COMPUTATIONAL IDENTIFICATION AND ANALYSIS OF EVOLUTIONARY BREAKPOINT REGIONS IN AMNIOTE GENOMES

Jitendra Narayan

Supervisors

Dr. Denis Larkin and Dr. Neil McEwan

A thesis submitted for the degree of

Doctor of Philosophy

at

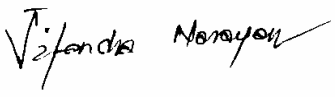
Institute of Biological, Environmental and Rural Sciences

Aberystwyth University

2014

DECLARATION

This work has not previously been accepted in substances for any degree and is not being concurrently submitted in candidature for any degree.

Signature ……  ……. [Candidate]

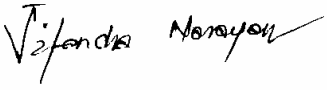
Date ……

STATEMENT 1

This thesis is the result of my own investigations, excepts where otherwise stated. Where *correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references.

A bibliography is appended.

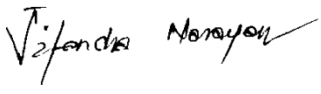
Signature ……  ……..… [Candidate]

Date ……

[*This refers to the extent to which the text has been corrected by others]

STATEMENT 2

I hereby give consents for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organization.

Signature ……  …… [Candidate]

Date ……

ABSTRACT

Genomes undergo mutation during evolution. Out of several mutational events, large-scale mutations, called genome rearrangements, mainly contribute to large-scale structural changes in chromosomes. My study of genome rearrangements mainly concentrates on identifying chromosomal evolutionary breakpoint regions and connects these to changes gained by each species during the course of evolution. In this thesis, I first focused on comparative genome analysis of seven mammalian genomes and discovered 192 evolutionary breakpoints in the pig genome. Subsequently, an extensive study demonstrated how chromosomal rearrangements produced variations in the gene networks potentially used by natural selection for adaptation. Thereafter, I developed a novel computational tool which uses a statistical method to find breakpoints in chromosomes with respect to various genome attributes, such as genome size, assembly type, and the phylogenetic relationship between species. The published cattle EBR dataset was used to test the algorithm, in which I was able to classify upto 95.55% of cattle specific EBRs. The comparative analysis of avian genomes demonstrates that there are lower rates of chromosome evolution as well as the presence of lower fractions of transposable elements in bird genomes compared to mammals. Our study revealed enrichment for Gene Ontology terms related to *regulation of gene expression* and *biosynthetic processes* in bird, crocodile and turtle HSBs. The archosaurian HSBs were found enriched for genes that are responsible for the similar retina structures in birds and crocodiles, while the avian HSBs contain genes involved in the bird skeleton and limb development. Moreover, the analysis of gene content in and around avian EBRs revealed enrichments for genes related to lineage-specific phenotypes, such as the GO terms "*regionalisation"* in the Adelie penguin and "*forebrain development"* in the Budgerigar. Our findings shed light on mechanisms underlying adaptation, development, and evolution at the genomic level.

ACKNOWLEDGMENTS

The submission of this thesis brings to an end a wonderful period, of almost 3 years, in which I was a student of IBERS, Aberystwyth University. On my way to complete this thesis, I have experience may joyous moments, as well as lovable hurdles. I would like to thank those who gave me the strength and courage to continue and press forward.

I am deeply grateful to my supervisors, Dr. Denis Larkin and Dr. Neil McEwan, for their guidance, encouragement, criticism, and faith throughout this PhD, without them there would have been no project to start, enjoy and complete. I especially would like to express my deepest gratitude to my mentor and advisor, Dr. Denis Larkin. He has been a role model for me, with his broad knowledge, inquisitive mind, uncompromising integrity, and enviable ability to conduct many diverse researches in parallel. He spends countless hours in training, encouraging, and improving my scientific writing skills. I would like to thank him for introducing me into the field of chromosomal breakpoints, and tolerance of my idiosyncrasies. Apart from him, Dr. Gancho Slavov's invaluable inputs in the chromosomal breakpoint algorithm development were impeccable. He assisted me with carrying out some of the complex statistical analyses. His support gave me a more profound knowledge of statistics. In addition, I would also like to thank Dr. Katie Fowler at University of Kent for FISH analysis and validation of pig breakpoints.

I would like to thank the incredibly multi-talented Dr. Marta Ferre Belmonte, Dr. Robert Vickerstaff, whose emotional support and friendship have been invaluable, Dr. Martin Swain for moral support and encouragements. Thank you for lending a sympathetic ear and giving useful advice. Thanks to all of my friends, especially Sarah Beynon, Vasileios Panagiotis Lenis, Stefani Dritsa, Martin Vickers, and Altan Kara for their love, patience, and most importantly their ability to keep me sane.

Last, but not least, I would like to thank my loving parents for instilling in me the love of learning and the continuous desire for more knowledge. Finally, I would like to thank all the teachers and professors that have transmitted passion for science and education to me. I hope someday I am able to have the same impact on my students.

DEDICATION

I dedicate my research work to my dearest parents and many friends, who has always been a source of love, encouragement and support throughout my study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

Standard international code of zoological nomenclature and international units of measurement are used. In addition, the following abbreviations are referred in this thesis.

Acronym	Definition
3C	Chromosome conformation capture
4C	Circular chromosome conformation capture
5C	3C-carbon copy
BAC	Bacterial artificial chromosome
BGI	Beijing genomics institute
Bp	Basepair
BRs	Breakpoint regions
BTA	*Bos taurus* chromosome
CNVs	Copy number variants
CS	Conserved segments
CSREES-USDA	Cooperative state research, education and extension service at the United States department of agriculture
DAVID	Database for annotation, visualization and integrated discovery
ddNTPs	dideoxynucleotide triphosphates
DNA	Deoxyribonucleic acid
EBA	Evolutionary breakpoint analyser
EBR	Evolutionary breakpoint region
EH	Evolution highway
ENCODE	Encyclopedia of DNA elements
FDR	False discovery rate
FISH	Fluorescence *in situ* hybridization
FM	Fragile model
G10K	Genome 10K

G10KCOS	Genome 10K community of scientists
Gb	Gigabasepair
GO	Gene ontology
HiC	Hydrophobic interaction chromatography
HSA	*Homo sapiens* chromosome
HGP	Human genome project
HSB	Homologous synteny block
ISEA	Island south east Asia
Kb	Kilobasepair
KEGG	Kyoto encyclopedia of genes and genomes
K-T	Cretaceous-Tertiary
Mb	Megabasepair
mRNA	Messenger RNA
msHSB	Multi species homologous synteny block
MT	Mitochondria
mtDNA	mitochondrial DNA
MYA	Million years ago
NAHR	non-allelic homologous recombination
NGS	Next generation sequencing
NCBI	National center for biotechnology information
NHGRI	National human genome research institute
NIH	National institute of health
OM	Optical mapping
OR	Olfactory receptor
PERL	Practical extraction and report language
PBS	Phosphate buffered saline
QSRA	Quality-value guided *de novo* Short Read Assembler
QTL	Quantitative trait loci
RACA	Reference assistant chromosome assembly
RBM	Random breakage model

RFLPs	Restriction fragment length polymorphisms
RH	Radiation hybrid
RNA	Ribonucleic acid
ROB	Robertsonian translocation
SF	Syntenic fragment
SINE	Short interspersed nuclear element
SGSC	Swine genome sequencing consortium
SMRT	Single molecule real time
SNA	Single-nucleotide addition
SNP	Single nucleotide polymorphism
SSC	*Sus scrofa* chromosome
TAS2R	Taste receptor, type 2
TCC	Tethered conformation capture
TE	Transposable element
TFBM	Turnover fragile breakage model
tRNA	transfer RNA
TT	Taste transduction
UN	Unplaced
UTRs	Untranslated regions
WGS	Whole genome sequencing
WHO	World health organization

PREFACE

The part of the work presented in this thesis is based on the collection of two articles published and two more are communicated throughout the PhD period in scientific journals and in refereed proceedings of conferences. These are included at the end of this thesis.

Peer-reviewed journal articles

- Groenen, M.A., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.J., Li, S., Larkin, D.M., Kim, H., Frantz, L.A., Caccamo, M., Ahn, H., Aken, B.L., Anselmo, A., Anthon, C., Auvil, L., Badaoui, B., Beattie, C.W., Bendixen, C., Berman, D., Blecha, F., Blomberg, J., Bolund, L., Bosse, M., Botti, S., Bujie, Z., Bystrom, M., Capitanu, B., Carvalho-Silva, D., Chardon, P., Chen, C., Cheng, R., Choi, S.H., Chow, W., Clark, R.C., Clee, C., Crooijmans, R.P., Dawson, H.D., Dehais, P., De Sapio, F., Dibbits, B., Drou, N., Du, Z.Q., Eversole, K., Fadista, J., Fairley, S., Faraut, T., Faulkner, G.J., Fowler, K.E., Fredholm, M., Fritz, E., Gilbert, J.G., Giuffra, E., Gorodkin, J., Griffin, D.K., Harrow, J.L., Hayward, A., Howe, K., Hu, Z.L., Humphray, S.J., Hunt, T., Hornshoj, H., Jeon, J.T., Jern, P., Jones, M., Jurka, J., Kanamori, H., Kapetanovic, R., Kim, J., Kim, J.H., Kim, K.W., Kim, T.H., Larson, G., Lee, K., Lee, K.T., Leggett, R., Lewin, H.A., Li, Y., Liu, W., Loveland, J.E., Lu, Y., Lunney, J.K., Ma, J., Madsen, O., Mann, K., Matthews, L., McLaren, S., Morozumi, T., Murtaugh, M.P., **Narayan, J.,** Nguyen, D.T., Ni, P., Oh, S.J., Onteru, S., Panitz, F., Park, E.W., Park, H.S., Pascal, G., Paudel, Y., Perez-Enciso, M., Ramirez-Gonzalez, R., Reecy, J.M., Rodriguez-Zas, S., Rohrer, G.A., Rund, L., Sang, Y., Schachtschneider, K., Schraiber, J.G., Schwartz, J., Scobie, L., Scott, C., Searle, S., Servin, B., Southey, B.R., Sperber, G., Stadler, P., Sweedler, J.V., Tafer, H., Thomsen, B., Wali, R., Wang, J., Wang, J., White, S., Xu, X., Yerle, M., Zhang, G., Zhang, J., Zhang, J., Zhao, S., Rogers, J., Churcher, C. and Schook, L.B. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491, 393-398.

