus 92.4% for the PJ (Table 1). These results demonstrate that these two rDNA flanking regions have different genomic characteristics in humans: The segmental duplication profile of the PJ resembles pericentromeric regions, while that of the DJ resembles euchromatic regions. The high level of segmental duplication likely explains the problems encountered using hybridization-based approaches with the PJ. Finally, this high level of segmental duplication prevented us from extending the PJ further, as we could not unambiguously assign additional sequences to the PJ region.

Perinucleolar positioning of the DJ

To better understand the relationship between these flanking regions and nucleolar architecture, we analyzed DJ and PJ positioning in interphase nuclei of HT1080 cells. 3D immuno-FISH revealed that, despite their close proximity to rDNA, DJ sequences

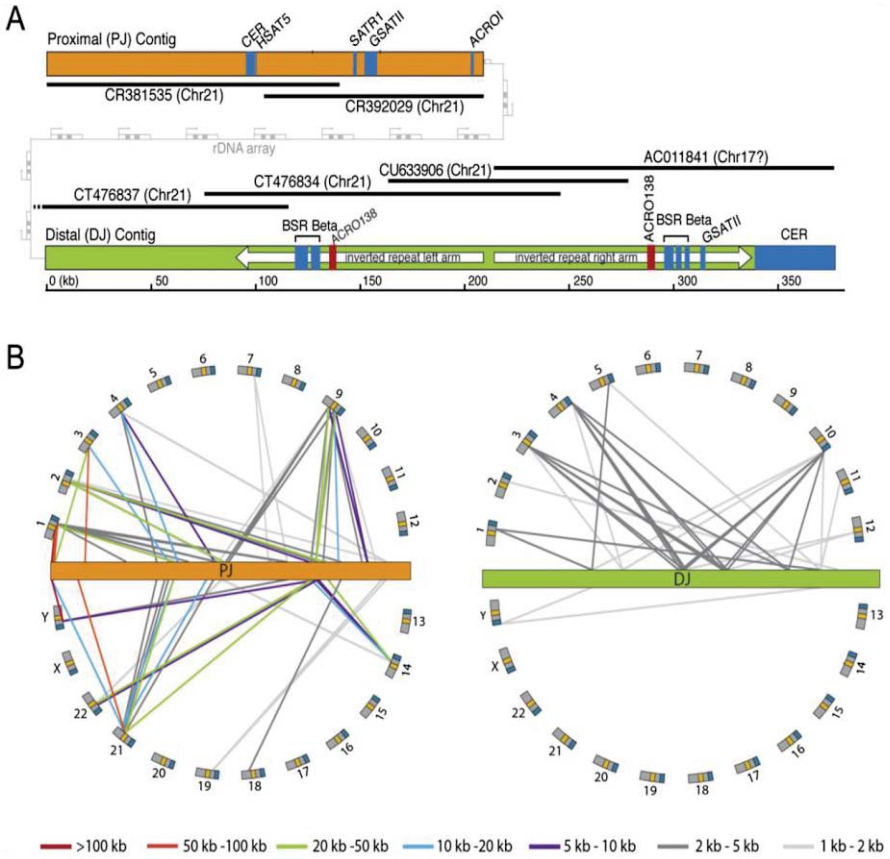


Figure 3. DJ and PJ sequence characterization. (*A*) Major genomic features of the PJ (orange) and DJ (green) contigs. BACs used to construct the contigs are shown as black lines, with BAC names and chromosomal origins indicated (chr17 annotation of AC011841 is incorrect). Satellites are shown in blue, the ACRO138 repeats in red, the large DJ inverted repeat as white arrows, and the rDNA array between the PJ and DJ in gray. (*B*) Segmental duplication analysis. Lines show segmental duplications from PJ (orange) and DJ (green), indicating the location of the duplicate on the human chromosomes (arranged around the flanking regions). Positions of centromeres (yellow) and telomeres (blue) are indicated. Segmental duplicate length is indicated by line color, as defined *below*.

are excluded from the nucleolar interior, instead appearing as discrete foci embedded in the perinucleolar heterochromatin (Fig. 4A). The majority of DJ foci associate with perinucleolar heterochromatin (Supplemental Videos 1, 2). The high level of PJ segmental duplication makes it difficult to determine which FISH signals derive from the PJ versus from unlinked segmentally duplicated regions (Supplemental Fig. 13); therefore, we focused on the DJ for the remainder of this study.

The rDNA is transcribed by RNA polymerase I (pol-I), and inhibition of pol-I transcription by low concentrations of actinomycinD (AMD) induces rapid and remarkable nucleolar reorganization (Hadjiolov 1985). This involves dissociation of nascent rRNA transcripts, resulting in collapse of the rDNA repeats into nucleolar caps that form at the nucleolar periphery and consist of rDNA, pol-I transcription machinery, and processing factors (Supplemental Fig. 14; Prieto and McStay 2007; Sirri et al. 2008). Given the localization of the DJ, we wondered whether it has any relationship to these nucleolar caps. Strikingly, 3D immuno-FISH on AMD-treated cells revealed that nucleolar caps form immediately adjacent to DJ sequences positioned in perinucleolar heterochromatin (Fig. 4B). Although some larger nucleolar caps are bilobed and associated with two DJ signals, it appears that the majority of caps are derived from individual NORs. These results suggest that the DJ anchors the linked rDNA to perinucleolar heterochromatin and that the retreat of the rDNA to the DJ upon AMD treatment provides an explanation for the positioning of nucleolar caps (Fig. 4B).

DJ sequences can target perinucleolar heterochromatin

rDNA arrays regress to perinucleolar heterochromatin in AMD-treated cells, rather than DJ sequences moving toward rDNA foci within the nucleolar interior (Fig. 4). This suggests that elements within the DJ may be responsible for its perinucleolar localization,

Table 1. Segmental duplication comparison between the DJ and PJ

Segmental duplication feature	DJ	PJ
Number of segmentally duplicated regions	31	98
Average segmental duplicate length (kb)	2.3	11.8
Average percent identity between segmental duplicates	88.2%	93.1%

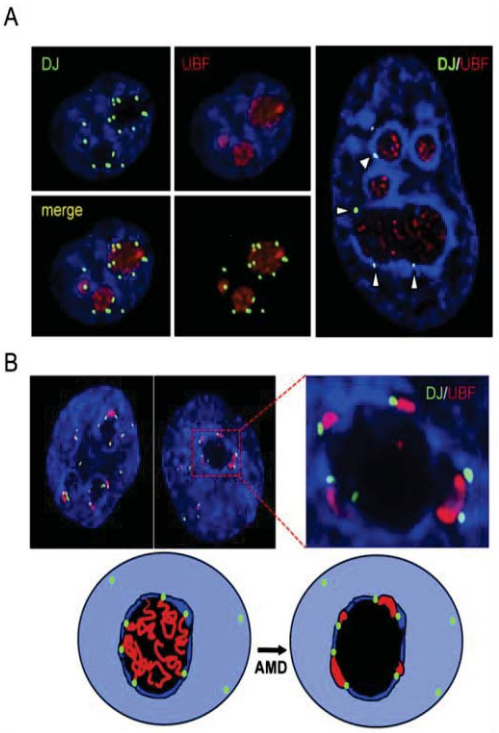
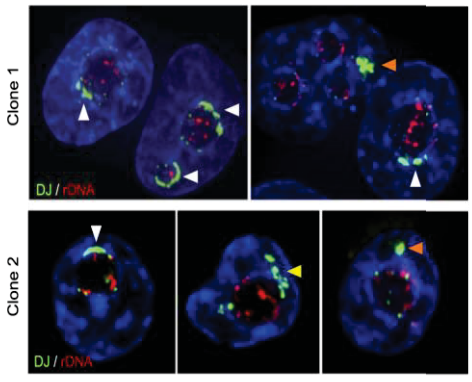


Figure 4. The DJ forms a perinucleolar anchor for rDNA repeats. (*A*) 3D immuno-FISH reveals that DJ sequences lie in perinucleolar heterochromatin in HT1080 cells. Nucleoli are visualized with UBF antibodies (red) and DJ with BAC CT476834 (green). The nucleus is DAPI-stained. The extended focus images (*left*) are stills from Supplemental Videos 1 and 2, while the image on the *right* shows a single focal plane. (*B*) Inhibition of rDNA transcription with AMD results in formation of nucleolar CAPs juxtaposed with DJ sequences in perinucleolar heterochromatin. Staining as in *A*. Two representative cells are shown, one with an enlargement. Cartoon models the transition between active and withdrawn rDNA upon AMD treatment. rDNA (red) retreats from the nucleolus (black) to the DJ (green) that is embedded in perinucleolar heterochromatin (dark blue).

rather than the DJ being simply linked to the rDNA. To investigate this, we transfected HT1080 cells with a mixture of three BACs that cover the DJ contig. Two stable clones containing large integrated arrays of this BAC mixture were selected for further analysis. FISH with DJ BAC CT476834 and rDNA probes revealed that ectopic DJ BAC arrays had integrated into metacentric (non-NOR bearing) chromosomes (Supplemental Fig. 15A). Quantitative PCR using primer pairs positioned across the DJ contig revealed that the sequence content of the ectopic arrays reflects that of the input BAC mixture (Supplemental Table 5). 3D FISH was performed on cells from these clones to determine whether the ectopic DJ arrays associate with nucleoli. In order to more clearly reveal the boundaries of the nucleolus, cells were treated with AMD (Fig. 5). For both clones, we observe a remarkable degree of association of the ectopic DJ arrays with perinucleolar heterochromatin. Moreover, in the majority of cells, the ectopic DJ array appears to spread through the perinucleolar heterochromatin, covering a significant fraction of the nucleolar surface. Transcription from the small amounts of rDNA derived from BAC AL592188 does not appear to explain this localization, as silver staining (which can detect activity at NORs with comparable rDNA content; A Grob and B McStay, in prep.) shows no activity. Additionally, transcriptionally active ectopic rDNA arrays (neo-NORs) that do not contain any DJ sequences are usually not associated with endogenous nucleoli (A Grob and B McStay, in prep.). Therefore, we conclude that sequences within the DJ contig specify association with perinucleolar heterochromatin even when positioned on metacentric chromosomes. (Supplemental Fig. 15B).

Chromatin profiling of the DJ

Our results suggest that the DJ region plays a role in nucleolar organization. Specifically, we hypothesize that DJ sequences provide an anchor point within perinucleolar heterochromatin for the linked rDNA array that is normally present in the nucleolar interior. If so, the DJ may have a chromatin organization that facilitates this role. To profile its chromatin organization, we mapped available histone modification and insulator binding protein CTCF ChIP-seq data sets from the ENCODE Project (Ernst et al. 2011) to the DJ. Discrete patterns of enrichment were observed at specific points across the DJ (Fig. 6A). Integration of these chromatin data sets using ChromHMM (Ernst and Kellis 2012) with multiple cell types revealed a complex chromatin landscape that is largely conserved among cell types (Fig. 6B; Supplemental Fig. 16). Strikingly, chromatin signatures characteristic of promoters are found at regular ~45-kb intervals across the DJ, interspersed among marks associated with heterochromatin. The periodicity (~45 kb) of these putative promoters is interesting, as it closely mirrors the size of the rDNA unit. Chromatin marks indicative of promoters (e.g., H3K4me3) coincide with DNase hypersensitive sites and FAIRE signals (Supplemental Fig. 17), and we experimentally confirmed that the H3K4me3 and FAIRE peaks are present in the HT1080 cell line used in our immuno-FISH experiments (Fig. 6C). The open chromatin peaks centered at 138 kb and 290 kb



	Associated ◁	Partial ◀	Non-associated ◀
Clone 1 cells n = 41	68%	27%	5%
Clone 2 cells n = 50	74%	18%	8%

Figure 5. Ectopic DJ arrays target perinucleolar heterochromatin. Positioning of DJ BAC arrays. 3D FISH was performed on AMD-treated cells from clones 1 and 2 with rDNA (red) and DJ BAC CT476834 (green) probes. The large green hybridization signals identified by arrowheads indicate the ectopic DJ array. Endogenous DJ signals are also visible. Classification of ectopic DJ arrays as nucleolar associated, partially associated, or nonassociated is indicated by white, yellow, and orange arrowheads, respectively, and is quantified *below*.

