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]