- Guojie Zhang, Cai Li, Qiye Li, Bo Li, Denis M. Larkin, Chul Lee, Jay F. Storz, Agostinho Antunes, Matthew J. Greenwold, Robert W. Meredith, Anders Ödeen, Jie Cui, Qi Zhou, Luohao Xu, Hailin Pan, Zongji Wang, Lijun Jin, Pei Zhang, Haofu Hu, Wei Yang, Jiang Hu, Jin Xiao, Zhikai Yang, Yang Liu, Qiaolin Xie, Hao Yu, Jinmin Lian, Ping Wen, Fang Zhang, Hui Li, Yongli Zeng, Zijun Xiong, Shiping Liu, Long Zhou, Zhiyong Huang, Na An, Jie Wang, Qiumei Zheng, Yingqi Xiong, Guangbiao Wang, Bo Wang, Jingjing Wang, Yu Fan, Rute R. da Fonseca, Alonzo Alfaro-Núñez, Mikkel Schubert, Ludovic Orlando, Tobias Mourier, Jason T. Howard, Ganeshkumar Ganapathy, Andreas Pfenning, Osceola Whitney, Miriam V. Rivas, Erina Hara, Julia Smith, Marta Farré, **Jitendra Narayan**, Gancho Slavov, Michael N Romanov, Rui Borges, João Paulo Machado, Imran Khan, Mark S. Springer, John Gatesy, Federico G. Hoffmann, Juan C. Opazo, Olle Håstad, Roger H. Sawyer, Heebal Kim, Kyu-Won Kim, Hyeon Jeong Kim, Seoae Cho, Ning Li, Yinhua Huang, Michael W. Bruford, Xiangjiang Zhan, Andrew Dixon, Mads F. Bertelsen, Elizabeth Derryberry, Wesley Warren, Richard K Wilson, Shengbin Li, David A. Ray, Richard E. Green, Stephen J. O'Brien, Darren Griffin, Warren E. Johnson, David Haussler, Oliver A. Ryder, Eske Willerslev, Gary R. Graves, Per Alström, Jon Fjeldså, David P. Mindell, Scott V. Edwards, Edward L. Braun, Carsten Rahbek, David W. Burt, Peter Houde, Yong Zhang, Huanming Yang, Jian Wang, Avian Genome Consortium, Erich D. Jarvis, M. Thomas P. Gilbert, and Jun Wang, Science 12 December 2014, Comparative genomics reveals insights into avian genome evolution and adaptation, 346 (6215), 1311-1320.

In addition, this thesis contains the following two articles, one of them is communicated recently for publication in a peer reviewed journal, and the second one is in preparation.

- Farré M., **Narayan J.**, Slavov G., Auvil L., Li C., Jarvis E.D., Burt D.W., Griffin D., Larkin D.M. 2014. Comparative analysis of 25 birds and reptile genomes reveals features of chromosome evolution in birds, archosaurians, and reptiles (2014) PNAS (Communicated).

- **Jitendra Narayan,** Marta Ferre Belmonte, Gancho Slavov, Denis Larkin. An algorithmic approach to identify and classify EBRs in sequenced amniote genomes (In preparation).

1. INTRODUCTION

A species genome is constantly changing in evolution to adapt the host organism to the ever-changing environment. Over time, genomes accumulate information about their evolutionary history. In animals, this information is passed to the next generations through cell division in the process called "meiosis". For a long time it was believed that the main evolutionary changes in genomes that have adaptive values are small changes in the coding parts of genes ("single nucleotide mutations") leading to changes of amino acids in proteins (Ackers and Smith 1985, Ng and Henikoff 2006). These, so called "point" mutations do indeed affect gene products by producing aberrant and non-functional proteins (mis-sense mutations), or by changing the physical properties of proteins (non-synonymous mutations) (Miyata *et al.* 1979, Betts and Russell 2003). If a gene accumulates too many non-synonymous mutations, the resulting protein could even change its function compared to the original protein leading to the birth of a novel gene and protein (Hoyle and Wickramasinghe 2000). With the growing understanding of genome function and evolution, it became clear that in addition to the point mutations other events might play an important role in the adaptive changes of organisms. One type of such event is the structural DNA changes called "chromosome rearrangements" (Griffiths *et al.* 1999). This event affects the order and position of genes in chromosomes and is often associated with gene duplications and deletions that occur at their boundaries (so called evolutionary breakpoint regions (EBRs). While the nature and mechanism of chromosome rearrangement formation are different from those of point mutations, multiple evidence collected from different taxa show that these events also play a crucial role in genome evolution and organism adaptation (Crombach and Hogeweg 2007, Bovine Genome *et al.* 2009, Larkin 2012). In addition, the exploration of the rearrangement history of a set of genomes allows for an in-depth understanding of the evolutionary history of the corresponding organisms (William J Murphy et al. 2005). Therefore, a study of genome organisation and chromosomal rearrangements using whole genome sequences is important to better understand the evolutionary history of organisms and ways of adaptations in clades and individual lineages.

There has been a long debate about whether chromosome rearrangements contribute to speciation and adaptation (Ohno 1973, King 1995, Loren H Rieseberg 2001, Pérez-Ortín *et al.* 2002, Navarro and Barton 2003, F. J. Ayala and M. Coluzzi 2005, Butlin

2005, Brown and O'Neill 2010, Faria and Navarro 2010, Chang *et al.* 2013, Ayala *et al.* 2014, Hou *et al.* 2014). While some chromosome rearrangements most likely contribute to the speciation process by building reproduction barriers between populations in lower taxa (Sites and Moritz 1987, Noor *et al.* 2001, Loren H Rieseberg 2001), their contribution to speciation and adaptive changes in higher taxa is still unclear (White 1969, Bush *et al.* 1977, Jian Lu *et al.* 2003, Navarro and Barton 2003, Faria and Navarro 2010, Servedio *et al.* 2011).

Recently, several genome sequencing projects have provided us with high quality genome sequences. These genomes of phylogenetically-related and distinct species are assembled to chromosomes or scaffolds and provide the basis for a detailed exploration of genome dynamics. The genomic information can be used to better understand the changes in the genomic architecture of organisms which happen during the course of evolution. In addition, genome resources provide a means for addressing questions about the influence of genomic rearrangements on adaptation in higher taxa at a new level (Pevzner and Tesler 2003b, W. J. Murphy et al. 2005, Larkin et al. 2009, Ruiz-Herrera et al. 2012).

The various novel computational methods and tools[1] have been recently developed to identify regions of shared synteny i.e., homologous synteny blocks (HSBs), and EBRs among the growing number of sequenced genomes of different species (Bourque *et al.* 2004, Ruiz-Herrera *et al.* 2004, Ruiz-Herrera *et al.* 2006, Larkin *et al.* 2009, Farre *et al.* 2011). The molecular and computational analysis of EBRs has revealed that they are not randomly distributed in genomes, but tend to cluster in break-prone genome intervals i.e., in hotspots of genome rearrangements (Bourque *et al.* 2004, Ruiz-Herrera *et al.* 2004, Ruiz-Herrera *et al.* 2006, Larkin *et al.* 2009, Farre *et al.* 2011). The EBRs are associated with several genomic features such as gene-rich regions (Everts-van der Wind *et al.* 2004, Ma *et al.* 2006), chromosome fragile sites (Ruiz-Herrera *et al.* 2006), and an elevated frequency of segmental duplications and repetitive elements (Bailey et al. 2004, William J Murphy et al. 2005). It addition, the GC content and CpG islands were found enriched in chicken EBRs. This, therefore, could highlight a potential role for these genomic features in evolutionary instability of genome structures. The evolutionary features mentioned above have nourished a growing fascination in chromosomal

[1] http://bioinformaticsonline.com/blog/view/4574/tools-to-detect-synteny-blocks-regions-among-multiple-genomes Accessed: 14/10/2014

evolution, particularly on the relationship between chromosome rearrangements and species adaptation to the environment.

Although tremendous progress has been made in recent years towards determining the relationship between EBRs and various sequence features and their association with probable mechanisms of chromosome breakage in evolution (William J Murphy et al. 2005, Ruiz-Herrera et al. 2006, Gordon et al. 2007, Larkin et al. 2009, Larkin 2012, Farré et al. 2013, Bose et al. 2014), the role of EBRs in adaptation to the environment is unclear. Henceforth, this poses several fundamental questions: How does one detect and classify EBRs across phylogenetically related species? Can EBRs be accurately classified using a statistical framework? Are EBRs enriched for genes underlying adaptation of species to the ever changing environment?

Recent molecular and computational advances, coupled with the availability of amniote (reptile, avian and mammalian) whole genome sequences[2] make it possible to start addressing the above mentioned questions. Hence, the main objective of this thesis was to understand how chromosome rearrangements affect amniotes evolution, focusing mainly on the relationship between chromosomal rearrangements and adaptation to the environment. This work therefore focuses on the detection and classification of EBRs and the role of evolutionary rearrangements in clade and species-specific biology in two classes – mammals (using pig as an example) and reptiles (using comparison of genomes from 21 bird species). The main objective can be sub-divided into three specific aims:

1. Identify chromosome rearrangements and detect pig-specific EBRs to elucidate their influence on the pig lineage-specific biology.
2. Develop a novel computational algorithm to automatically detect and assign EBRs to phylogenetic nodes by taking into account phylogenetic relationships of the genomes involved in the analysis.
3. Application of the novel tool developed to the detection and study of the role of rearrangements in *de novo* sequenced bird genomes.

The work presented in this thesis has permitted the first computational analysis of the relationship between chromosome organisation, genome rearrangements, and

[2] https://genome10k.soe.ucsc.edu/ Accessed: 14/10/2014

adaptation in pigs and birds. Overall, several scientific findings and a computational method will be reported:

- The application of comparative genomics methods to several mammalian genomes revealed a large number of EBRs in the pig lineage and their impact on the pig genome evolution.

- The development of a novel algorithm to identify EBRs and assign them to proper phylogenetic nodes. The algorithm was implemented into a user-friendly tool which identifies and assigns EBRs to phylogenetic nodes based on the phylogenetic relationships provided by user or downloaded from the NCBI.

- The use of the algorithm for the comparative study of 21 avian, and five non-avian species to address fundamental questions of genome organisation and chromosome evolution in birds and reptiles.

This thesis is organised and proceeds as follows:

- The chapter 2 will cover an in depth literature review on the genome structure, genome organization followed by an introduction to genome mapping techniques. It then covers a general background of genome evolution, synteny, and chromosomal rearrangements.

- The work described in the chapter 3 covers the pig chromosome evolution analysis using seven sequenced and assembled mammalian genomes. This chapter focuses on the chromosomal rearrangement events that have occurred in artiodactyl species with a particular focus on the evolutionary events present in the pig genome. It also covers the computational analysis of the gene content in and around pig EBRs and demonstrates that chromosomal rearrangements introduce changes in the gene networks and these changes are likely to be used by the natural selection for adaptation.