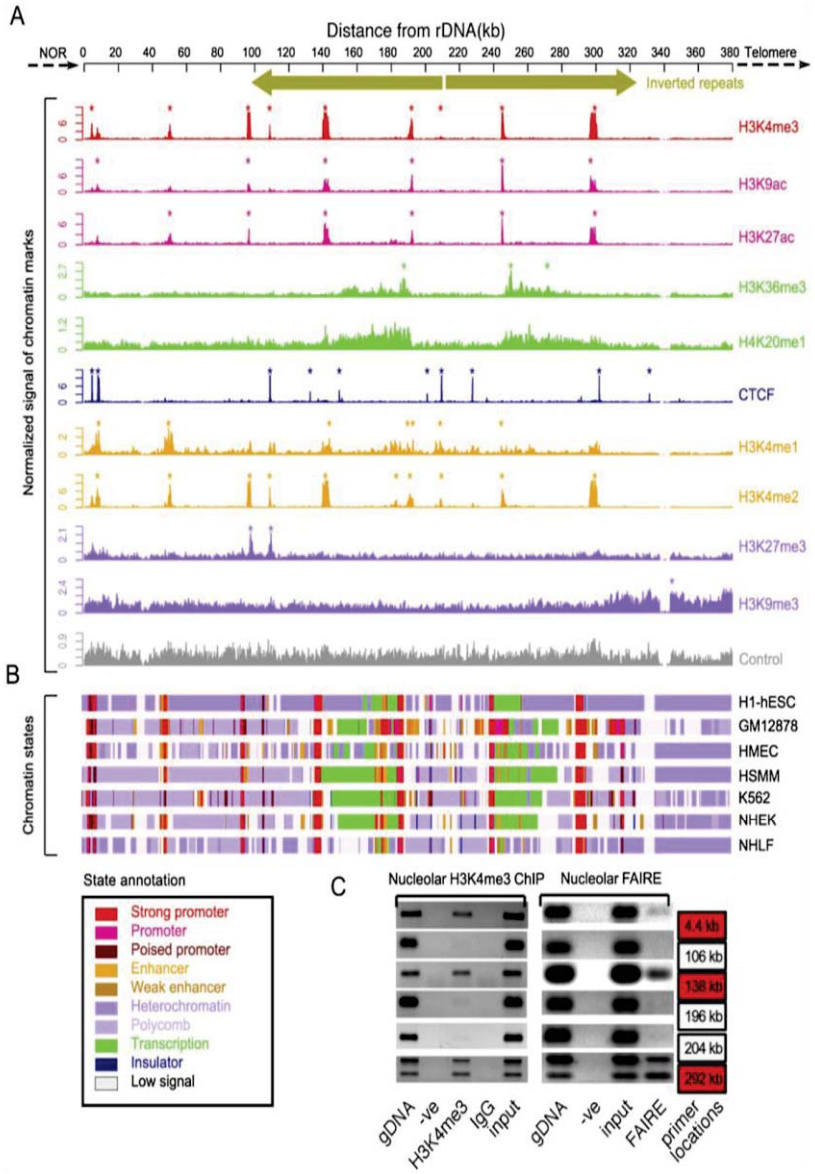


Figure 6. Chromatin landscape of the DJ. (*A*) ChIP-seq signals of different chromatin features (*right*) in H1-hESC cells, normalized to tags per million mapped reads are shown *below* a schematic of the DJ, including inverted repeats. Asterisks indicate enrichment sites. (*Bottom*) Control signal is shown in gray. (*B*) Chromatin states derived from the multivariate HMM analysis for seven different cell types (*right*). Each colored bar represents a specific chromatin state, as annotated *below left*. (*C*) Nucleolar H3K4me3 ChIP-PCR and nucleolar FAIRE-PCR using HT1080 cells validate the presence of H3K4me3 and FAIRE in the DJ. DJ positions of the primers used are shown to the *right*, and red boxes correspond to peaks of H3K4me3 from *A*. Genomic DNA (gDNA), input and negative controls (-ve and IgG) are shown.

correspond to the ACRO138 repeat blocks identified in the repeat analysis (Fig. 3A). Moreover, chromatin signatures associated with actively transcribed gene bodies (e.g. H3K36me3 and H3K20me1) (Ernst et al. 2011) are observed extending leftward and rightward from the promoters at 187 kb and 238 kb, respectively, (Fig. 6A). CTCF, a multivalent DNA binding protein involved in many cellular processes (Phillips and Corces 2009), has recently been shown to be involved in the transcriptional regulation of ribosomal genes (van de Nobelen et al. 2010) and human nucleolar organization (Hernandez-Hernandez et al. 2012). Multiple CTCF binding peaks were observed across the DJ, coinciding with CTCF consensus sequences (Supplemental Fig. 18). Interestingly, CTCF binding sites are positioned close to the DJ/rDNA boundary and frame the DJ transcription units described above. Together, these results reveal that the DJ has a complex and structured chromatin landscape.

Transcription profiling of the DJ

The chromatin profiling results suggest that despite being embedded in perinucleolar heterochromatin, DJ sequences are tran-

scriptionally active. To directly investigate transcription in the DJ, we mapped RNA-seq, RNA Pol II, and TAF1 ChIP-seq data (from ENCODE) and mRNA and EST data (from GenBank) onto the DJ (Supplemental Fig. 19). Strong evidence for transcription originating from the majority of putative promoters was obtained, confirming that the DJ is transcriptionally active. The putative promoters at 187 kb and 238 kb in particular are supported by multiple lines of evidence (Fig. 7A), with RNA-seq data indicating that the transcripts from these promoters are spliced, and cDNA clones of these transcripts being present in GenBank (accession nos. AK026938 and BX647680). We used RT-PCR to experimentally confirm the existence of these spliced polyadenylated transcripts, which we term disnor187 and disnor238 (Fig. 7B). In addition, we show using RNA-seq data that they have low to medium expression levels (Supplemental Fig. 20). The largest open reading frames present in disnor187 and disnor238 are 120 and 144 amino acids, respectively. Therefore their size and limited coding capacity suggest that they may function as long noncoding RNAs (lncRNAs). We also confirmed that the ACRO138 repeats within the open chromatin peak at 138 kb are actively transcribed (Supplemental Fig. 21). These three transcripts all lie within the DJ inverted repeat arms and, together with another putative promoter in the ACRO138 repeat block centered at 297 kb, form a symmetrical arrangement of transcriptional units that mirror the inverted repeat structure. Therefore the DJ, rather than being a passive block of heterochromatin, shows a specific pattern of localization, a distinct genomic and chromatin organization, and transcriptional activity.

Discussion

NORs were originally defined in 1934 as chromosomal regions that organize formation of the nucleolus (McClintock 1934). Now ~80 yr later, we can begin to appreciate their genomic architecture. In this study, we have identified >550 kb of sequence from the regions flanking the rDNA array and have performed an in-depth characterization of these regions. We reveal that the sequences flanking the rDNA are conserved across all five acrocentric chromosomes and have a complex sequence feature composition.

Sequences proximal to the rDNA almost entirely consist of segmentally duplicated regions, like those surrounding centromeres (She et al. 2004), and thus are unlikely to contain nucleolus-specific functional elements. The high level of interchromosomal sequence conservation of the PJ, coupled with its high level of segmental duplication, strongly suggests that, far from being a recombinationally inert region of the genome, the PJ experiences frequent and ongoing recombination. This recombination appears to occur predominantly with other peri-/centromeric regions of the genome, implying colocalization of these regions, and may be responsible for Robertsonian translocations that appear to derive from the pericentromeric regions of acrocentric short arms (Therman et al. 1989) and are associated with genetic disorders. Our identification of the sequence of the PJ provides a means to investigate this.

In contrast, the DJ region is replete with unique sequences and displays evidence of functionality. We show that the DJ is localized to perinucleolar heterochromatin, where it appears to anchor the rDNA array to this region, and sequences located within the DJ are likely to be important for this localization. We propose that the DJ acts as a "control panel" for the entire NOR where it can determine the transcriptional status of the linked rDNA array. In this way, the DJ may be involved in regulating the coalescence of nucleoli around individual NORs. Our model is that active NORs are localized to perinucleolar heterochromatin, where they form nucleoli, while inactive NORs lose this localization and form silent arrays that do not participate in nucleoli.

The sequences we describe begin to close one of the major remaining gaps in the human genome, the short arms of the acrocentric chromosomes, and their identification is an important step toward a complete understanding of nucleolar biology. The level of segmental duplication we observe, particularly of the PJ, suggests that some previously identified nucleolus-associated chromatin domains (NADs) may actually be segmental duplicates (Nemeth et al. 2010; van Koningsbruggen et al. 2010), and our sequences will allow a more refined picture of NADs to be developed. Additionally, the DJ sequence will allow researchers to use hybridization-based approaches to determine whether human nucleoli contain multiple NORs (Supplemental Videos 1, 2), something that has remained difficult to prove. The degree of heterogeneity in nucleolar morphology observed between cancers (Derenzini et al. 2009) suggests that mechanisms other than direct up-regulation of rDNA transcription have a role in the development of malignancy. As acrocentric short arms underpin both nucleolar form and function, we hypothesize that genetic alterations on these chromosome arms contribute to tumorigenesis and other human diseases. Therefore the sequences that we

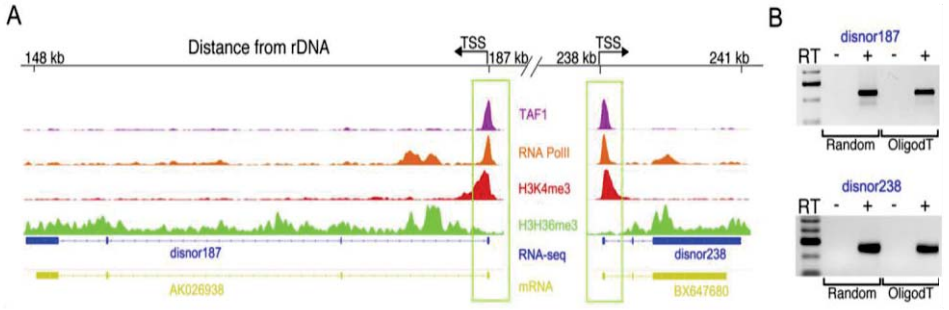


Figure 7. DJ transcript profiling. (*A*) ChIP-seq reveals chromatin features consistent with transcription originating from promoters at 187 kb and 238 kb (boxed) in the DJ. *Top* four tracks are an enlargement of selected chromatin features from Figure 6A. *Bottom* two tracks show RNA-seq and cDNA mapping results. Exons are indicated by blocks. These identify spliced transcripts (disnor187 and disnor238) similar to cDNA clones AK026938 and BX647690. (*B*) RT-PCR using primers to detect disnor187 and disnor238 transcripts in HT1080 cells. Random and oligo(dT)-primed RT-PCR products of the expected sizes for spliced transcripts were produced.

report here lay the foundations for addressing the roles that genetic and epigenetic changes in the DJ and PJ play in human disease, as well as providing a wealth of new tools for studying nucleolar biology.

