Identification and Annotation of Recombinant Repeats In Mammals Indicates They Are Experimental Products For Creating Novel Transposable Element Families

Sim Lin Lim

A thesis submitted for the degree of Doctor of Philosophy

Discipline of Genetics

School of Molecular and Biomedical Science

The University of Adelaide

October 2013

Table of Contents

Abstract

About 40-50% of mammalian genomes are made up of repetitive elements, primarily transposable elements. Transposable elements' activities not only drive genome evolution, they contribute to the creation of novel recombinant repeats. Recombinant repeats have largely remained uncharacterized due to their complexity. Initially, I developed a pipeline for the genome wide identification of recombinant repeats in four different mammals: human, mouse, cow and horse. The pipeline identified 1,336,824 copies, but only 37,830 sequences were able to be clustered into 6,116 families. The majority of the recombinant repeats were simple recombinant repeat families and only a small proportion were complex recombinant repeat families. My analysis showed that recombinant repeat families only covered a small fraction of the genomes examined (0.30% in human, 0.13% in mouse, 0.217% in horse and 0.464% in cow), indicating most of the recombinant repeats were singletons. Further analysis has shown that both classes of RR were created via transposon-into-transposon events, indicating that novel transposable elements are likely to be created via this mechanism. I found that simple recombinant repeats were probably retrotranspositionally active because they contained polyA tails and target site duplications, showing that they integrated into the genome via retrotransposition events. However, complex recombinant repeat families were only replicated via segmental duplications. My analysis showed that complex recombinant repeat families are excellent candidates for the identification of genome segmental duplication regions that cannot be found through standard methods. In addition, I used the RR identification pipeline to annotate possible RR in pig genome. I discovered a

novel RR family (LTR2i_SS) that contained > 1,000 copies. Repeat annotation showed that it was a chimeric LTR2_SS that contained ~300bp of un-annotated sequence, only found in the pig genome. Further investigation revealed that some LTR2i_SS flanked β3 proviruses, but these proviruses were unable to replicate autonomously as they did not encode a functional, complete polyprotein. My phylogenetic tree analysis of the LTR2i_SS and LTR2_SS familis suggested that LTR2i_SS was the ancestral form of LTR2_SS. In conclusion, I was able to identify the recombinant repeat distributions in different mammals and determine their most probable origin as TinT events. I have shown that recombinant repeats could serve as an important model to explain the origin of novel transposable elements in genomes, or could be used as markers to identify structural variations, or segmental duplications in different species. However, my data have also shown that we have to be cautious when annotating novel recombinant repeats in genomes, as they could be the ancestral form of other known transposable elements rather than novel forms generated through TinT.

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In additon, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicatble, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis (as listed below) resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed.……………………………….…………....Date……………………………….

Acknowledgements

I would like to express my sincere gratitude to the following people:

Professor David Adelson, for supervision and guidance, and for giving me the opportunity and support to do the PhD. I am so lucky to have Dave to be my mentor. The knowledge, lessons and experiences that I learned from you is not only valuable for my PhD, but it is an invaluable experiences for my future career.

Professor Hamish Scott, my co-supervisor, for providing me the sequencing data for analysis.

Dr Dan Kortchak, who guided me in solving computational and bioinformatic questions, read my manuscripts and always provided valuable suggestions; Joy Raison, for helping me to solve statistical questions, and all other members of the Adelson lab, past and present, for making it such a supportive and enjoyable environments to work.

Dong Wang in the Timmis lab, and everyone else in the MLS building that has helped me along my PhD research.

My wife, Sanny and my parents, for always encouraging and supporting me through my PhD, and for their endless love throughout my life. My sister and brother who have always providing me supports and encouragements.