- Chapter 4 introduces and discusses the algorithm developed to perform an automated identification and classification of EBRs from a large number of

genomes taking into account their phylogenetic relationships. This software tool named as "Evolutionary Breakpoint Analyzer" (EBA) can be used not only for the genomes assembled to chromosomes, but also with the genomes that have fragmented scaffold-based assemblies.

- Chapter 5 covers the application of the EBA tool to a set of 21 avian, and five non-avian genomes. The EBA tool detects and classifies lineage- and group-specific EBRs. Later, the enrichment analysis for transposable elements (TE) and genes related to lineage-specific phenotypes were done and patterns similar to those observed in mammalian genomes were observed. Our first comprehensive and large scale genome analysis of bird and reptile genome rearrangements provides a resource for studying the nature of karyotype stability in birds. In addition, our results demonstrate how the chromosome rearrangements could have contributed to the maintenance of ancestral and formation of novel phenotypes in reptiles.

- Finally, chapter 6 is an in depth discussion of the results presented in this thesis and outlines some future directions.

2. LITERATURE REVIEW

SECTION 1: GENOME STRUCTURE AND MAPPING

1.1 AMNIOTE GENOME: AN OVERVIEW

The nuclear genome is composed of deoxyribonucleic acid (DNA), which holds information about an organism's development, physiology, and evolution. Additionally, eukaryotes also bear organelles genomes contained within mitochondria and chloroplast. (Schwartz and Dayhoff 1978). Each genome of an organism contains genes that encode for proteins with particular structures and functions, and these proteins are a building block of living organisms. The phenotype of any organism is determined by their genetic makeup and the environmental pressures to which the organism is subject. Both the number of base pairs and the number of genes vary widely from one species to another, and there is only a rough correlation between the two, an observation known as the C-value paradox (Thomas Jr 1971). At present, the organism with the most known genes is the trichomoniasis-causing protozoan, which has a genome containing approximately 60,000 genes, almost three times as many as found in the human genome[3]. In the early 1970s the discovery of non-coding DNA resolved the question of the C-value paradox to some extends (Thomas Jr 1971, Elgar and Vavouri 2008). It has been hypothesised that genome size does not reflect the number of genes in eukaryotes. This is because most of the DNA is non-coding (i.e., does not code for proteins) and henceforth does not consist of genes. Such cases are clearly visible in the human genome, in which protein-coding regions comprise less than 2% of the nuclear genome. However, the Encyclopedia of DNA Elements (ENCODE)[4] project has built a comprehensive list of functional elements in the human genome and states that while the gene coding portion of the genome is only 2% of base pairs, 80% of the human genome is still comprised of "functional DNA" (Consortium 2004). These functional DNA regions or elements are biologically relevant; they may be promoters or parts of other regulatory elements. Moreover, a positive correlation between biological complexity and the amount of non-coding DNA has been reported, which suggests

[3] http://www.genomesize.com/statistics.php
[4] https://www.encodeproject.org/

introns, intergenic sequences, repeat elements have far more importance than thought previously (Taft and Mattick 2004).

Prokaryotes are distinguished from eukaryotes in many ways. In the genetic material of prokaryotes is not bound by a membrane, whereas eukaryotic cells contain membrane-bound organelles, such as the nucleus. Additionally, the differences in cellular structure of prokaryotes and eukaryotes include the presence of mitochondria and chloroplasts, the cell wall, and the structures of their chromosomal DNA. Prokaryote genomes contain only a single loop of stable chromosomal DNA stored in an area named the nucleoid, whereas eukaryotic genomes are tightly bound and organised into chromosomes found within the nucleus. Prokaryotic genes lack intron and the majority of their genomes code for proteins, whereas a large portion of eukaryotic genome does not encode for proteins or transcribed RNA. Prokaryotic genes are expressed in groups known as operons, while eukaryotes express genes individually (Lodish 2008).

The amniotes, which includes turtles, lizards, birds, dinosaurs, and mammals, last shared a common ancestor approximately 310 MYA and diversified dramatically during the Carboniferous period (Deakin and Ezaz 2014). Amniotes have been laying eggs for millions of years. Their eggs consist of a membrane bound shell filled with an amnios to prevent developing embryos from drying out. These adaptations enable them to lay eggs on land rather than in water as anamniotes do. While most modern mammal do not lay eggs, one group of mammals, the monotremes, still do (Hall 2008). The amniote embryos are protected and aided by several membranes. These membranes contain the amniotic sac that surrounds the foetus in eutherians (placental mammals). The first known basal amniotes resembled small lizards. The unique ability of small lizard eggs to survive out of water, "breathe", and cope with wastes empowered amniotes to diversify, adapt to drier environments, and evolve into larger forms (Hall 2008). Interestingly, despite the common origin of amniote lineages, they have strikingly different chromosomes (Figure 2.1). This genomic diversity directly suggests that amniote genomes have undergone a considerable amount of chromosomal rearrangement since they last shared a common ancestor (Deakin and Ezaz 2014).

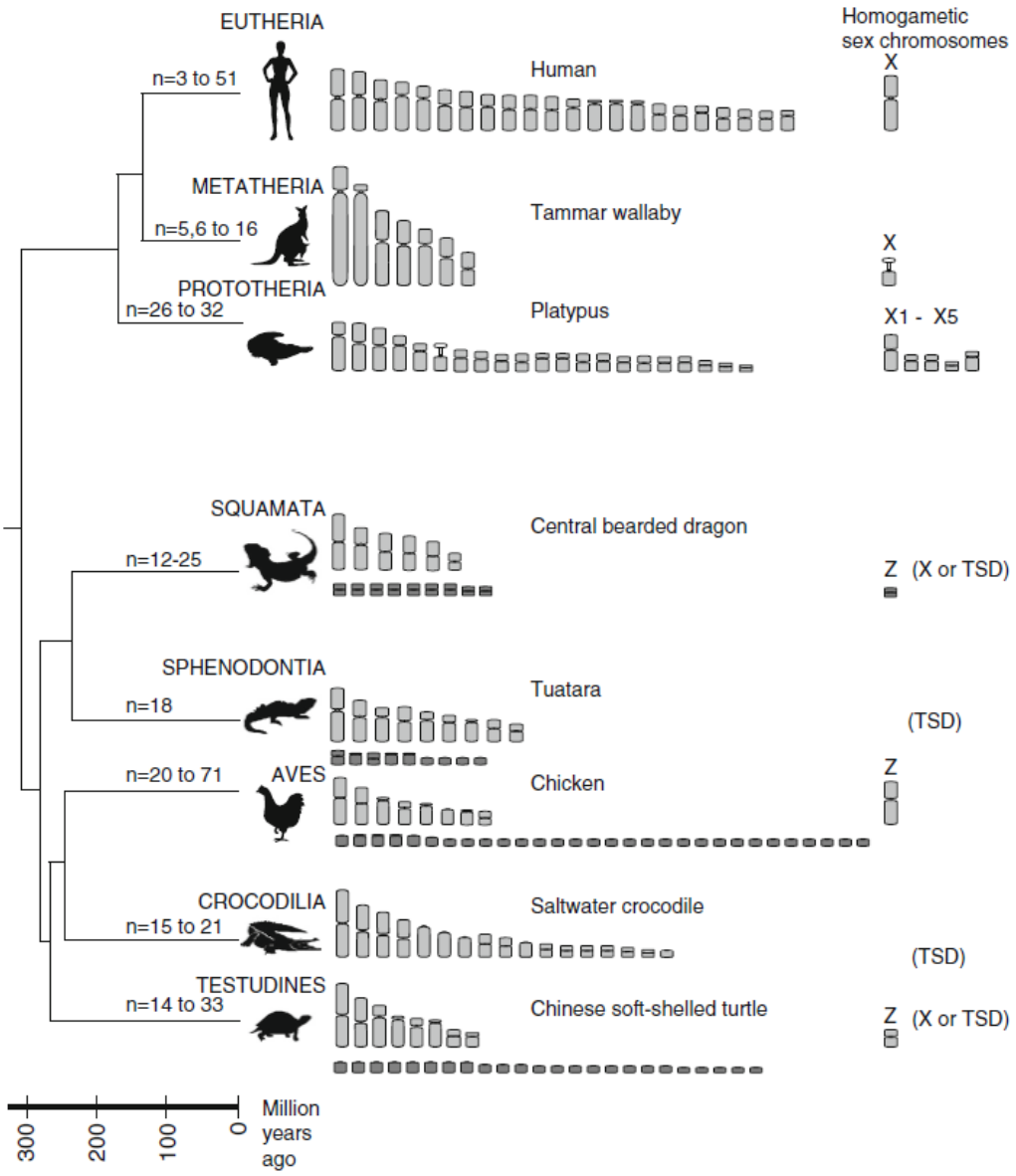


Figure 2.1 Amniote phylogeny with representative karyotypes. The haploid chromosome number and range are indicated on respective branches. The micro-chromosomes are denoted by a dark grey colour (Deakin and Ezaz 2014).

Genomics has been a boon to evolutionary biologists, as it has enabled the exploration of the evolution of genomes amongst taxa such as amniotes. Compared to other animals, avian have fewer repetitive elements, lower GC content, and genome size variation, and they also have comparatively small genomes as well (Shedlock *et al.* 2007). Such cases are also reported in alligator and turtle, in which the genome sizes are 30% smaller than human (David W Burt et al. 1999). Additionally, alligator, turtles, and chicken genomes have a significant number of micro-chromosomes (David W Burt et

al. 1999). Avian genomes are gene rich, as reported in chickens (Ellegren 2005). Soft-shelled turtles exhibit an extinction of repetitive elements (Shedlock *et al.* 2007).

1.2 GENOME ORGANISATION

The genomic DNA segments that encodes for a polypeptide or a functional RNA are called genes. Genes are sometime also called "protein coding DNA", but recently it has been determined that a gene does not need to code for a protein. The flow of genetic material from DNA to RNA to protein is known as the central dogma of biology. In other words, "DNA makes RNA, RNA makes proteins, which in turn facilitate the previous two steps as well as the replication of DNA" with a few notable exceptions. The entire process is further broken down into the following steps: transcription, splicing, translation, and replication. The first step is transcription, in which a section of DNA is transferred to a newly assembled piece of messenger RNA (mRNA) by RNA polymerase and transcription factors. In eukaryotic cells the primary transcript (pre-mRNA) is processed, and one or more sequences (introns) are cut out via the mechanism of alternative splicing. Thereafter, the mature mRNA is read by the ribosome as triplet codons. Triplet codons usually begin with an AUG, or initiator methonine codon downstream of the ribosome binding site. Complexes of initiation factors and elongation factors bring aminoacylated transfer RNAs (tRNAs) into the ribosome-mRNA complex, matching the codon of the mRNA to the anti-codon of the tRNA, thereby adding the correct amino acid into the sequence encoding the gene. The final element of the Central Dogma is transmission of genetic information from parents to progeny, that is, the DNA must be replicated faithfully. Replication is carried out by a complex group of proteins that unwind the double-stranded DNA helix, and, using DNA polymerase and its associated proteins, copy or replicate the master template itself so the cycle can be repeated, from DNA to RNA to proteins in a new generation of cells or organisms.