Methods

Genomic cosmid and BAC clones

Acrocentric chromosome cosmid libraries LA13 NC01, LA14 NC01, LA15 NC01, LL21 NC02, and LL22 NC03 were obtained from the UK HGMP resource center. To obtain cosmids spanning the DJ, libraries were screened with a 638-bp PCR product generated using DJUf/DJUr primers (Worton et al. 1988). To obtain cosmids spanning the PJ, libraries were screened with a 220-bp PCR product generated using PJf/PJr primers (Sakai et al. 1995). Clone names for DJ and PJ cosmids identified and used in this study are shown in Supplemental Figure 4. We identified BAC clones in GenBank representing the DJ and PJ regions flanking the rDNA using BLAST. Cosmids LA14 138F10 and N 29M24 were used as the initial query sequences to search for BACs representing the DJ and PJ, respectively (Supplemental Methods). BAC clones were obtained from BACPAC Resources.

DNA sequencing and assembly

DNA sequencing of cosmid clones was performed through a combination of standard Sanger and next-generation sequencing (NGS). The sequence of the insert in cosmid LA14 138F10 and most of the insert in N 29M24 were determined by Sanger sequencing. The inserts of the remaining cosmids were end sequenced using Sanger sequencing and then subjected to NGS. Indexed libraries were prepared from individual cosmids using a Nextera DNA sample prep kit and Nextera barcodes (Epicentre NGS). NGS was performed on an Illumina Genome Analyzer IIx, using 54-bp singleton processing (Ambry Genetics). Sequences of cosmids were assembled using ABySS v1.2.7 (for parameters, see Supplemental Methods) (Simpson et al. 2009). Velvet v1.101 (Zerbino and Birney 2008) was used to refine the ABySS assemblies.

Cell lines

HeLa cells and HT1080 human fibrosarcoma cells were grown in DMEM supplemented with 10% fetal bovine serum (FBS). RPE-1 cells were maintained in DMEM:F12 medium supplemented with 10% FBS, 2 mM L-glutamine, and 0.348% sodium bicarbonate. Mouse A9 cells containing individual human acrocentric chromosomes were previously described (Sullivan et al. 2001), and those containing X/21 reciprocal translocation products (GM09142 and GM10063) were obtained from Coriell Cell Repositories. To generate cells that contain ectopic DJ arrays, HT1080 cells were cotransfected using a standard calcium phosphate protocol with BACs AL592188, CT476834, and AC011841 together with a blasticidin selection marker in a 200:1 w/w ratio. Stable transfectants were maintained as above but supplemented with 5 μg/mL blasticidin.

FISH and 3D immuno-FISH and RNA FISH

Probes for FISH experiments were labeled using spectrum red or green dUTP (Abbott Molecular). For chromosome mapping experiments, slides of human normal male metaphase chromosome spreads (Applied Genetics) were denatured in 70% formamide/2× SSC for 5 min at 73°C. Slides were then dehydrated through a 70%–100% ethanol series, washed, and air dried. Denatured probe (50 ng/slide) combined with human COT-1 DNA (10 μg/slide) in 20 μL/slide Hybrisol VII (Qbiogene) was then added to the slides and allowed to hybridize for 24–48 h at 37°C in a humidified chamber. For CER satellite detection, hybridizations were performed with a 5′ FITC-labeled oligo and herring sperm DNA. Post-hybridization washes were 0.4× SSC/0.3% NP-40 for 2 min at 74°C followed by 2× SSC/0.1% NP-40 at ambient temperature for 1 min. Slides were air dried and mounted in Vectashield, including DAPI (Vector Laboratories). For 3D immuno-FISH experiments, cells were fixed, denatured, probed, and antibody stained as described previously (Mais et al. 2005; Prieto and McStay 2007). Z-stacks of fluorescent images were captured using a Photometric Coolsnap HQ camera and Volocity 5 imaging software (Improvision) with a 63 Plan Apochromat Zeiss objective mounted on a Zeiss Axioplan2 imaging microscope. In some cases, extended focus projections of deconvolved Z-stacks (iterative restoration) are presented; in other cases, individual focal planes are shown. Movies (Supplemental Videos 1, 2) were prepared from 3D images constructed from deconvolved Z-stacks using Volocity 6 (Improvision). 3D images were rotated to create a series of bookmarks. The movies are an animation of the transitions between these bookmarks.

Nucleolar DNA combing

Nucleoli, prepared from HeLa cells as previously described (Andersen et al. 2002), were resuspended at a concentration of 1×10^6 to 2×10^6 cell equivalents/mL in TE (10 mM Tris at pH 8.0, 1 mM EDTA). Resuspended nucleoli were mixed with an equal volume of 1% low melting point agarose in TE at 50°C. The mixture was pipetted into a plug mould (BioRad, 100 μL/slot). Embedded nucleoli were deproteinized, and encapsulated high-molecular-weight nucleolar DNA was combed onto silanized coverslips as previously described (Caburet et al. 2005) using a Molecular Combing apparatus supplied by Genomic Vision Paris. Coverslips were then hybridized with biotin-labeled DJ cosmid LA14 138F10 and a digoxygenin-labeled 5.8-kb EcoRI restriction fragment that contains 5′ ETS and 18S rRNA from human rDNA. Hybridization and detection were performed as described previously (Caburet et al. 2005).

Bioinformatic analyses

The bioinformatics pipeline that was used to identify potential junction regions from whole-genome sequence data is described in Supplemental Methods, as is the method for determining intra-/interchromosomal sequence identity between DJ and PJ clones.

Construction and analysis of DJ and PJ contigs

To construct the DJ, four BACs were merged (CT476837, CT476834, CU633906, and AC011841). The overlapping regions between BACs CT476837 and CT476834 and BACs CT476834 and CU633906 are 100% identical, but the identity decreases to ~98% between CU633906 and AC011841 (Supplemental Fig. 7A). To form the PJ, two BACs (CR392039 and CR381535) with an identical overlapping region were merged to obtain a single contig (Supplemental Fig. 7B). The PJ is identical to that previously published (Lyle et al. 2007) and is deposited in GenBank under accession no. NT113958. The repeat content analysis method and the gene identification pipeline that utilizes gene prediction, mRNA, EST, and protein sequence data to identify potential DJ and PJ genes are both presented in Supplemental Methods. Segmental duplicates in the DJ and PJ contigs were detected using a modified BLAST-based detection scheme called the "whole genome assembly comparison" (Bailey et al. 2001). The human genome assembly (hg19) was

broken into 400-kb pieces. Repeats in this fragmented human genome and in the DJ/PJ contigs were masked using RepeatMasker (Smit et al. 2010). DJ and PJ contigs were then matched to the fragmented masked genome using BLAST, with a cutoff of ≥85% identity >1 kb. Next, the repeats were reinserted into these matched sequences, and global alignments were created. All steps were performed using a series of Perl scripts (J. Bailey, University of Massachusetts Medical School). Low identity ends of the sequences were identified from the alignments and trimmed. Where the ends of two human genome fragments match a single region or where a fragment is interrupted by repeats, these fragments were merged together. The merged sequences were then aligned again to recalculate the identity.

ENCODE data

ChIP-seq data for 10 chromatin marks (CTCF, H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K9me3, H3K27me3, H3K9ac, H3K27ac, and H4K20me1) and input were obtained for seven different cell types (GM12878, H1-hESC, HMEC, HSMM, K562, NHEK, and NHLF) from ENCODE Broad Histone (Ernst et al. 2011). DNase-seq and FAIRE-seq data were obtained from ENCODE UNC/Duke (Song et al. 2011). PolyA tailed RNA-seq data were obtained for 11 different cell types (GM12878, H1-hESC, HCT-116, HeLa-S3, HepG2, HSMM, HUVEC, K562, MCF-7, NHEK, NHLF) from ENCODE Caltech RNA-seq. GenBank mRNAs and ESTs data were downloaded from the UCSC Genome Browser (Fujita et al. 2011) on January 1, 2012. These data were mapped to the human genome to which the DJ sequences had been added (Supplemental Methods).

Nucleolar ChIP and FAIRE

ChIP was performed on nucleolar chromatin isolated from HT1080 cells with H4K4me3 antibodies (Millipore, catalog no. 04-745) as described previously (Mais et al. 2005). We adapted a FAIRE protocol (Giresi et al. 2007) for cross-linked nucleolar chromatin. One hundred microliters of HT1080 nucleolar chromatin was extracted using an equal volume of phenol/chloroform. DNA was recovered from the aqueous phase by ethanol precipitation and resuspended in 100 µL of TE buffer. PCR was performed using 2 µL of recovered DNA. DNA recovered from input nucleolar chromatin served as a control for ChIP and FAIRE experiments.

Transcriptome profiling

Paired-end RNA-seq data from the 11 cell types was mapped to the human genome with DJ sequences added using TopHat (v1.2.0) (Trapnell et al. 2009) with mostly default parameters (-r 50 -a 8). We then merged the output alignments from all replicates using SAMtools (Li et al. 2009). The result is used as the input for Cufflinks (v1.3.0), with default parameters, to assemble the transcriptome of the custom genome. Finally we merged all 11 assembled transcriptomes, corresponding to the 11 cell types, using Cuffmerge (Trapnell et al. 2010) to obtain the final transcriptome. We also used this final transcriptome to estimate DJ transcript abundance. We used BLAT (Kent 2002) to map the mRNA and EST data to the DJ using the parameters "-fine -q=rna -minIdentity=95 -maxIntron=70000", and "-minIdentity=97 -maxIntron=70000", respectively.

Data access

The DJ contig nucleotide sequence and feature list are available in Supplemental Data 2 and 3. Assembled DJ and PJ cosmid sequences are available from GenBank (http://www.ncbi.nlm.nih.gov/genbank) under accession numbers KC876024–KC876030. Raw NGS data have been submitted to the Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/Traces/sra/) under accession number SRP024282.

Acknowledgments

B.M. acknowledges Science Foundation Ireland (PI grant 07/IN.1/B924) for funding work in his laboratory. A.R.D.G. acknowledges the Auckland Medical Research Foundation, the Marsden Fund, the Royal Society of NZ, and the Bio-Protection Research Centre for funding. C.S. thanks Science Foundation Ireland (07/SK/M1211b) for funding. T.T.N. received a PhD fellowship from the Irish Research Council for Science, Engineering and Technology. We also thank Samantha Raggett, Jane Wright, and Mark Anderson for help in library screening and DNA sequencing during the early stages of this work, and Kevin Sullivan and Carol Duffy for comments on the manuscript.

References

Andersen JS, Lyon CE, Fox AH, Leung AK, Lam YW, Steen H, Mann M, Lamond AI. 2002. Directed proteomic analysis of the human nucleolus. *Curr Biol* **12:** 1–11.

Andersen JS, Lam YW, Leung AK, Ong SE, Lyon CE, Lamond AI, Mann M. 2005. Nucleolar proteome dynamics. *Nature* **433:** 77–83.

Bailey JA, Eichler EE. 2006. Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nat Rev Genet* **7:** 552–564.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res* **11:** 1005–1017.