In order to store the entire genome within the microscopic nuclear space, DNA molecules must undergo many levels of structural and biochemical compactisation resulting in discrete nuclear 'environments' (Figure 2.2). Moreover, this complex genome organisation must be dynamically responsive as cells go about the process of producing functional proteins and respond to environmental challenges. In other words,

the genome should be organised in such a way that the execution of various biological processes such as gene expression, protein interaction and gene regulation should be possible. Moreover, the genome organisation, expression and regulation complexities increase with chromosome numbers and also with the number of genes present in an organism (Assis *et al.* 2008). The condensed and systematically packed chromosomes in nucleus have special spatial organisation with territories, which tend to change with increased gene expression and after chromosomal rearrangements (Finlan and Bickmore 2008).

The high resolution mapping technologies for spatial chromatin structure such as chromosome conformation capture (3C), circular chromosome conformation capture (4C), 3C-carbon copy (5C), hydrophobic interaction chromatography (HiC), tethered conformation capture (TCC) techniques guide researchers in exploring spatial genome organisation with respect to structure and functions (Göndör and Ohlsson 2009, Belton *et al.* 2012, Gibcus and Dekker 2013). Even before the above mentioned high-throughput molecular biology methods, the microscopy and ChIP (chromatin immunoprecipitation) was the main approach to study arrangement of chromosomes and their interactions in the nucleus. The newly developed 3C technique combined with ultra-high-throughput DNA sequencing, dramatically increased the scale relative to the ChIP method, at which physical interactions between genomic elements can be studied (Splinter et al. 2004). The 4C is an upgraded version of 3C which allows for the detection of *unknown DNA regions* of interaction with the region of interest (Ohlsson and Göndör 2007). The 5C, a high-throughput version of 3C for large-scale mapping of chromatin interaction networks, which employs quantitative DNA sequencing using 454-technology or microarray as detection methods (Dostie et al. 2006). In order to enable the research community to adopt 5C, to study, visualise and analyse the large chromatin interaction a new technology the 'my5C'[5] has been developed. It allows detailed insights into the three-dimensional arrangements of complete genomes at kilobase resolution (Lajoie et al. 2009). Later, a genome-wide and unbiased method, Hi-C technology, came into existence which combines 3C with deep sequencing. In other words, the 5C method is more or less similar to Hi-C but the comparison is genome wide. These techniques enabled scientists to reveal both known hallmarks of nuclear organization such as chromosome territories formation, and preferred co-locations of

[5] http://my5c.umassmed.edu/welcome/welcome.php

particular pairs of chromosomes, as well as novel folding principles of chromosomes (Van Berkum et al. 2010, Nagano et al. 2013). In addition, the new molecular biology techniques and recently completed genome projects are assisting to reveal a great deal about how genomes are organised, expressed and genes are regulated in a cell (Belton *et al.* 2012). A recent study by Dixon *et al.* reported that topological boundaries of chromatin interaction are enriched for an insulator binding protein, housekeeping genes and short interspersed elements (SINE) retro-transposons suggesting their role in establishing the topological domain structures of the genome (Dixon *et al.* 2012). Additionally, the genome organisation study also reveals the modular organisation and their triggering effect on dynamic chromosome structure and role in genome activity (Nagano *et al.* 2013). The genomic spatial heterogeneity and their contribution in recurrent chromosomal translocations have also been reported, in which translocation was found to be significantly enriched in cis along single chromosomes containing target DNA double-strand breaks (DSBs) and within other chromosomes and sub-chromosomal domains. These findings suggest the role of spatial heterogeneity, which allowed recurrent DSBs to drive translocation (Zhang *et al.* 2012). Moreover, the recent ENCODE project has also reported hundreds of long range interactions, which show strong correlation between gene expressions and the region of functional classes such as enhancers (Malin *et al.* 2013). In ENCODE they reported 2,324 and 19,813 genes involved in "single-gene" enhancer-promoter interactions and "multi-genes" interactions complex respectively. The multigene complexes found spanned up to several megabases, including promoter-promoter and enhancer-promoter interactions (Birney 2012, Hoffman *et al.* 2012).

Every living organism possesses a genome which contains the encrypted biological information needed to construct and maintain a living organism. This cellular life form's genome is made up of DNA (deoxyribonucleic acid), with a few exceptions like viruses have ribonucleic acid (RNA) genomes (Brown 2002). The DNA and RNA are polymeric molecules made up of chains of monomeric subunits called nucleotides (Watson and Crick 1953). The extensively studied, explored, and annotated human genome, is in many respects a fairly good model for eukaryotic genomes and analytical studies in general. All of the studied eukaryotic nuclear genomes are divided into two or more linear DNA molecules, each organised and arranged in a different chromosome (Pray 2008). In addition to the nuclear genetic material, the eukaryotes also possess

smaller and circular mitochondrial genomes with very few known genes (Cooper and Hausman 2000) . However, the human genome is not suitable to illustrate unique plant photosynthetic organelles which were specifically located in the chloroplasts of each plant cell nucleus.

All discovered and cytogenetically studied eukaryotes organisms are known to have at least two chromosomes with linear DNA molecules without any exceptions (Strachan and Read 1999). However, the only known variability noticed at this level of prokaryotic and eukaryotic genome structure lies with the number of chromosomes, which appears not to be correlated to the biological features of the organism (Pray 2008). Later, the structural genome organization and packaging system were further explored to better understand these complex mechanisms.

1.2.1 Packaging of DNA into chromosomes

The DNA molecules are much longer than the chromosomes they packed in. Henceforth, in order to store large DNA molecule a highly organised and sophisticated biological packaging mechanism were deployed to keep all of DNA molecules in chromosomes. To understand this complex biological packaging mechanism, Clark and Felsenfeld in 1971 carried out research on nuclear protection and organisation using biochemical analysis and electron microscopy techniques (Clark and Felsenfeld 1971). They used DNA-histone complexes to understand the packaging of single uninterrupted molecule of DNA which is tightly bound to a group of small, essential proteins called histones (Clark and Felsenfeld 1971). The DNA in the eukaryotic nucleus exists mainly in combination with histone proteins. These DNA–histone biological complexes with other protein that makes up the chromosome are termed as "chromatin". The chromatin, half DNA and half protein, can be envisioned as a repeat of structural units called "nucleosomes" which appeared similar to beads on a string through electron microscope (Figure 2.2) (Kornberg 1974). A nucleosome core particle is composed of histone octamer (H2A, H2B, H3, and H4) plus the DNA that wraps around the octamer in a left-handed supercoil in about 1.75 turns which encloses about 146bp (Clark and Felsenfeld 1971). Histone octamer H1 is a linker which works along with linker DNA (the DNA in between two nucleosome core particles) to physically connect the adjacent nucleosome core particles (Van Hoide et al. 1974, Wolffe 1998, Schwarzacher and Heslop-Harrison 2000).

The 30nm chromatin fibre is formed in the nucleus during interphase, the period between nuclear divisions (Luger and Hansen 2005, Woodcock 2005). The DNA adopts a more compact configuration during the nuclear division and packaging, resulting in the highly condensed metaphase structure. These condensed and compact structures can be seen with the light microscope, which have the appearance generally associated with the word 'chromosome'. In most of the bacteria these chromosomes are single in number and size 2 to 4.6 Mbp with up to 8288 genes (*E. Coli*). However, the chromosome size is much larger in eukaryotes which might go up to 1440 chromosomes as in *Ophioglossum reticulatum* (Khandelwal 1990, Grubben 2004)[6].

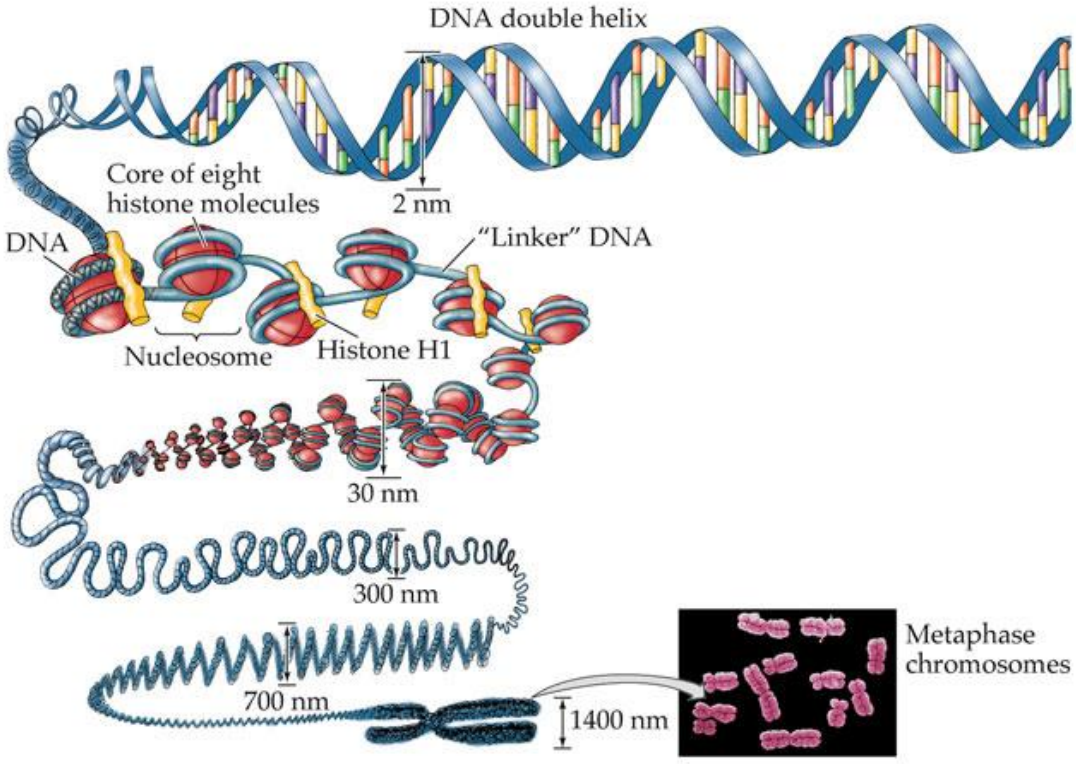


Figure 2.2 Packaging of DNA molecule. The chromatin (DNA-histone complexes) has a highly complex structure with several levels of structural compactisation. The simplest level is the double-helical structure of DNA. The image was adapted from (Purves et al. 2003).

The cells in an organism contain the highly ordered and packed DNA content; however the expression activity of genes in a genome changes during organism development leading to formation of specialised cells and tissues. Moreover, the spatial organisation

[6] http://www.genomesize.com/

studies indicate that chromosomes occupy distinct territories in the eukaryotic nucleus (Cremer and Cremer 2010). The gene expression regulation is seems to be correlated to the folding pattern and territories (Pederson 2004, Bártová and Kozubek 2006, Cremer and Cremer 2010, Halverson *et al.* 2014). The gene rich and poor regions tend to occupy different nuclear areas (Tanabe *et al.* 2002, Cremer and Cremer 2010). However, the dynamic organisation of chromosomes and repositioning of genome within territories are believe to play an important role in gene expression (Gasser 2002). The spatial organisation or the "3D genome organisation" bring together the genes located on different chromosomes, which is called as 'gene kissing' (Lanctôt *et al.* 2007, Bantignies and Cavalli 2011). These gene-gene interactions either contribute to transcriptional silencing (Francis *et al.* 2004) or activation (Lomvardas *et al.* 2006) or epigenetic gene network regulation (Murrell *et al.* 2004). It has also been reported that silencing or mutation in one of the kissing pair gene can affect the expression of the pair (Zhao *et al.* 2006). Moreover, the mounting biological evidence indicates the role of spatial organization of the genome, and their role in biological gene networks (Smallwood and Ren 2013). However, the chromosomal organization and dynamic nature of chromatin still a puzzle and scientist are trying to explore more about how these orchestrate vital role in the maintaining biological systems and controlling gene activity.

1.2.2 Genes

1.2.2.1 Gene duplication

Any set of two or more similar genes in one genome with similar biochemical function is known as gene family. These are generally formed by gene duplication, also known as chromosomal duplication or gene amplification (Figure 2.3) (Ohno 1970). The amplification might involve either large DNA segments or individual genes or exons (Betrán and Long 2002). Various natural biological events, such as homologous recombination, chromosome duplication, and retrotransposition events, promote the formation of new gene families (Meyer and Schartl 1999, Jiang *et al.* 2004, Volff 2006, Ranz *et al.* 2007). The duplicated gene families and their ancestry are generally identified using rigorous sequence similarities, phylogenetic analysis, and functional analysis techniques. Furthermore, these duplicated gene families and their ancestry are verified by examination of their secondary or tertiary structural organisation, which is conserved even if the sequences have diverged considerably (Roth *et al.* 2007). The expansion or

contraction of gene family that appears in lineage or order might be due to natural selection or random changes (Hahn *et al.* 2005, Demuth *et al.* 2006). One recent study found between ten and thousands of gene are duplicated every millions years throughout the vertebrate genome and reported that over the last 200 Myr the rate of duplication was 0.00115 Myr^{1} (Cotton and Page 2005).

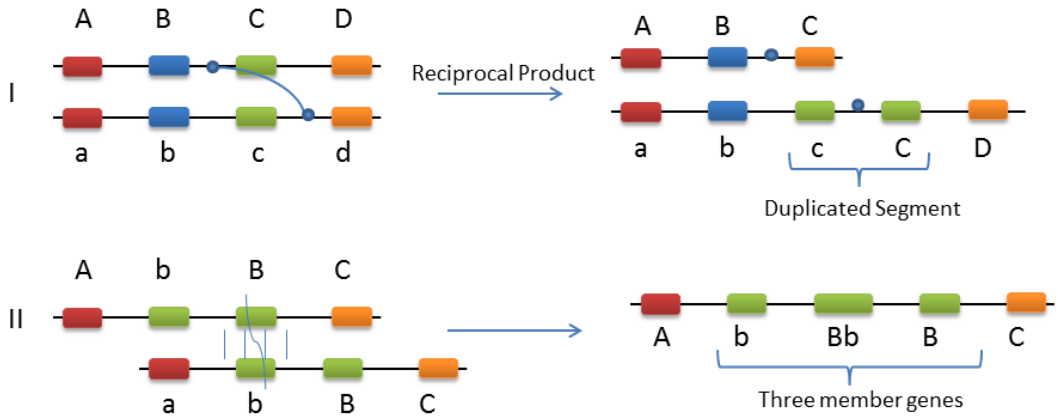


Figure 2.3 Schematic representation of evolution of gene family. The unequal crossing over generates new gene families. I) An initial duplication of a single copy region demonstrates an unequal crossing over event and the two products that are generated. One product is deleted and the other is duplicated for the same region. In this example, the duplicated region contains a second complete copy of a single gene. The blue dot indicates the genetic exchange site. II) Expansion from a two repeat cluster illustrates a second round of unequal crossing over that can occur in a genome that is homozygous for the original duplicated chromosome. In this case, the crossover event has occurred between the two copies of the original gene. The vertical blue line indicates the region of pairing and the cross-over site. Only the duplicated product generated by this event is shown. Over time new gene members can diverge into new gene families.

Gene duplication plays an important role in evolution; it is one source of the raw material from which natural selection produces adaptations in response to environmental conditions (Yamanaka et al. 1998, Hughes 2002, Zhang 2003, Bailey et al. 2004, W. J. Murphy et al. 2005, Larkin et al. 2009). Some notable examples of adaptation via selection on duplicated genes include accelerated expansion of immune-related genes, which were previously known to be evolving in mammalian genome (Barreiro

and Quintana-Murci 2009, Elsik *et al.* 2009). The Toll-like receptors (TLRs) play a key role in the innate immune system, which have been reported in eutherian mammals (Armant and Fenton 2002). The expansion in TLRs family occurred in mammals 300 Mya (Beutler and Rehli 2002), birds 147 Mya (Brownlie and Allan 2011), and in chicken as recently as 65 Mya (Temperley *et al.* 2008), producing TLR2A and TLR2B (Temperley *et al.* 2008). The comparative mammalian genome analysis reported accelerated evolution in certain families, such as cathelicidin in cetartiodactyl, and β-defensins and C-type lyzozymes in ruminants. Moreover, the I interferon (IFN) and interferon tau (*IFNT*) genes have been duplicated in the pig and cattle genomes respectively (Elsik *et al.* 2009).

Orthology and paralogy are both evolutionary concepts that are defined by speciation and duplication events. Orthologous genes are genes that have become distinct copies through a speciation events (Lechner *et al.* 2014). Similarly, copies of genes that arise through duplication events are paralogs (Jensen 2001). In order to detect orthologs, several algorithms, tools such as orthobench (Trachana *et al.* 2011), BLASTO (Zhou and Landweber 2007), OrthoMCL (Li *et al.* 2003), OrthoSelect (Schreiber *et al.* 2009), MSOAR 2.0 (Shi *et al.* 2010), OrthologID (Chiu *et al.* 2006), MetaPhOrs (Pryszcz *et al.* 2010), PHOG (Datta *et al.* 2009), have been implemented. Most of the commonly used tools implement phylogenetic approaches to reconstruct the best evolutionary view of orthologous and paralogous relationships (Trachana *et al.* 2011). The tools with tree reconciliation algorithms are expected to provide fine-grained predictions but are computational very expensive and not free of artefacts.

Depending on the number of genes found in each species, EnsEMBL classifies the genes and differentiates them into *one2one*, *one2many* and *many2many* relationships (Figure 2.4). The *one2one* label indicates that one copy of the gene is present in both species; whereas *one2many* represent occurrences of one gene in one species and its multiple duplications in another species. The *many2many* label denotes the occurrences of multiple duplications within a gene family in both species being compared. The *apparent one2one homologs* were counted in the *one2one* homologs list.

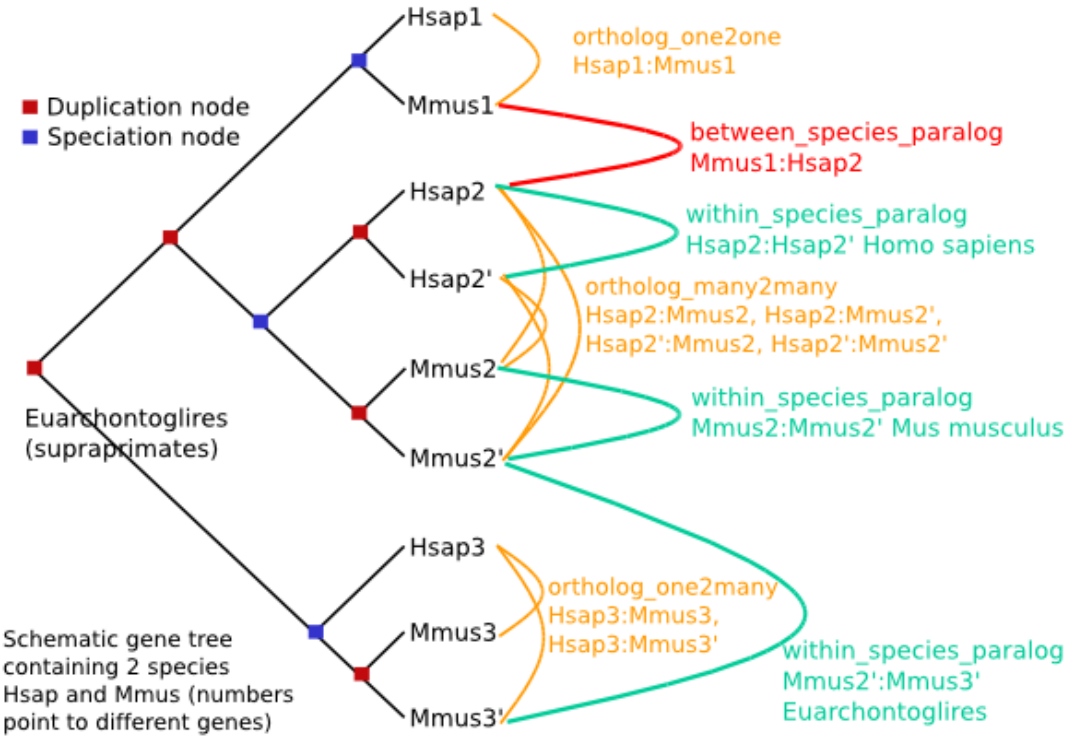


Figure 2.4 Schematic gene tree of homolog relationships between *Homo sapiens* (Hsap) and *Mus musculus* (Mmus) genes. These pairwise relationships between genes can be inferred with EnsEMBL's GeneTree algorithms. The duplication nodes are denoted by red, whereas speciation nodes are blue. Orthologous and paralogous relationships are indicated by coloured lines. The *one2one* relationship indicates the presence of one copy of a gene in both species; whereas *one2many* relationships represent occurrences of a single gene in one of the species and many copies of the same gene with similar function in other species. The *many2 many* denotes the occurrences of multiple genes in both the species for single functions.[7]

1.2.2.2 Gene ontology and enrichment analysis

The several ongoing projects discussed above and availability of many annotated genomes empower the biological science with enormous data, but also cause confusion regarding the annotation, expression, and protein products of genes (Lewis 2005). The Gene Ontology (GO) consortium, therefore, has come into existence to rescue, unify and manage the huge amount of biological information with a certain set of well-defined and universal vocabularies for biological domains (Ashburner et al. 2000, Consortium 2008). These consortiums believe in the fact that certain biological functions are shared

[7] http://www.ensembl.org/info/genome/compara/homology_method.html

amongst eukaryotes and that those functions slowly evolve over time (Ashburner *et al.* 2000). In other words, there is a unified universe of genes and their products that are dispersed across living organism. For example, such unified and important biological processes are DNA replication, transcription, and metabolism, which are functionally conserved across all eukaryotes. In order to systematically manage all this information, three main extensive ontologies have been designed to describe the molecular function, biological processes, and cellular component[8] of genes. The GO consortium keeps all the GO data cross-linked with several genes and protein keyword databases in the public domain, which can be further scrutinised by scientists around the world and thus improve over time[9] (Ashburner et al. 2000, Hill et al. 2008, Consortium 2010, Consortium 2013).