Bensimon A, Simon A, Chiffaudel A, Croquette V, Heslot F, Bensimon D. 1994. Alignment and sensitive detection of DNA by a moving interface. *Science* **265:** 2096–2098.

Boisvert FM, van Koningsbruggen S, Navascues J, Lamond AI. 2007. The multifunctional nucleolus. *Nat Rev Mol Cell Biol* **8:** 574–585.

Budde A, Grummt I. 1999. p53 represses ribosomal gene transcription. *Oncogene* **18:** 1119–1124.

Bywater MJ, Poortinga G, Sanij E, Hein N, Peck A, Cullinane C, Wall M, Cluse L, Drygin D, Anderes K, et al. 2012. Inhibition of RNA polymerase I as a therapeutic strategy to promote cancer-specific activation of p53. *Cancer Cell* **22:** 51–65.

Caburet S, Conti C, Schurra C, Lebofsky R, Edelstein SJ, Bensimon A. 2005. Human ribosomal RNA gene arrays display a broad range of palindromic structures. *Genome Res* **15:** 1079–1085.

Derenzini M, Montanaro L, Trere D. 2009. What the nucleolus says to a tumour pathologist. *Histopathology* **54:** 753–762.

Dundr M, Hoffmann-Rohrer U, Hu Q, Grummt I, Rothblum LI, Phair RD, Misteli T. 2002. A kinetic framework for a mammalian RNA polymerase in vivo. *Science* **298:** 1623–1626.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Ernst J, Kellis M. 2012. ChromHMM: Automating chromatin-state discovery and characterization. *Nat Methods* **9:** 215–216.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473:** 43–49.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, et al. 2011. The UCSC Genome Browser database: Update 2011. *Nucleic Acids Res* **39:** D876–D882.

Ganley AR, Ide S, Saka K, Kobayashi T. 2009. The effect of replication initiation on gene amplification in the rDNA and its relationship to aging. *Mol Cell* **35:** 683–693.

Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res* **17:** 877–885.

Gonzalez IL, Sylvester JE. 1997. Beyond ribosomal DNA: On towards the telomere. *Chromosoma* **105:** 431–437.

Grandori C, Gomez-Roman N, Felton-Edkins ZA, Ngouenet C, Galloway DA, Eisenman RN, White RJ. 2005. c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I. *Nat Cell Biol* **7:** 311–318.

Grummt I. 2003. Life on a planet of its own: Regulation of RNA polymerase I transcription in the nucleolus. *Genes Dev* **17:** 1691–1702.

Hadjiolov AA. 1985. *The nucleolus and ribosome biogenesis*. Springer-Verlag, New York.
Hannan KM, Hannan RD, Smith SD, Jefferson LS, Lun M, Rothblum LI. 2000. Rb and p130 regulate RNA polymerase I transcription: Rb disrupts the interaction between UBF and SL-1. *Oncogene* **19:** 4988–4999.
Henderson AS, Warburton D, Atwood KC. 1972. Location of ribosomal DNA in the human chromosome complement. *Proc Natl Acad Sci* **69:** 3394–3398.
Hernandez-Hernandez A, Soto-Reyes E, Ortiz R, Arriaga-Canon C, Echeverria-Martinez OM, Vazquez-Nin GH, Recillas-Targa F. 2012. Changes of the nucleolus architecture in absence of the nuclear factor CTCF. *Cytogenet Genome Res* **136:** 89–96.
International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.
Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:** 462–467.
Kent WJ. 2002. BLAT: The BLAST-like alignment tool. *Genome Res* **12:** 656–664.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.
Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326:** 289–293.
Lyle R, Prandini P, Osoegawa K, ten Hallers B, Humphray S, Zhu B, Eyras E, Castelo R, Bird CP, Gagos S, et al. 2007. Islands of euchromatin-like sequence and expressed polymorphic sequences within the short arm of human chromosome 21. *Genome Res* **17:** 1690–1696.
Mais C, Wright JE, Prieto JL, Raggett SL, McStay B. 2005. UBF-binding site arrays form pseudo-NORs and sequester the RNA polymerase I transcription machinery. *Genes Dev* **19:** 50–64.
McClintock B. 1934. The relationship of a particular chromosomal element to the development of the nucleoli in Zea Mays. *Zeit Zellforsch Mik Anat* **21:** 294–328.
McStay B, Grummt I. 2008. The epigenetics of rRNA genes: From molecular to chromosome biology. *Annu Rev Cell Dev Biol* **24:** 131–157.
Metzdorf R, Gottert E, Blin N. 1988. A novel centromeric repetitive DNA from human chromosome 22. *Chromosoma* **97:** 154–158.
Nemeth A, Langst G. 2011. Genome organization in and around the nucleolus. *Trends Genet* **27:** 149–156.
Nemeth A, Conesa A, Santoyo-Lopez J, Medina I, Montaner D, Peterfia B, Solovei I, Cremer T, Dopazo J, Langst G. 2010. Initial genomics of the human nucleolus. *PLoS Genet* **6:** e1000889.
Olson MOJE. 2011. *The nucleolus*. Springer, Berlin.
Phillips JE, Corces VG. 2009. CTCF: Master weaver of the genome. *Cell* **137:** 1194–1211.
Pianese G. 1896. Beitrag zur histologie und aetiologie der carcinoma. Histologische und experimentelle untersuchungen. *Beitr Pathol Anat Allgem Pathol* **142:** 1–193.
Piccini I, Ballarati L, Bassi C, Rocchi M, Marozzi A, Ginelli E, Meneveri R. 2001. The structure of duplications on human acrocentric chromosome short arms derived by the analysis of 15p. *Hum Genet* **108:** 467–477.
Prieto JL, McStay B. 2007. Recruitment of factors linking transcription and processing of pre-rRNA to NOR chromatin is UBF-dependent and occurs independent of transcription in human cells. *Genes Dev* **21:** 2041–2054.
Sakai K, Ohta T, Minoshima S, Kudoh J, Wang Y, de Jong PJ, Shimizu N. 1995. Human ribosomal RNA gene cluster: Identification of the proximal end containing a novel tandem repeat sequence. *Genomics* **26:** 521–526.
Savino TM, Gebrane-Younes J, De Mey J, Sibarita JB, Hernandez-Verdun D. 2001. Nucleolar assembly of the rRNA processing machinery in living cells. *J Cell Biol* **153:** 1097–1110.
Schmickel RD. 1973. Quantitation of human ribosomal DNA: Hybridization of human DNA with ribosomal RNA for quantitation and fractionation. *Pediatr Res* **7:** 5–12.
She X, Horvath JE, Jiang Z, Liu G, Furey TS, Christ L, Clark R, Graves T, Gulden CL, Alkan C, et al. 2004. The structure and evolution of centromeric transition regions within the human genome. *Nature* **430:** 857–864.
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Res* **19:** 1117–1123.
Sirri V, Urcuqui-Inchima S, Roussel P, Hernandez-Verdun D. 2008. Nucleolus: The fascinating nuclear body. *Histochem Cell Biol* **129:** 13–31.
Smit AFA, Hubley R, Green P. 2010. RepeatMasker. http://www.repeatmasker.org.
Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* **21:** 1757–1767.
Stults DM, Killen MW, Pierce HH, Pierce AJ. 2008. Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res* **18:** 13–18.
Sullivan GJ, Bridger JM, Cuthbert AP, Newbold RF, Bickmore WA, McStay B. 2001. Human acrocentric chromosomes with transcriptionally silent nucleolar organizer regions associate with nucleoli. *EMBO J* **20:** 2867–2874.
Therman E, Susman B, Denniston C. 1989. The nonrandom participation of human acrocentric chromosomes in Robertsonian translocations. *Ann Hum Genet* **53:** 49–65.
Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.
van de Nobelen S, Rosa-Garrido M, Leers J, Heath H, Soochit W, Joosen L, Jonkers I, Demmers J, van der Reijden M, Torrano V, et al. 2010. CTCF regulates the local epigenetic state of ribosomal DNA repeats. *Epigenetics Chromatin* **3:** 19.
van Koningsbruggen S, Gierlinski M, Schofield P, Martin D, Barton GJ, Ariyurek Y, den Dunnen JT, Lamond AI. 2010. High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. *Mol Biol Cell* **21:** 3735–3748.
Visintin R, Hwang ES, Amon A. 1999. Cfi1 prevents premature exit from mitosis by anchoring Cdc14 phosphatase in the nucleolus. *Nature* **398:** 818–823.
Worton RG, Sutherland J, Sylvester JE, Willard HF, Bodrug S, Dube I, Duff C, Kean V, Ray PN, Schmickel RD. 1988. Human ribosomal RNA genes: Orientation of the tandem array and conservation of the 5′ end. *Science* **239:** 64–68.
Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18:** 821–829.
Zhang LF, Huynh KD, Lee JT. 2007. Perinucleolar targeting of the inactive X during S phase: Evidence for a role in the maintenance of silencing. *Cell* **129:** 693–706.

Received March 21, 2013; accepted in revised form August 28, 2013.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**(3): 403-410.

Anastassova-Kristeva M. 1977. The nucleolar cycle in man. *Journal of cell science* **25**: 103-110.

Arnheim N, Krystal M, Schmickel R, Wilson G, Ryder O, Zimmer E. 1980. Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. *Proceedings of the National Academy of Sciences* **77**(12): 7323-7327.

Audas TE, Jacob MD, Lee S. 2012. Immobilization of proteins in the nucleolus by ribosomal intergenic spacer noncoding RNA. *Molecular cell* **45**: 147-157.

Bachman NJ, Yang TL, Dasen JS, Ernst RE, Lomax MI. 1996. Phylogenetic footprinting of the human cytochrome c oxidase subunit VB promoter. *Arch Biochem Biophys* **333**(1): 152-162.

Bailey JA, Gu ZP, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**(5583): 1003-1007.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001a. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11**(6): 1005-1017.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001b. Segmental duplications: Organization and impact within the current Human Genome Project assembly. *Genome Res* **11**(6): 1005-1017.

Barbouti A, Stankiewicz P, Nusbaum C, Cuomo C, Cook A *et al.* 2004. The Breakpoint Region of the Most Common Isochromosome, i(17q), in Human Neoplasia Is Characterized by a Complex Genomic Architecture with Large, Palindromic, Low-Copy Repeats. *The American Journal of Human Genetics* **74**: 1-10.

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

Bartsch I, Schoneberg C, Grummt I. 1987. Evolutionary changes of sequences and factors that direct transcription termination of human and mouse ribsomal genes. *Molecular and Cellular Biology* **7**: 2521-2529.

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES. 2002. ARACHNE: a whole-genome shotgun assembler. *Genome Res* **12**(1): 177-189.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**(2): 573-580.

Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, Cuppen E. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**(1): 21-24.

Besemer J, Borodovsky M. 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* **33**: W451-W454.

Betran E, Demuth JP, Williford A. 2012. Why chromosome palindromes? *Int J Evol Biol* **2012**: 207958.

Bierhoff H, Schmitz K, Maass F, Ye J, Grummt I. 2010. Noncoding transcripts in sense and antisense orientation regulate the epigenetic state of ribosomal RNA genes. *Cold Spring Harbor symposia on quantitative biology* **75**: 357-364.

Birnstiel ML, Chipchase MIH, Hyde BB. 1963. The nucleolus, a source of ribosomes. *Biochimica et Biophysica Acta (BBA) - Specialized Section on Nucleic Acids and Related Subjects* **76**: 454-462.