Despite having highly curated and freely available GO data, scientists need some specialised tools and software to capture localised genes and their products with annotation references. GO data is most often accessed with some specialised software developed by the GO consortium, such as AmiGO (Carbon *et al.* 2009), QuickGO (Binns *et al.* 2009), GO browse, etc. Similarly, several other independent pieces of software have been developed by research groups to accomplish their GO analysis research, and such common software and tools are mentioned in table 2.1. In addition, some tools like GOFigure (Khan *et al.* 2003) and Goblet (Groth *et al.* 2004) have been developed to automate the annotation of GO terms (Zhou *et al.* 2005).

[8] http://www.geneontology.org/GO.doc.shtml
[9] http://www.geneontology.org

Table 2.1 List of GO analysis and visualisation tools, open source software, plugins, modules and web servers

Tool name		Remarks
BiNGO	(Maere *et al.* 2005)	Biological Networks Gene Ontology tool (BiNGO) is an open-source Java tool
FatiGO	(Al-Shahrour *et al.* 2004)	Web application, FatiGO, allowing for easy and interactive querying
MAPPFinder	(Doniger *et al.* 2003)	Gene Ontology and GenMAPP to create a global gene-expression profile
GO:TermFinder	(Boyle *et al.* 2004)	Identify GO nodes that annotate a group of genes with a significant p-value
GOStats	(Beißbarth and Speed 2004)	Find statistically overrepresented Gene Ontologies within a group of genes
GOTree Machine (GOTM)	(Zhang *et al.* 2004)	Web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies
AmiGO	(Carbon *et al.* 2009)	Online access to GO consortium database
GOEAST	(Zheng and Wang 2008)	Web-based software toolkit for Gene Ontology enrichment analysis
ClueGO	(Bindea *et al.* 2009)	Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks
DAVID	(Dennis Jr *et al.* 2003)	Database for annotation, visualisation, and integrated discovery
CLENCH	(Shah and Fedoroff 2004)	Calculate Cluster ENriCHment using the Gene Ontology
EasyGO	(Zhou and Su 2007)	Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species

Several tools and web server have been developed to recognise genes and their product, which invariably contributes to a better understanding of complex biological processes. Each tools has several advantages over others. For example, the Biological Networks

Gene Ontology tool (BiNGO)(Maere *et al.* 2005) is an open-source Java tool that is easy to use and provides interactive cytoscape[10] visualisation interface, whereas Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis Jr *et al.* 2003) is a web based server with cross connectivity with multiple databases, Ids conversion, and pathway analysis features. The Perl module GO:TermFinder has also have been developed to analyse GO term with significant P-values (Boyle *et al.* 2004). Similarly, ClueGO also come into existence to decipher functionally grouped gene ontology and pathway annotation networks (Bindea *et al.* 2009). However, scientifically readable biological information about molecular systems is not only dependent on the software type, but also dependent upon the quality of the database being used. Therefore, GeneGO MetaCore was produced, which provides a highly curated database with interactive gene and protein analysis via a visualisation interfaces.

1.2.3 Transposable elements (TEs)

Repetitive DNA, DNA sequence with a high number of copies, is found in all prokaryotes and eukaryotes, and it makes up a significant fraction of the entire genome of most organisms (Tautz and Renz 1984, Lupski and Weinstock 1992, van Belkum *et al.* 1998, Jurka *et al.* 2005). These significant fractions of DNA repeats in genomes are of two types, tandem repeats and interspersed repeats. An array or copies of adjacent motif DNA sequences are called tandem repeats, whereas interspersed repeats are dispersed throughout the genome as a single unit flanked by unique sequence. The interspersed repeats generally originate by a process of transposition, which is a "jumping" movement of DNA from one location to another in a genome, albeit with low frequency. Transposition can occur either directly by a cut-and-paste mechanism (transposons) or indirectly through an RNA intermediate (retrotransposons), such as short interspersed repeat elements (SINEs), long interspersed repeat elements (LINEs), and retrovirus-like elements with long-terminal repeats (LTRs) (Munoz-Lopez and García-Pérez 2010, Levin and Moran 2011). The segments of DNA with this unique ability to move are called TEs, also known as transposons or "jumping genes" (Figure 2.5). These mobile elements were first discovered by maize geneticist Barbara McClintock, and she hypothesised that they play a regulatory role as they can move to different chromosomes. She also posited that they can contribute to the creation of new

[10] http://www.cytoscape.org/

genes and determine which genes are turned on and when this activation takes place (McClintock 1950, McClintock 1965). While, this ground-breaking finding was largely dismissed by the scientific community at that period, Roy Britten and Eric Davidson supported it and further speculated that these mobile elements not only play a role in gene expression regulation, but also generate different cell types and biological structures (Britten and Davidson 1969). Later, it was shown that TEs can inactivate any gene by inserting and thereby interrupting the coding part of its sequence. For example, insertion of an Alu retroelement into the exon of the CMP gene disrupted the normal open reading frame, which resulted in a lack of *N*-glycolyl neuraminic acid (Neu5Gc) on a surface of human cell membranes (Chou *et al.* 1998, Irie *et al.* 1998). Insertional inactivation of genes is useful for isolating mutants defective in specific functions and for mapping genes (Nowacki *et al.* 2009). Alternatively, TEs can also activate adjacent genes by altering the promoter or transcriptional activator to the gene. A study of *Pseudomonas cepacia* showed that the insertion of certain TEs in the upstream region of a poorly expressed gene can increases its expression by more than 30-fold (Scordilis *et al.* 1987).

Moreover, TEs were formerly thought to be found only in a few species, but we now know that TEs (both active and inactive) constitute a large amount of the DNA in many higher eukaryotes, 40% in human (Smit 1999), 27% in cattle (Elsik *et al.* 2009), and 37% in mouse (Chinwalla *et al.* 2002). Moreover, fish and bird genomes consist of 10% TEs (Abrusán *et al.* 2008), whereas the genome of *C. elegans* is having 12% TEs (Consortium 1998, Stein et al. 2003). However, in some plants, such as maize, the TE percentage exceeds 80% of the entire genome (SanMiguel *et al.* 1996). These TEs are omnipresent in the biosphere and are self-trained to efficiently propagate themselves. Moreover, the impact of TEs in genomic instability and reconfiguration of gene expression networks is costly, as they may cause several diseases (Kazazian Jr 1998, Kazazian 2004, Reilly *et al.* 2013). Approximately 0.27% of human diseases are attributed to retrotransposable elements (Callinan and Batzer 2006, Fedoroff 2012).

1.2.3.1 Impact of TE in genome evolution

Transposable elements (TEs) and their fingerprints are found throughout genomes, ranging from the coarsest features of genomic landscapes to gene dense regions. These elements are not just junk DNA or mutagens, but instead an "operating system" or

fertile ground for genome evolution (Biémont and Vieira 2006, Fedoroff 2012). In addition, as predicted by Barbara McClintock and others, TEs play a vital role in genome evolutions by controlling or interfering with gene structure, function, regulation, and expression (Tautz and Renz 1984, Lupski and Weinstock 1992, van Belkum *et al.* 1998, Jurka *et al.* 2005, Fedoroff 2012, Chang *et al.* 2013). Such alterations are being made by the insertion of transposons or retrotransposons into the functional regions of genes (Medstrand *et al.* 2005). This can either damage or alter the gene functions. For example, insertion of Alu repeats can obstruct the chromosomal pairing which results in unequal crossover, mediating further duplications (Chandley 1989).

The role of TEs in the evolution of various amniote genomes, such as those of human (Mills *et al.* 2007), great apes (Warnefors *et al.* 2010), cow (Bovine Genome *et al.* 2009), mouse (Nellåker *et al.* 2012, Rebollo *et al.* 2012), reptiles, and birds (Kordis 2010) have been studied extensively. These studies show the profound impact of TEs on structure, function, and genome evolutions by interfering with respective genomes. It has been shown that some of the TEs that were found more active in non-mammalian vertebrates during Silurian period are the source of ultra-conserved elements within mammalian genomes, with some exceptions (Sela *et al.* 2010). In addition, the vertebrates exhibit a high abundance of TEs in intrinsic sequences and introns in comparisons to invertebrates (Sela *et al.* 2010)

1.2.3.2 Activation and deactivation over the period of genome evolution

Many TEs were reported to be inactive or active at specific periods in evolutionary time. As reported in the cattle genome, the non-LTR LINE retrotransposon were found lacking an open reading frame (ORF) suggesting their inactive nature (Malik and Eickbush 1998). On the other hand, a few of the BovB repeats were found containing intact ORF suggesting they are actively expanding and evolving in the cattle genome (Elsik *et al.* 2009). The older repeats are believed to be destroyed by insertion of new and highly active repeats. The bovine genome consortium reported a lower number of ancestral repeat families in cattle-specific EBRs, whereas there are significantly more repeats in ancestral EBRs (Elsik *et al.* 2009). These findings suggest that either repeat elements were more recently inserted into regions lacking ancient repeats or that older repeats were destroyed by such insertions. Another evolutionary study employing a genome-wide defragmentation approach has revealed the early activity of some MER2

transposons and the relatively recent activity of MER1 transposons during the evolution of primate lineages (Giordano *et al.* 2007). These bouts of activation and inactivation contribute to evolution of the genome, providing raw material to natural selection.

1.2.3.3 Retrotransposed genes

The evolutionary dynamics of genomes are influenced by various genomic processes that give rise to novel sequences; one such process is retrotransposition. In this process the mRNA transcript is spontaneously reverse-transcribed and reintegrated into the genome (Boeke *et al.* 1985). The large-scale retrotransposition of mRNAs into mammalian genomes has been revealed by the detection of thousands of obvious retrotransposition in mouse, rat, and human (Zhang *et al.* 2002, Zhang and Gerstein 2003, Zhang *et al.* 2003). The retropseudogenes (processed pseudogenes) mostly lack promoters and introns and possess relics of the poly-(A) tail at their 3' tail (Harrison and Gerstein 2002). Retropseudogenes also include short direct repeats flanking their sequences (Betrán *et al.* 2002), frequent a truncation at the 5' ends, and at a genomic location different from that of the parent gene (Zhang *et al.* 2002, Zhang *et al.* 2003). These are the hallmark characteristics of retrotransposition, which often deteriorate or inactivate gene sequence copies. Henceforth, retropseudogenes are generally considered non-functional and "dead on arrival" from the moment they reintegrate into the genome (Harrison and Gerstein 2002). Contrary to this, a few events have been reported in which insertions may have contributed exons to existing genes (Baertsch *et al.* 2008). A growing number of studies have been carried out on spontaneous substitutions, deletions, and insertions in retropseudogenes (Ophir and Graur 1997). It has been discovered in human that these processes are mainly mediated by reverse-transcriptase (Mathias *et al.* 1991) and endonuclease (Feng *et al.* 1996) functions of the LINE-1 ORF2 protein. These processes work in assistance with the ORF1 protein, which binds RNA (Hohjoh and Singer 1997) and functions as a chaperone (Martin and Bushman 2001). Additionally, LINE-1 mobilises other transcripts including Alu (Dewannieux *et al.* 2003), SINE-VNTR-Alu (Hancks *et al.* 2011) and processed pseudogenes (Esnault *et al.* 2000). The processed pseudogene formation through reverse-transcriptase varies among species, and mainly depends on the retroelement content of the genome. Many genes with novel function may have originated via the retrotransposition process, as few of the genes in mammalian genomes were reported to bear the characteristics of

retrosequences (Brosius 1999, Emerson *et al.* 2004). Some of the retrotransposed genes have been annotated in human and mouse and are known to be expressed in testis, which may be a driving force for rapid testis evolution in primates (Emerson *et al.* 2004, Marques *et al.* 2005).

1.2.3.4 Role of TE in genome instability and rearrangements

There have been reports describing the association between TEs and chromosomal breakpoints in several plants and animals (Nevers and Saedler 1977, Gray 2000, Lönnig and Saedler 2002, Bennetzen *et al.* 2005); however, this was first studied by McClintock to better understand the mechanisms of chromosome breakage and fusion in maize. In her research, she identified a locus on chromosome 9, which is called "*Ds*" or "dissociation" locus and has repeatedly broken over time. Later, she discover the locus *Activator*, which initiates its own transposition and can activate chromosomal breakage (McClintock 1947). Similarly, Collins and Rubin (1983) first reported an aggressive case of chromosomal rearrangements in *Drosophila*, in which a 10Kb fold back TE with a complex inverted shape contributed to rearrangements (Collins and Rubin 1983). The association between TEs and chromosomal breakage has been verified by several groups of scientist in various organisms, such as *Drosophila melanogaster* (Lim and Simmons 1994, Ladeveze *et al.* 1998), yeast (Roeder and Fink 1980), cattle (Elsik *et al.* 2009), gibbon (Girirajan *et al.* 2009) and other mammals (Schibler *et al.* 2006). Moreover there is ample research that confirms the significant enrichment of TE in chromosomal breakpoints beyond that expected by random chance, suggesting a probable role of TEs in chromosomal rearrangements (Longo *et al.* 2009, Penny 2012).

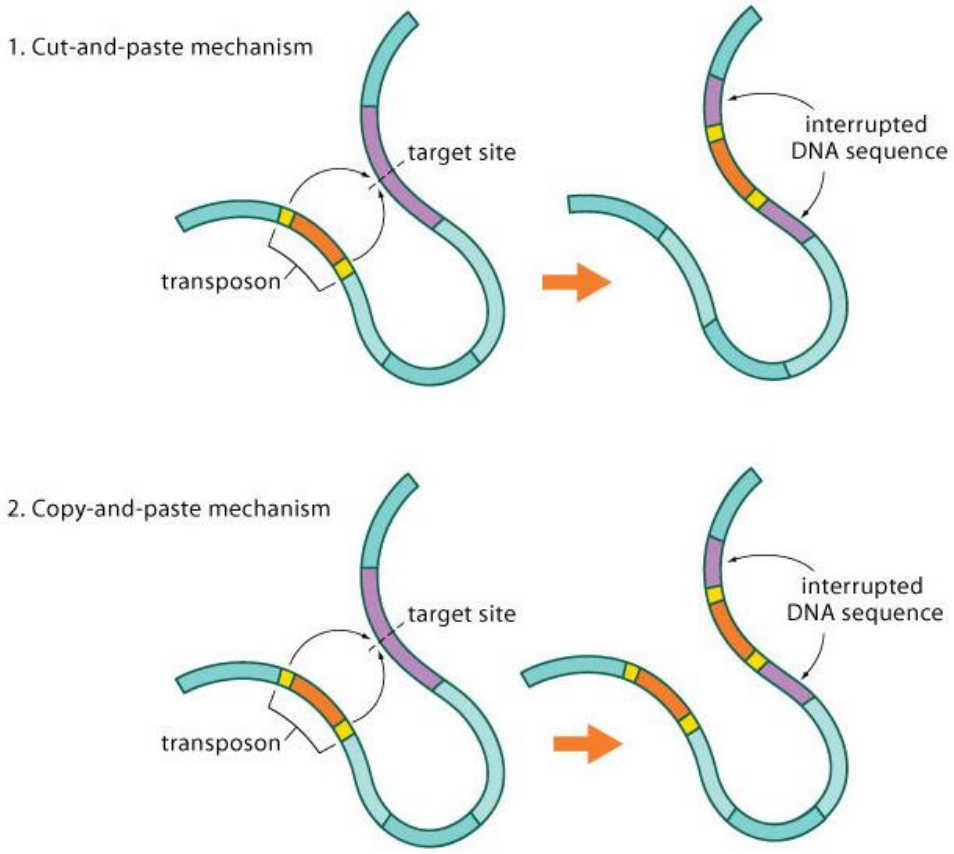


Figure 2.5 A schematic representation of transposable elements (TEs) movements A TE (shown in orange) is inserted via a cut-and-paste mechanism, disrupting the existing target DNA sequence. The second TE mechanism makes a copy of a transposon and inserts into another location of the genome and interrupting DNA sequences[11]

1.3 GENOME MAPPING

To quickly navigate the features of interest and detect their relative positions in the genome, genome maps have been developed. Genome mapping, also called gene mapping, is the assignment of DNA fragments to specific chromosome locations and the determination of the relative distances between genes on those chromosomes (Sturtevant 1913). The gene for eye-colour was first located by Thomas Hunt Morgan on the X chromosome of fruit fly. Shortly thereafter, E.B. Wilson identified sex-linked genes underlying colour blindness and haemophilia in humans, similar to the many X-

[11] https://www.broadinstitute.org/education/glossary/transposable-elements

linked factors that were being described by the Morgan group in flies. Later, Donis-Keller *et al.* (1987) generated the first comprehensive genetic linkage map of human chromosomes using restriction fragment length polymorphism (RFLP) techniques. This genetic map was based on 400 RFLPs, which are variations in DNA sequence observed by digesting DNA with restriction enzymes (Donis-Keller *et al.* 1987). These types of maps organise valuable annotations, which assists in further understanding of genomes. Additonally, genetic variation can be used to locate genes responsible for diseases. These genetic variants can either occur in genes (coding), regulatory regions or non-coding (and non-regulatory) sequences. These genetic variants that are identified and mapped throughout genomes are called markers (Brown 2002). Henceforth, the accuracy of genome maps entirely depends on the quality of the markers detected and the methods applied.

There are two distinct types of molecular maps— physical and genetic-linkage—that can be derived for each chromosome in the genome. These maps provide the likely order of markers along a chromosome. The physical maps can also be divided into three general types: Chromosomal (also known as cytogenetic maps), Radiation hybrid (RH) maps, and Sequence maps. Figure 2.6 not only illustrates and distinguishes the methods that are used to create maps, but also the metrics used for measuring distances within them. Linkage maps, also called recombination maps, are constructed from loci that occur in two or more heritable forms, or alleles. Therefore, monomorphic loci, those with only a single allele, cannot be mapped using this technique. On the other hand, chromosomal map use size and banding pattern inferred from direct cytogenetic analysis or by linkage and physical positions that are associated with observable chromosomal banding patterns. This is the most direct mapping approach. The resolution of chromosomal maps is low compared to linkage or physical approaches and therefore it is less frequently used. Physical maps use the direct analysis of DNA, in which physical distances between and within loci is measured in basepairs (bp), kilobasepairs (kb) or megabasepairs (mb). There are several physical mapping techniques available. One such technique is fluorescent *in situ* hybridization, which directly observes the relative position of markers in the genome (Iacia and Pinto-Maglio 2013). Other methods are also useful, but use less direct approaches to map genetic markers. However, almost physical mapping techniques use a common approach to isolate a portion/gene of interest from the genome and map relevant markers. Out of all three aforementioned

mapping techniques, only the basepair distances measured by physical maps provide an accurate description of the actual length of DNA that separates loci from each other. Each of these types of maps provides the same information regarding chromosomal assignment and the order of loci, but the relative distance between the loci generally varies (see more about all map types in subsequent sub-sections).