Birnstiel ML, Hyde BB. 1963. Protein synthesis by isolated pea nucleoli. *The Journal of cell biology* **18**: 41-50.
Boseley P, Moss T, Machler M, Portmann R, Birnstiel M. 1979. Sequence organization of the spacer DNA in a ribosomal gene unit of Xenopus laevis. *Cell* **17**(1): 19-31.
Boyle AP, Guinney J, Crawford GE, Furey TS. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**(21): 2537-2538.
Brewer BJ, Lockshon D, Fangman WL. 1992. The arrest of replication forks in the rDNA of yeast occurs independently of transcription. *Cell* **71**(2): 267-276.
Budde A, Grummt I. 1999. p53 represses ribosomal gene transcription. *Oncogene* **18**: 1119-1124.
Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**(1): 78-94.
Bywater MJ, Poortinga G, Sanij E, Hein N, Peck A, Cullinane C, Wall M, Cluse L, Drygin D, Anderes K. 2012. Inhibition of RNA polymerase I as a therapeutic strategy to promote cancer-specific activation of p53. *Cancer cell* **22**(1): 51-65.
Caburet S, Conti C, Schurra C, Lebofsky R, Edelstein SJ, Bensimon A. 2005. Human ribosomal RNA gene arrays display a broad range of palindromic structures. *Genome Res* **15**(8): 1079-1085.
Caparros M-L, Fisher A, Merkenschlager M, Wang Z, Schones DE, Zhao K. 2009. Characterization of human epigenomes. *Current Opinion in Genetics & Development* **19**: 127-134.
Carbone L, Vessere GM, ten Hallers BFH, Zhu B, Osoegawa K *et al.* 2006. A high-resolution map of synteny disruptions in gibbon and human genomes. *PLoS Genetics* **2**: e223.
Carvalho CMB, Lupski JR. 2008. Copy number variation at the breakpoint region of isochromosome 17q. *Genome Res* **18**: 1724-1732.
Clark CG, Tague BW, Ware VC, Gerbi SA. 1984. Xenopus-Laevis-28s Ribosomal-Rna - a Secondary Structure Model and Its Evolutionary and Functional Implications. *Nucleic Acids Res* **12**(15): 6197-6220.
Coffman FD, Georgoff I, Fresa KL, Sylvester J, Gonzalez I, Cohen S. 1993. In vitro replication of plasmids containing human ribosomal gene sequences: origin localization and dependence on an aprotinin-binding cytosolic protein. *Exp Cell Res* **209**(1): 123-132.
Coffman FD, He M, Diaz M-L, Cohen S. 2006. Multiple Initiation Sites within the Human Ribosomal RNA Gene. *Cell Cycle* **5**: 1223-1233.
Conconi A, Widmer RM, Koller T, Sogo J. 1989. Two different chromatin structures coexist in ribosomal RNA genes throughout the cell cycle. *Cell* **57**: 753-761.
Connallon T, Clark AG. 2010. Gene duplication, gene conversion and the evolution of the Y chromosome. *Genetics* **186**(1): 277-286.
Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11**: 485.
Dai MS, Lu H. 2008. Crosstalk Between c-Myc and Ribosome in Ribosomal Biogenesis and Cancer. *J Cell Biochem* **105**(3): 670-677.
Dammann R, Lucchini R, Koller T, Sogo JM. 1995. Transcription in the yeast rRNA gene locus: distribution of the active gene copies and chromatin structure of their flanking regulatory sequences. *Mol Cell Biol* **15**(10): 5294-5303.
Dang CV. 1999. c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Mol Cell Biol* **19**(1): 1-11.
Deininger P. 2011. Alu elements: know the SINEs. *Genome Biology* **12**: 236.
Dellino GI, Cittaro D, Piccioni R, Luzi L, Banfi S, Segalla S, Cesaroni M, Mendoza-Maldonado R, Giacca M, Pelicci PG. 2013. Genome-wide mapping of human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing. *Genome Res* **23**: 1-11.
Dimitrova DS. 2011. DNA replication initiation patterns and spatial dynamics of the human ribosomal RNA gene loci. *Journal of cell science* **124**: 2743-2752.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T *et al.* 2012. Landscape of transcription in human cells. *Nature* **489**(7414): 101-108.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1): 15-21.

Drygin D, Rice WG, Grummt I. 2010. The RNA polymerase I transcription machinery: an emerging target for the treatment of cancer. *Annu Rev Pharmacol Toxicol* **50**: 131-156.

Dunham A Matthews LH Burton J Ashurst JL Howe KL *et al.* 2004. The DNA sequence and analysis of human chromosome 13. *Nature* **428**: 522-528.

Dunham I Kundaje A Aldred SF Collins PJ Davis C *et al.* 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR *et al.* 1999. The DNA sequence of human chromosome 22. *Nature* **402**: 489-495.

Ebersberger I, Metzler D, Schwarz C, Paabo S. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. *The American Journal of Human Genetics* **70**(6): 1490-1497.

Edstrom J-e. 1960. Composition of ribonucleic acid from various parts of spider oocytes. *J Biophys Biochem Cytol* **8**: 47-51.

Eichler EE, Clark RA, She X. 2004. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nature reviews Genetics* **5**: 345-354.

Elder JF, Jr., Turner BJ. 1995. Concerted evolution of repetitive DNA sequences in eukaryotes. *The Quarterly review of biology* **70**(3): 297-320.

Elion EA, Warner JR. 1984. The major promoter element of rRNA transcription in yeast lies 2 kb upstream. *Cell* **39**(3 Pt 2): 663-673.

Faghihi MA, Wahlestedt C. 2009. Regulatory roles of natural antisense transcripts. *Nature reviews Molecular cell biology* **10**: 637-643.

Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ *et al.* 2009. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**(7232): 1028-1032.

Floutsakou I, Agrawal S, Nguyen TT, Seoighe C, Ganley ARD, McStay B. 2013. The shared genomic architecture of human nucleolar organizer regions. *Genome Res*.

Forrest ARR Kawaji H Rehli M Baillie JK de Hoon MJL *et al.* 2014. A promoter-level mammalian expression atlas. *Nature* **507**(7493): 462-+.

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32**: W273-W279.

Gabellini D, D'Antona G, Moggio M, Prelle A, Zecca C *et al.* 2006. Facioscapulohumeral muscular dystrophy in mice overexpressing FRG1. *Nature* **439**(7079): 973-977.

Gagnon-Kugler T, Langlois F, Stefanovsky V, Lessard F, Moss T. 2009. Loss of human ribosomal gene CpG methylation enhances cryptic RNA polymerase II transcription and disrupts ribosomal RNA processing. *Mol Cell* **35**(4): 414-425.

Ganley ARD, Hayashi K, Horiuchi T, Kobayashi T. 2005. Identifying gene-independent noncoding functional elements in the yeast ribosomal DNA by phylogenetic footprinting. *Proceedings of the National Academy of Sciences* **102**: 11787-11792.

Gencheva M, Anachkova, B., Russev G. 1996. Mapping the Sites of Initiation of DNA Replication in Rat and Human rRNA Genes. *The Journal of biological chemistry* **271**: 2608-2614.

Georgiev OI, Nikolaev N, Hadjiolov AA, Skryabin KG, Zakharyev VM, Bayev AA. 1981. The structure of the yeast ribosomal RNA genes. 4. Complete sequence of the 25 S rRNA gene from Saccharomyces cerevisae. *Nucleic Acids Res* **9**(24): 6953-6958.

Gibbs RA Rogers J Katze MG Bumgarner R Weinstock GM *et al.* 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**: 222-234.

Gimelli G, Zuffardi O, Giglio S, Zeng C, He D. 2000. CENP-G in neocentromeres and inactive centromeres. *Chromosoma* **109**: 328-333.

Gögel E, Längst G, Grummt I, Kunkel E, Grummt F. 1996. Mapping of replication initiation sites in the mouse ribosomal gene cluster. *Chromosoma* **104**(7): 511-518.

Goidts V, Szamalek JM, Hameister H, Kehrer-Sawatzki H. 2004. Segmental duplication associated with the human-specific inversion of chromosome 18: a further example of the impact of segmental duplications on karyotype and genome evolution in primates. *Human Genetics* **115**(2): 116-122.

Gonzalez IL, Gorski JL, Campen TJ, Dorney DJ, Erickson JM, Sylvester JE, Schmickel RD. 1985. Variation among human 28S ribosomal RNA genes. *Proceedings of the National Academy of Sciences* **82**: 7666-7670.

Gonzalez IL, Petersen R, Sylvester JE. 1989. Independent insertion of Alu elements in the human ribosomal spacer and their concerted evolution. *Mol Biol Evol* **6**(4): 413-423.

Gonzalez IL, Sylvester JE. 1995. Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. *Genomics* **27**: 320-328.

Gonzalez IL, Sylvester JE. 1997. Beyond ribosomal DNA: on towards the telomere. *Chromosoma* **105**: 431-437.

Gonzalez IL, Sylvester JE. 2001. Human rDNA: evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. *Genomics* **73**(3): 255-263.

Gonzalez IL, Tugendreich S, Hieter P, Sylvester JE. 1993. Fixation times of retroposons in the ribosomal DNA spacer of human and other primates. *Genomics* **18**(1): 29-36.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Res* **8**(3): 195-202.

Gorski JJ, Pathak S, Panov K, Kasciukovic T, Panova T, Russell J, Zomerdijk JC. 2007. A novel TBP-associated factor of SL1 functions in RNA polymerase I transcription. *The EMBO journal* **26**(6): 1560-1568.

Gorski JL, Gonzalez IL, Schmickel RD. 1987. The secondary structure of human 28S rRNA: The structure and evolution of a mosaic rRNA gene. *Journal of Molecular Evolution* **24**: 236-251.

Gottgens B, Gilbert JG, Barton LM, Grafham D, Rogers J, Bentley DR, Green AR. 2001. Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res* **11**(1): 87-97.

Grandori C, Cowley SM, James LP, Eisenman RN. 2000. The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu Rev Cell Dev Biol* **16**: 653-699.

Grandori C, Gomez-Roman N, Felton-Edkins ZA, Ngouenet C, Galloway DA, Eisenman RN, White RJ. 2005. c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I. *Nature cell biology* **7**: 311-318.

Grozdanov P, Georgiev O, Karagyozov L. 2003. Complete sequence of the 45-kb mouse ribosomal DNA repeat: analysis of the intergenic spacer☆. *Genomics* **82**: 637-643.

Grummt I. 1999. Regulation of mammalian ribosomal gene transcription by RNA polymerase I. *Prog Nucleic Acid Res Mol Biol* **62**: 109-154.

Grummt I. 2007. Different epigenetic layers engage in complex crosstalk to define the epigenetic state of mammalian rRNA genes. *Human molecular genetics* **16 Spec No**: R21-27.

Grummt I, Kuhn A, Bartsch I, Rosenbauer H. 1986. A transcription terminator located upstream of the mouse rDNA initiation site affects rRNA synthesis. *Cell* **47**(6): 901-911.

Grzmil M, Hemmings BA. 2012. Translation regulation as a therapeutic target in cancer. *Cancer Research* **72**(16): 3891-3900.

Guilbaud G, Rappailles A, Baker A, Chen CL, Arneodo A, Goldar A, d'Aubenton-Carafa Y, Thermes C, Audit B, Hyrien O. 2011. Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. *PLoS Comput Biol* **7**(12): e1002322.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology* **9**(1).

Hadjiolov A. 1985. Ribosome Biogenesis in the Life Cycle of Normal and Cancer Cells. In *The Nucleolus and Ribosome Biogenesis*, Vol 12, pp. 165-196. Springer Vienna.

Haltiner MM, Smale ST, Tjian R. 1986. Two distinct promoter elements in the human rRNA gene identified by linker scanning mutagenesis. *Molecular and cellular biology* **6**: 227-235.

Hamperl S, Wittner M, Babl V, Perez-Fernandez J, Tschochner H, Griesenbeck J. 2013. Chromatin states at ribosomal DNA loci. *Biochimica et biophysica acta* **1829**: 405-417.

Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler SM, Stamatoyannopoulos JA. 2010. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences* **107**(1): 139-144.

Harrison PM, Zheng DY, Zhang ZL, Carriero N, Gerstein M. 2005. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res* **33**(8): 2374-2383.

Hassouna N, Michot B, Bachellerie JP. 1984. The complete nucleotide sequence of mouse 28S rRNA gene. Implications for the process of size increase of the large subunit rRNA in higher eukaryotes. *Nucleic Acids Res* **12**: 3563-3583.

Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T *et al.* 2000. The DNA sequence of human chromosome 21. *Nature* **405**: 311-319.

He D, Zeng C, Woods K, Zhong L, Turner D, Busch RK, Brinkley BR, Busch H. 1998. CENP-G: a new centromeric protein that is associated with the α-1 satellite DNA subfamily. *Chromosoma* **107**(3): 189-197.

Heilig R, Eckenberg R, Petit J-L, Fonknechten N, Da Silva C *et al.* 2003. The DNA sequence and analysis of human chromosome 14. *Nature* **421**: 601-607.

Henderson A, Warburton D, Atwood KC. 1974a. Localization of rDNA in the chimpanzee (Pan troglodytes) chromosome complement. *Chromosoma* **46**(4): 435-441.

Henderson AS, Atwood KC, Warburton D. 1976. Chromosomal distribution of rDNA in Pan paniscus, Gorilla gorilla beringei, and Symphalangus syndactylus: comparison to related primates. *Chromosoma* **59**(2): 147-155.

Henderson AS, Warburton D, Atwood KC. 1972. Location of ribosomal DNA in the human chromosome complement. *Proceedings of the National Academy of Sciences* **69**(11): 3394-3398.

Henderson AS, Warburton D, Atwood KC. 1974b. Localization of rDNA in the chromosome complement of the rhesus (Macaca mulatta). *Chromosoma* **44**(4): 367-370.

Henderson AS, Warburton D, Megraw-Ripley S, Atwood KC. 1977. The chromosomal location of rDNA in selected lower primates. *Cytogenet Cell Genet* **19**(5): 281-302.

Henderson AS, Warburton D, Megraw-Ripley S, Atwood KC. 1979. The chromosomal location of rDNA in the Sumatran orangutan, Pongo pygmaeus albei. *Cytogenet Cell Genet* **23**(3): 213-216.

Hernández-Hernández A, Soto-Reyes E, Ortiz R, Arriaga-Canon C, Echeverría-Martinez OM, Vázquez-Nin GH, Recillas-Targa F. 2012. Changes of the nucleolus architecture in absence of the nuclear factor CTCF. *Cytogenetic and Genome Research* **136**: 89-96.

Hillis DM, Dixon MT. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. *The Quarterly review of biology* **66**: 411-453.

Høgset A, Øyen TB. 1984. Correlation between suppressed meiotic recombination and the lack of DNA strand-breaks in the rRNA genes of Saccharomyces cerevisiae. *Nucleic Acids Res* **12**(18): 7199-7213.

Holland EC. 2004. Regulation of translation and cancer. *Cell Cycle* **3**(4): 452-455.

Houseley J, Kotovic K, El Hage A, Tollervey D. 2007. Trf4 targets ncRNAs from telomeric and rDNA spacer regions and functions in rDNA copy number control. *The EMBO journal* **26**: 4996-5006.

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM *et al.* 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**(7280): 536-539.

Hughes JF, Skaletsky H, Pyntikova T, Minx PJ, Graves T, Rozen S, Wilson RK, Page DC. 2005. Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* **437**(7055): 101-104.

Huh GS, Hynes RO. 1994. Regulation of alternative pre-mRNA splicing by a novel repeated hexanucleotide element. *Genes & Development* **8**(13): 1561-1574.

Hwang CJ, Fields JR, Shiao Y-H. 2011. Non-coding rRNA-mediated preferential killing in cancer cells is enhanced by suppression of autophagy in non-transformed counterpart. *Cell death & disease* **2**: e239.

Hyrien O, Mechali M. 1993. Chromosomal Replication Initiates and Terminates at Random Sequences but at Regular Intervals in the Ribosomal DNA of Xenopus Early Embryos. *The EMBO journal* **12**(12): 4511-4520.

Ide S, Miyazaki T, Maki H, Kobayashi T. 2010. Abundance of Ribosomal RNA Gene Copies Maintains Genome Integrity. *Science* **327**(5966): 693-696.

Iglesias AR, Kindlund E, Tammi M, Wadelius C. 2004. Some microsatellites may act as novel polymorphic< i> cis</i>-regulatory elements through transcription factor binding. *Gene* **341**: 149-165.

Jacob MD, Audas TE, Mullineux S-T, Lee S. 2012. Where no RNA polymerase has gone before: novel functional transcripts derived from the ribosomal intergenic spacer. *Nucleus* **3**: 315-319.

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES. 2003. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* **13**: 91-96.

Johnsson P, Ackley A, Vidarsdottir L, Lui W-O, Corcoran M, Grandér D, Morris KV. 2013. A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. *Nature structural & molecular biology* **20**: 440-446.

Kalmárová M, Smirnov E, Mašata M, Koberna K, Ligasová A, Popov A, Raška I. 2007. Positioning of NORs and NOR-bearing chromosomes in relation to nucleoli. *Journal of Structural Biology* **160**: 49-56.

Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu Y-M, Cao X, Asangani IA, Kothari V, Prensner JR, Lonigro RJ. 2012. Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* **149**(7): 1622-1634.

Kaplan FS, Murray J, Sylvester JE, Gonzalez IL, O'Connor JP, Doering JL, Muenke M, Emanuel BS, Zasloff MA. 1993. The topographic organization of repetitive DNA in the human nucleolus. *Genomics* **15**(1): 123-132.

Kassavetis GA, Braun BR, Nguyen LH, Peter Geiduschek E. 1990. S. cerevisiae TFIIIB is the transcription initiation factor proper of RNA polymerase III, while TFIIIA and TFIIIC are assembly factors. *Cell* **60**: 235-245.

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M *et al.* 2005a. Antisense transcription in the mammalian transcriptome. *Science* **309**: 1564-1566.

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M *et al.* 2005b. Antisense transcription in the mammalian transcriptome. *Science* **309**(5740): 1564-1566.

Katinakis PK, Slater A, Burdon RH. 1980. Non-polyadenylated mRNAs from eukaryotes. *Febs Lett* **116**(1): 1-7.

Katoh K, Asimenos G, Toh H. 2009. Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol* **537**: 39-64.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**(2): 511-518.

Kehrer-Sawatzki H, Sandig C, Chuzhanova N, Goidts V, Szamalek JM, Tanzer S, Muller S, Platzer M, Cooper DN, Hameister H. 2005. Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (Pan troglodytes). *Human Mutation* **25**(1): 45-55.

Kern SE, Kinzler KW, Bruskin A, Jarosz D, Friedman P, Prives C, Vogelstein B. 1991. Identification of p53 as a sequence-specific DNA-binding protein. *Science* **252**(5013): 1708-1711.

Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26**(12): 1351-1359.

Koch CM, Andrews RM, Flicek P, Dillon SC, Karaöz U *et al.* 2007. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res* **17**: 691-707.

Krystal M, Arnheim N. 1978. Length heterogeneity in a region of the human ribosomal gene spacer is not accompanied by extensive population polymorphism. *J Mol Biol* **126**(1): 91-104.

Krystal M, D'Eustachio P, Ruddle FH, Arnheim N. 1981. Human nucleolus organizers on nonhomologous chromosomes can share the same ribosomal gene variants. *Proceedings of the National Academy of Sciences* **78**: 5744-5748.

Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, Makeev VJ. 2013. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res* **41**(Database issue): D195-202.

Lamaye F, Galliot S, Alibardi L, Lafontaine DL, Thiry M. 2011. Nucleolar structure across evolution: the transition between bi-and tricompartmentalized nucleoli lies within the class Reptilia. *Journal of Structural Biology* **174**(2): 352-359.

Lander ES Linton LM Birren B Nusbaum C Zody MC *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**(4): 357-359.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**(3): R25.

Learned RM. 1983. Regulation of Human Ribosomal RNA Transcription. *Proceedings of the National Academy of Sciences* **80**: 3558-3562.

Learned RM, Learned TK, Haltiner MM, Tjian RT. 1986. Human rRNA transcription is modulated by the coordinate binding of two factors to an upstream control element. *Cell* **45**: 847-857.

Lebofsky R, Bensimon A. 2005. DNA replication origin plasticity and perturbed fork progression in human inverted repeats. *Molecular and cellular biology* **25**: 6789-6797.

Leffers H, Andersen AH. 1993. The sequence of 28S ribosomal RNA varies within and between human cell lines. *Nucleic Acids Res* **21**: 1449-1455.

Léger I, Guillaud M, Krief B, Brugal G. 1994. Interactive computer-assisted analysis of chromosome 1 colocalization with nucleoli. *Cytometry* **16**: 313-323.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL *et al.* 2007. The diploid genome sequence of an individual human. *PLoS biology* **5**(10): e254.

Lewis SE, Searle SM, Harris N, Gibson M, Lyer V *et al.* 2002. Apollo: a sequence annotation editor. *Genome Biology* **3**(12): RESEARCH0082.

Li WH, Gojobori T, Nei M. 1981. Pseudogenes as a Paradigm of Neutral Evolution. *Nature* **292**(5820): 237-239.

Li Y-C, Korol AB, Fahima T, Nevo E. 2004. Microsatellites within genes: structure, function, and evolution. *Molecular biology and evolution* **21**: 991-1007.

Linardopoulou EV, Williams EM, Fan Y, Friedman C, Young JM, Trask BJ. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437**(7055): 94-100.

Linskens MH, Huberman JA. 1988. Organization of replication of ribosomal DNA in Saccharomyces cerevisiae. *Mol Cell Biol* **8**(11): 4927-4935.

Little RD, Platt TH, Schildkraut CL. 1993. Initiation and termination of DNA replication in human rRNA genes. *Mol Cell Biol* **13**: 6600-6613.

Locke DP Hillier LW Warren WC Worley KC Nazareth LV *et al.* 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**: 529-533.
López-Estraño C, Schvartzman JB, Krimer DB, Hernández P. 1998. Co-localization of polar replication fork barriers and rRNA transcription terminators in mouse rDNA 11Edited by M. Gottesman. *J Mol Biol* **277**: 249-256.
Lucchini R, Sogo JM. 1992. Different chromatin structures along the spacers flanking active and inactive Xenopus rRNA genes. *Mol Cell Biol* **12**(10): 4288-4296.
Lutz W, Leon J, Eilers M. 2002. Contributions of Myc to tumorigenesis. *Biochim Biophys Acta* **1602**(1): 61-71.
Lyle R, Béna F, Gagos S, Gehrig C, Lopez G, Schinzel A, Lespinasse J, Bottani A, Dahoun S, Taine L. 2008. Genotype–phenotype correlations in Down syndrome identified by array CGH in 30 cases of partial trisomy and partial monosomy chromosome 21. *European Journal of Human Genetics* **17**(4): 454-466.
Lyle R, Prandini P, Osoegawa K, ten Hallers B, Humphray S *et al.* 2007. Islands of euchromatin-like sequence and expressed polymorphic sequences within the short arm of human chromosome 21. *Genome Res* **17**: 1690-1696.
Lyle R, Wright TJ, Clark LN, Hewitt JE. 1995. Fshd-Associated Repeat, D4z4, Is a Member of a Dispersed Family of Homeobox-Containing Repeats, Subsets of Which Are Clustered on the Short Arms of the Acrocentric Chromosomes. *Genomics* **28**(3): 389-397.
Maggi LB, Weber JD. 2005. Nucleolar adaptation in human cancer. *Cancer investigation* **23**: 599-608.
Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**(16): 2878-2879.
Manuelidis L. 1978. Chromosomal localization of complex and simple repeated human DNAs. *Chromosoma* **66**(1): 23-32.
Manuelidis L, Borden J. 1988. Reproducible compartmentalization of individual chromosome domains in human CNS cells revealed by in situ hybridization and three-dimensional reconstruction. *Chromosoma* **96**: 397-410.
Marais GAB, Campos PRA, Gordo I. 2010. Can Intra-Y Gene Conversion Oppose the Degeneration of the Human Y Chromosome? A Simulation Study. *Genome biology and evolution* **2**: 347-357.
Marques-Bonet T, Ryder OA, Eichler EE. 2009. Sequencing primate genomes: what have we learned? *Annual review of genomics and human genetics* **10**: 355-386.
Mayan MD. 2013. RNAP-II molecules participate in the anchoring of the ORC to rDNA replication origins. *PLoS one* **8**: e53405.
Mayán MD. 2013. RNAP-II transcribes two small RNAs at the promoter and terminator regions of the RNAP-I gene in Saccharomyces cerevisiae. *Yeast*: 25-32.
Mayer C, Bierhoff H, Grummt I. 2005. The nucleolus as a stress sensor: JNK2 inactivates the transcription factor TIF-IA and down-regulates rRNA synthesis. *Genes & Development* **19**(8): 933-941.
Mayer C, Neubert M, Grummt I. 2008. The structure of NoRC-associated RNA is crucial for targeting the chromatin remodelling complex NoRC to the nucleolus. *EMBO reports* **9**: 774-780.
Mayer C, Schmitz K-M, Li J, Grummt I, Santoro R. 2006. Intergenic transcripts regulate the epigenetic state of rRNA genes. *Molecular cell* **22**: 351-361.
Mayer MP, Bukau B. 2005. Hsp70 chaperones: cellular functions and molecular mechanism. *Cell Mol Life Sci* **62**(6): 670-684.
McClintock B. 1934. The relation of a particular chromosomal element to the development of the nucleoli in Zea mays. *Zeitschrift für Zellforschung und Mikroskopische Anatomie* **21**: 294-326.
McConkey E, Hopkins J. 1964. The relationship of the nucleolus to the synthesis of ribosomal RNA in HeLa cells. *Proceedings of the National Academy of Sciences*.

McCue LA, Thompson W, Carmack CS, Lawrence CE. 2002. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res* **12**(10): 1523-1532.

McStay B, Grummt I. 2008. The epigenetics of rRNA genes: from molecular to chromosome biology. *Annu Rev Cell Dev Biol* **24**: 131-157.

Mekhail K, Gunaratnam L, Bonicalzi ME, Lee S. 2004. HIF activation by pH-dependent nucleolar sequestration of VHL. *Nat Cell Biol* **6**(7): 642-647.

Mekhail K, Rivero-Lopez L, Khacho M, Lee S. 2006. Restriction of rRNA synthesis by VHL maintains energy equilibrium under hypoxia. *Cell Cycle* **5**(20): 2401-2413.

Metzdorf R, Gottert E, Blin N. 1988. A novel centromeric repetitive DNA from human chromosome 22. *Chromosoma* **97**(2): 154-158.

Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang S-P, Enard W, Hellmann I, Lindblad-Toh K, Altheide TK. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*: 69-87.

Miller OJ, Tantravahi R, Miller DA, Yu LC, Szabo P, Prensky W. 1979. Marked Increase in Ribosomal-Rna Gene Multiplicity in a Rat Hepatoma Cell-Line. *Chromosoma* **71**(2): 183-195.

Montanaro L, Treré D, Derenzini M. 2008. Nucleolus, ribosomes, and cancer. *The American journal of pathology* **173**: 301-310.

Morris EE, Amria MY, Kistner-Griffin E, Svenson JL, Kamen DL, Gilkeson GS, Nowling TK. 2010. A GA microsatellite in the Fli1 promoter modulates gene expression and is associated with systemic lupus erythematosus patients without nephritis. *Arthritis research & therapy* **12**: R212.

Mosgoeller W. 2004. Nucleolar ultrastructure in vertebrates. *The nucleolus*: 10-20.

Moss T, Langlois F, Gagnon-Kugler T, Stefanovsky V. 2007. A housekeeper with power of attorney: the rRNA genes in ribosome biogenesis. *Cellular and molecular life sciences* **64**(1): 29-49.

Muleris M, Salmon R, Zafrani B, Girodet J, Dutrillaux B. 1984. Consistent deficiencies of chromosome 18 and of the short arm of chromosome 17 in eleven cases of human large bowel cancer: a possible recessive determinism. In *Annales de genetique*, Vol 28, pp. 206-213.

Muller M, Lucchini R, Sogo JM. 2000. Replication of yeast rDNA initiates downstream of transcriptionally active genes. *Mol Cell* **5**(5): 767-777.

Murao S-i, Horita Y, Maeda S, Takahashi R, Kano Y, Sugiyama T. 1982. Amplification and abnormal chromosomal distribution of ribosomal genes (rDNA) in rat erythroleukemia cells. *Cancer genetics and cytogenetics* **6**(4): 303-312.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP *et al.* 2000. A whole-genome assembly of Drosophila. *Science* **287**(5461): 2196-2204.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**(7485): 635-+.

Németh A, Perez-Fernandez J, Merkl P, Hamperl S, Gerber J, Griesenbeck J, Tschochner H. 2012. RNA polymerase I termination: Where is the end? *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*.

Netchvolodov KK, Boiko AV, Ryskov AP, Kupriyanova NS. 2006. Evolutionary divergence of the pre-promotor region of ribosomal DNA in the great apes: Full Length Research Paper. *Mitochondrial DNA* **17**(5): 378-391.

Noe L, Kucherov G. 2005. YASS: enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res* **33**(Web Server issue): W540-543.

Oakes M, Aris JP, Brockenbrough JS, Wai H, Vu L, Nomura M. 1998. Mutational analysis of the structure and localization of the nucleolus in the yeast Saccharomyces cerevisiae. *J Cell Biol* **143**(1): 23-34.

Oakes ML, Johzuka K, Vu L, Eliason K, Nomura M. 2006. Expression of rRNA genes and nucleolus formation at ectopic chromosomal sites in the yeast Saccharomyces cerevisiae. *Molecular and Cellular Biology* **26**(16): 6223-6238.

Oren M. 2003. Decision making by p53: life, death and cancer. *Cell death and differentiation* **10**(4): 431-442.
Ota T Suzuki Y Nishikawa T Otsuki T Sugiyama T *et al.* 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet* **36**(1): 40-45.
Pain D, Chirn G-W, Strassel C, Kemp DM. 2005. Multiple retropseudogenes from pluripotent cell-specific gene expression indicates a potential signature for novel gene identification. *The Journal of biological chemistry* **280**: 6265-6268.
Palumbo SL, Memmott RM, Uribe DJ, Krotova-Khan Y, Hurley LH, Ebbinghaus SW. 2008. A novel G-quadruplex-forming GGA repeat region in the c-myb promoter is a critical regulator of promoter activity. *Nucleic Acids Res* **36**(6): 1755-1769.
Pape LK, Windle JJ, Mougey E, Sollner-Webb B. 1989. The Xenopus ribosomal DNA 60- and 81-base-pair repeats are position-dependent enhancers that function at the establishment of the preinitiation complex: analysis in vivo and in an enhancer-responsive in vitro system. *Molecular and Cellular Biology* **9**(11): 5093-5104.
Paule MR, White RJ. 2000. Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res* **28**(6): 1283-1298.
Pearson CE, Zorbas H, Price GB, Zannis-Hadjopoulos M. 1996. Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. *J Cell Biochem* **63**(1): 1-22.
Pelham HRB. 1984. Hsp70 Accelerates the Recovery of Nucleolar Morphology after Heat-Shock. *The EMBO journal* **3**(13): 3095-3100.
Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE *et al.* 2011. A molecular phylogeny of living primates. *PLoS Genetics* **7**: e1001342.
Petes TD, Botstein D. 1977. Simple Mendelian inheritance of the reiterated ribosomal DNA of yeast. *Proceedings of the National Academy of Sciences* **74**(11): 5091-5095.
Pfleiderer C, Smid A, Bartsch I, Grummt I. 1990. An undecamer DNA sequence directs termination of human ribosomal gene transcription. *Nucleic Acids Res* **18**: 4727-4736.
Phillips RL, Kleese RA, Wang SS. 1971. The nucleolus organizer region of maize (Zea mays L.): Chromosomal site of DNA complementary to ribosomal RNA. *Chromosoma* **36**: 79-88.
Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Carter DRF. 2011. Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA* **17**: 792-798.
Pipes L, Li S, Bozinoski M, Palermo R, Peng XX *et al.* 2013. The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic Acids Res* **41**(D1): D906-D914.
Poliseno L. 2012. Pseudogenes: newly discovered players in human cancer. *Science signaling* **5**: re5.
Poliseno L, Haimovic A, Christos PJ, Vega Y Saenz de Miera EC, Shapiro R, Pavlick A, Berman RS, Darvishian F, Osman I. 2011. Deletion of PTENP1 pseudogene in human melanoma. *The Journal of investigative dermatology* **131**: 2497-2500.
Proux-Wera E, Byrne KP, Wolfe KH. 2013. Evolutionary Mobility of the Ribosomal DNA Array in Yeasts. *Genome biology and evolution* **5**(3): 525-531.
Pusch CM, Wundrack I, Müllenbach R, Schempp W, Blin N. 2002. The 48 bp centromeric repeat is a functionally conserved motif in great apes and man showing protein-binding properties. *Electrophoresis* **23**: 20-26.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics* **16**(6): 276-277.
Rieder CL, Salmon E. 1998. The vertebrate cell kinetochore and its roles during mitosis. *Trends in Cell Biology* **8**(8): 310-318.
RITOSSA FM, SPIEGELMAN S. 1965. Localization of DNA complementary to ribosomal RNA in the nucleolus organizer region of Drosophila melanogaster. *Proceedings of the National Academy of Sciences* **53**: 737-745.
Roberts A, Pimentel H, Trapnell C, Pachter L. 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27**: 2325-2329.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**(1): 24-26.

Rohlfs EM, Puget N, Graham ML, Weber BL, Garber JE, Skrzynia C, Halperin JL, Lenoir GM, Silverman LM, Mazoyer S. 2000. An Alu-mediated 7.1 kb deletion of BRCA1 exons 8 and 9 in breast and ovarian cancer families that results in alternative splicing of exon 10. *Genes, Chromosomes and Cancer* **28**(3): 300-307.

Roussel P. 1996. The rDNA transcription machinery is assembled during mitosis in active NORs and absent in inactive NORs. *The Journal of Cell Biology* **133**: 235-246.

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423**(6942): 873-876.

Sáfrány G, Kominami R, Muramatsu M, Hidvégi EJ. 1989. Transcription of human ribosomal DNA may terminate at multiple sites. *Gene* **79**: 299-307.

Saka K, Ide S, Ganley AR, Kobayashi T. 2013. Cellular Senescence in Yeast Is Regulated by rDNA Noncoding Transcription. *Current Biology* **23**(18): 1794-1798.

Sakai K, Ohta T, Minoshima S, Kudoh J, Wang Y, de Jong PJ, Shimizu N. 1995. Human ribosomal RNA gene cluster: identification of the proximal end containing a novel tandem repeat sequence. *Genomics* **26**: 521-526.

Salamov AA, Solovyev VV. 2000. Ab initio gene finding in Drosophila genomic DNA. *Genome Res* **10**(4): 516-522.

Sambrook J, Russell DW. 2001. Molecular cloning: a laboratory manual (3-volume set).

Sanchez JA, Kim S-M, Huberman JA. 1998. Ribosomal DNA Replication in the Fission Yeast,Schizosaccharomyces pombe. *Experimental Cell Research* **238**(1): 220-230.

Santoro R, Grummt I. 2001. Molecular Mechanisms Mediating Methylation-Dependent Silencing of Ribosomal Gene Transcription. *Molecular Cell* **8**: 719-725.

Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N. 2013. Microsatellite Tandem Repeats Are Abundant in Human Promoters and Are Associated with Regulatory Elements. *PLoS one* **8**(2).

Sawaya SM, Bagshaw AT, Buschiazzo E, Gemmell NJ. 2012. Promoter Microsatellites as Modulators of Human Gene Expression. *Adv Exp Med Biol* **769**: 41-54.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I *et al.* 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**: 169-175.

Schaub M, Myslinski E, Schuster C, Krol A, Carbon P. 1997. Staf, a promiscuous activator for enhanced transcription by RNA polymerases II and III. *The EMBO journal* **16**(1): 173-181.

Schmickel RD. 1973. Quantitation of human ribosomal DNA: hybridization of human DNA with ribosomal RNA for quantitation and fractionation. *Pediatr Res* **7**(1): 5-12.

Schuster-Bockler B, Schultz J, Rahmann S. 2004. HMM Logos for visualization of protein families. *BMC Bioinformatics* **5**.

Schuster C, Krol A, Carbon P. 1998. Two distinct domains in Staf to selectively activate small nuclear RNA-type and mRNA promoters. *Mol Cell Biol* **18**(5): 2650-2658.

Scott SA, Cohen N, Brandt T, Warburton PE, Edelmann L. 2010. Large inverted repeats within Xp11.2 are present at the breakpoints of isodicentric X chromosomes in Turner syndrome. *Human molecular genetics* **19**: 3383-3393.

Shapiro JA, von Sternberg R. 2005. Why repetitive DNA is essential to genome function. *Biol Rev* **80**(2): 227-250.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T *et al.* 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences* **100**(26): 15776-15781.

Shklover J, Weisman-Shomer P, Yafe A, Fry M. 2010. Quadruplex structures of muscle gene promoter sequences enhance in vivo MyoD-dependent gene expression. *Nucleic Acids Res* **38**(7): 2369-2377.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**(6): 1117-1123.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L *et al.* 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**(6942): 825-U822.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**.

Sorek R, Ast G, Graur D. 2002. Alu-containing exons are alternatively spliced. *Genome Res* **12**(7): 1060-1067.

Stahl A, Hartung M, Vagner-Capodano AM, Fouet C. 1976. Chromosomal constitution of nucleolus-associated chromatin in man. *Human Genetics* **35**(1): 27-34.

Stone EA, Cooper GM, Sidow A. 2005. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu Rev Genomics Hum Genet* **6**: 143-164.

Stults DM, Killen MW, Pierce HH, Pierce AJ. 2008. Genomic architecture and inheritance of human ribosomal RNA gene clusters. *Genome Res* **18**: 13-18.

Sylvester JE, Petersen R, Schmickel RD. 1989. Human ribosomal DNA: novel sequence organization in a 4.5-kb region upstream from the promoter. *Gene* **84**(1): 193-196.

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT. 1988. Embryonic ε and γ globin genes of a prosimian primate (< i> Galago crassicaudatus</i>): Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* **203**(2): 439-455.

Tanaka Y, Okamoto K, Teye K, Umata T, Yamagiwa N, Suto Y, Zhang Y, Tsuneoka M. 2010. JmjC enzyme KDM2A is a regulator of rRNA transcription in response to starvation. *The EMBO journal* **29**: 1510-1522.

Tantravahi R, Miller DA, Dev VG, Miller OJ. 1976. Detection of nucleolus organizer regions in chromosomes of human, chimpanzee, gorilla, orangutan and gibbon. *Chromosoma* **56**(1): 15-27.

Tautz D, Hancock JM, Webb DA, Tautz C, Dover GA. 1988. Complete sequences of the rRNA genes of Drosophila melanogaster. *Mol Biol Evol* **5**(4): 366-376.

Thiry M, Lafontaine DL. 2005. Birth of a nucleolus: the evolution of nucleolar compartments. *Trends Cell Biol* **15**(4): 194-199.

Thornton BR, Ng TM, Matyskiela ME, Carroll CW, Morgan DO, Toczyski DP. 2006. An architectural map of the anaphase-promoting complex. *Genes & Development* **20**: 449-460.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinformatics* **14**(2): 178-192.

Torrano V, Navascues J, Docquier F, Zhang R, Burke LJ *et al.* 2006. Targeting of CTCF to the nucleolus inhibits nucleolar transcription through a poly(ADP-ribosyl)ation-dependent mechanism. *Journal of cell science* **119**(9): 1746-1759.

Tschentscher F, Prescher G, Horsman DE, White VA, Rieder H *et al.* 2001. Partial deletions of the long and short arm of chromosome 3 point to two tumor suppressor genes in uveal melanoma. *Cancer Research* **61**(8): 3439-3442.

Tugendreich S, Boguski MS, Seldin MS, Hieter P. 1993. Linking yeast genetics to mammalian genomes: identification and mapping of the human homolog of CDC27 via the expressed sequence tag (EST) data base. *Proceedings of the National Academy of Sciences* **90**(21): 10031-10035.

Tugendreich S, Tomkiel J, Earnshaw W, Hieter P. 1995. CDC27Hs colocalizes with CDC16Hs to the centrosome and mitotic spindle and is essential for the metaphase to anaphase transition. *Cell* **81**: 261-268.

van de Nobelen S, Rosa-Garrido M, Leers J, Heath H, Soochit W *et al.* 2010. CTCF regulates the local epigenetic state of ribosomal DNA repeats. *Epigenetics & chromatin* **3**: 19.

Venter JC Adams MD Myers EW Li PW Mural RJ *et al.* 2001. The sequence of the human genome. *Science* **291**: 1304-1351.

Versteeg R, van Schaik BDC, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AHC. 2003. The human transcriptome map reveals

extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* **13**: 1998-2004.

Vowles EJ, Amos W. 2004. Evidence for widespread convergent evolution around human microsatellites. *PLoS biology* **2**(8): 1157-1167.

Wai H, Johzuka K, Vu L, Eliason K, Kobayashi T, Horiuchi T, Nomura M. 2001. Yeast RNA polymerase I enhancer is dispensable for transcription of the chromosomal rRNA gene and cell growth, and its apparent transcription enhancement from ectopic promoters requires Fob1 protein. *Molecular and Cellular Biology* **21**(16): 5541-5553.

Wang Q, Moyret-Lalle C, Couzon F, Surbiguet-Clippe C, Saurin J-C, Lorca T, Navarro C, Puisieux A. 2003. Alterations of anaphase-promoting complex genes in human colon cancer cells. *Oncogene* **22**: 1486-1490.

Warburton D, Henderson AS, Atwood KC. 1975. Localization of rDNA and Giemsa-banded chromosome complement of white-handed gibbon, Hylobates lar. *Chromosoma* **51**(1): 35-40.

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

Watt JL, Olson IA, Johnston AW, Ross HS, Couzin DA, Stephen GS. 1985. A familial pericentric inversion of chromosome 22 with a recombinant subject illustrating a 'pure' partial monosomy syndrome. *Journal of Medical Genetics* **22**: 283-287.

Welch WJ, Feramisco JR. 1984. Nuclear and Nucleolar Localization of the 72,000-Dalton Heat-Shock Protein in Heat-Shocked Mammalian-Cells. *The Journal of biological chemistry* **259**(7): 4501-4513.

Welch WJ, Suhan JP. 1986. Cellular and Biochemical Events in Mammalian-Cells during and after Recovery from Physiological Stress. *Journal of cell science* **103**(5): 2035-2052.

Werner A. 2013. Biological functions of natural antisense transcripts. *BMC Biology* **11**: 31.

Wevrick R, Willard HF. 1989. Long-Range Organization of Tandem Arrays of Alpha-Satellite DNA at the Centromeres of Human-Chromosomes - High-Frequency Array-Length Polymorphism and Meiotic Stability. *Proceedings of the National Academy of Sciences* **86**(23): 9394-9398.

Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AF, Finn RD. 2013. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res* **41**(Database issue): D70-82.

Wiesendanger B, Lucchini R, Koller T, Sogo JM. 1994. Replication fork barriers in the Xenopus rDNA. *Nucleic Acids Res* **22**: 5038-5046.

Worton RG, Sutherland J, Sylvester JE, Willard HF, Bodrug S, Dubé I, Duff C, Kean V, Ray PN, Schmickel RD. 1988. Human ribosomal RNA genes: orientation of the tandem array and conservation of the 5' end. *Science* **239**: 64-68.

Yang L, Duff MO, Graveley BR, Carmichael GG, Chen L-L. 2011. Genomewide characterization of non-polyadenylated RNAs. *Genome Biology* **12**: R16.

Yelin R, Dahary D, Sorek R, Levanon EY, Goldstein O *et al.* 2003. Widespread occurrence of antisense transcription in the human genome. *Nature biotechnology* **21**: 379-386.

Yoon Y, Sanchez J, Brun C, Huberman J. 1995. Mapping of replication initiation sites in human ribosomal DNA by nascent-strand abundance analysis. *Mol Cell Biol* **15**: 2482-2489.

Zentner GE, Saiakhova A, Manaenkov P, Adams MD, Scacheri PC. 2011. Integrative genomic analysis of human ribosomal DNA. *Nucleic Acids Res* **39**: 4949-4960.

Zentner GE, Scacheri PC. 2012. The chromatin fingerprint of gene enhancer elements. *The Journal of biological chemistry* **287**: 30888-30896.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**(5): 821-829.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS *et al.* 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* **9**(9): R137.

Zheng Z, Jia J-L, Bou G, Hu L-L, Wang Z-D, Shen X-H, Shan Z-Y, Shen J-L, Liu Z-H, Lei L. 2012. rRNA genes are not fully activated in mouse somatic cell nuclear transfer embryos. *The Journal of biological chemistry* **287**: 19949-19960.

Zody MC, Garber M, Sharpe T, Young SK, Rowen L *et al.* 2006. Analysis of the DNA sequence and duplication history of human chromosome 15. *Nature* **440**: 671-675.

[Blank Page]

[Blank Page]