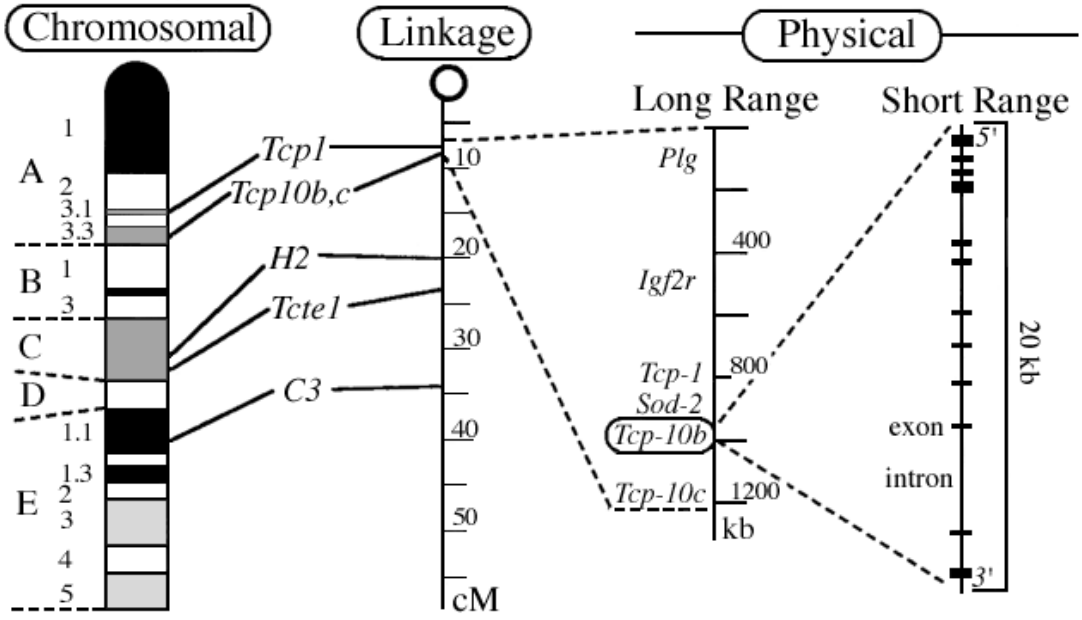


Figure 2.6 Comparative image of physical and genetic-linkage maps. The relative sizes of molecular maps -- linkage, chromosomal, and physical are shown for a 1,200 Kb genomic interval around the Tcp10b locus on mouse chromosome 17 (Barlow *et al.* 1991). The lines connect the relative positions of the same loci as mapped in linkage, physical and chromosomal maps.

The recent rapid advancement in various genome technologies has allowed the exploration and elucidation of the underlying molecular mechanisms of genome evolution. This has changed the way molecular biology research is conducted. The Human Genome Project (HGP) (E. S. Lander et al. 2001) had a profound impact on biomedical research and revolutionised a wide spectrum of biological research and clinical medicine programs; it also provoked the generation of genome sequences from other mammals. Many genome projects have leveraged new technology and produced an unprecedented wealth of genomic data for comparative analysis (Haussler *et al.* 2009).

The National Institute of Health (NIH) has funded several projects to expand the current understanding of molecular and evolutionary mechanisms by sequencing more mammalian genomes. The Broad Institute is currently sequencing ~150 mammal species, while other centres are generating an additional ~150 mammalian genomes. For example, the National Human Genome Research Institute (NHGRI), a large-scale sequencing centre, has sequenced the genomes of 24 species to low (~2x) sequence coverage[12]. Similarly, the 1000 Genomes Project is the first project to sequence the genomes of a large number of humans, in order to provide a comprehensive resource of human genetic variation (Siva 2008). The Genome 10K Community of Scientists (G10KCOS) have a long-term goal of generating and assembling ~10,000 vertebrate genomes of fishes, mammals, amphibians, reptiles and birds (Haussler *et al.* 2009). These sequencing projects will help us to understand the genetic basis of adaptive evolutionary changes within related species and also understanding the evolutionary mechanisms behind adaptation. G10K will enable the study of genetics in threatened and endangered species, disease risk factors within non-model organisms and help to reconstruct ancestral genomes for different clades. Additionally, it will assist in predicting the response of species to climate change, pollution, emerging diseases and invasive competitors (Bell *et al.* 2004, Kohn *et al.* 2006).

By comparing all annotated genomes, scientist can infer the order and relative positions of the markers. Maps annotated with marker information are an invaluable source for comparative genome mapping, which uses genome maps of various phylogenetically related species to reveal conservation of genes and synteny relationship amongst them. These map-based comparative techniques provide an insight into genome evolution and also assist in annotating the gene's location in new target species. These maps are also an invaluable asset for genome sequencing (Table 2.2). Genome maps are frequently used to guide and validate the multi-step procedure of genome assembly. This multi-step procedure of genome assembly first requires the cloning of DNA fragments, that are then sequenced and computationally assembled based on the markers the sequence contain. In order to obtain full coverage of genomes, I need to use fully-annotated physical and genetic maps (Beyer *et al.* 2007). High-resolution physical maps of several

[12] http://www.genomesonline.org/cgi-bin/GOLD/index.cgi

species' chromosomes empowers comparative genomics discovery and are indispensable for sequence assembly precision (Lewin *et al.* 2009).

Table 2.2 Physical and linkage map and genome assemblies. Physical and linkage maps have been used as anchors for mammalian genome assemblies in various whole genome sequencing projects (Lewin *et al.* 2009).

Species*	Genome Size(Gbp)	Sequence mapped	Type of physical maps	Number of markers
Human	2.8	99%	Fingerprint map	25,241
			Fluorescent in situ hybridisation map / Radiation hybrid map / Linkage map	924
Macaque	3.1	92.2	Radiation hybrid map	802
			Linkage map	241
Mouse	2.6	97.6	Radiation hybrid map	11,109
			Linkage map	7,377
Cattle	2.8	90.3	Radiation hybrid map / Linkage map	1,680
			Radiation hybrid map	3,484

*Physical and linkage maps have been used to anchor sequences to chromosomes for mammalian genome assemblies in various genome sequencing projects (Lewin *et al.* 2009).

1.3.1 Genetic linkage mapping

Mendel's conclusions were drawn from a series of experiments on *Pisum sativum*. His "law of independent assortment" states that factors (later identified as genes) are transmitted from parents to offspring independent of one another (Mendel 1865). However, not all genes are inherited independently. Thomas H. Morgan postulated that

linked genes are present in a linear order along a chromosome and depending upon the distance, during first meiotic prophase, a variable amount of reciprocal exchanges may occur between genes; he later confirmed this postulation in *Drosophila melanogaster* (Morgan 1910). Genes that are present on the same chromosome were described as "linked" genes by Bateson and Punnett (Bateson and Punnett 1911). On the basis of recombination frequency, Sturtevant (1913a, 1913b) published the first linkage map, placing three genes on the X chromosome of *Droshophila melanogaster*. In the 20th century, scientists were able to construct genome linkage maps using the log score technique (Haldane and Smith 1947), polymerase chain reaction (Mullis 1994, Mullis *et al.* 1995), Restriction Fragment Length Polymorphisms (RFLPs) (Botstein *et al.* 1980) and many other techniques. The drawback of genetic linkage mapping is its inability to accurately fine map closely located linked genes (i.e., genes with the lowest recombination frequency) and also the very coarse resolution of most genetic linkage maps. Despite this, linkage maps were extensively used for mapping marker intervals associated with phenotypic, disease, and economically important traits (Heyen *et al.* 1999).

Because of their low resolution, genetic maps do not make a strong basis for the sequencing phase of eukaryotic genome projects. However, due to the short life cycle of microorganisms, recombination events can be obtained in ample amounts, resulting in a highly detailed genetic map where the markers are a few kilobase (kb) apart, and thus microorganism linkage maps assist in genome sequencing and assembly. Besides being low resolution in eukaryotes, genetic maps also limited by their accuracy, as seen in comparative analysis of *S.cerevisiae* genetic map to the actual positions of markers as shown by DNA sequencing. Multiple markers, including glk1 and cha1, mapped to different locations in the genetic and linkage maps (Oliver *et al.* 1992, Dujon *et al.* 1994). In order to address such problems, a plethora of physical mapping techniques have been developed.

1.3.2 Physical mapping

A physical map shows the physical location of markers on the chromosomes. The most common methods used in physical mapping are fluorescent *in situ* hybridisation (FISH) mapping, radiation hybrid (RH) mapping, bacterial artificial chromosome fingerprinting and DNA sequencing. RH mapping and FISH mapping were widely used techniques for

physical mapping, but each of them has its own benefits and limitations. RH mapping makes use of RH panels and statistical methods to determine the order of and distances between DNA markers on chromosomes (Walter and Goodfellow 1993). RH mapping techniques have become a general way to construct high-resolution, contiguous physical maps for several species, such as human, rat, mouse, cat and pig (Murphy *et al.* 2000, Chowdhary *et al.* 2003, Kwitek *et al.* 2004, Wind *et al.* 2005). FISH mapping utilises hybridisation of fluorescent-labeled DNA probes to find the order of markers on chromosomes. Lorenzi *et al.* (2010) corrected the gene location in Btau_4.0 assembly using FISH (De Lorenzi *et al.* 2010). However, most FISH techniques generally provide insufficient resolution to map closely located markers.

Schwartz *et al.* (1990) developed a new method, optical mapping (OM), to construct an ordered, high-resolution restriction map from DNA. The unique feature of OM is that it preserves the order of DNA fragments. In this method the cells are lysed to retrieve genomic DNA and the DNA is randomly sheared to produce a "library" of large genomic molecules for optical mapping. Single genomic DNA molecules are placed onto a microfluidic device and digested by restriction enzymes. Later, the DNA fragments are stained with intercalating dye and are visualised by fluorescence microscopy. The fragment sizes are measured by their fluorescence intensity. Finally, all optical maps are combined to produce a consensus optical genomic map. This technique has been mostly used for the construction of whole-genome restriction maps of several eukaryotes (Schwartz *et al.* 1993, Lin *et al.* 1999). The main advantage of optical mapping includes its high throughput and resolution, safety and low cost.

1.4 SEQUENCING APPROACHES

Maxam and Gilbert (1973) developed the first method to determine DNA sequences and reported the sequence of 24 base pairs using a method known as "wandering-spot analysis" (Gilbert and Maxam 1973). The Maxam and Gilbert sequencing protocol is based on preferential, base-specific methylation of nucleotides, followed by chemical cleavage to generate a nested set of end-labelled derivatives at the final stage (Maxam and Gilbert 1977). The Maxam and Gilbert sequencing approach has a major disadvantage, because it dependents on the use of radioactive reagents. In the meantime, Frederick Sanger and co-workers (1977) develop a new method, known as "dideoxy

sequencing" or the "chain termination method". The principle of this method was based on the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators. This reaction results in all fragments ending in one of the four fluorescent dye-labeled terminators. Later, these fragments are separated by electrophoresis, in which the fluorescence is detected by laser excitation and a CCD camera (Figure 2.7) (F. Sanger et al. 1977). Later, this technique became the "workhorse" for genome sequencing because of its practicality. Technological advancements since the 1970s have made the Sanger method not commonly used for high-throughput sequencing, but still widely used for small, low throughput sequencing (Hert et al. 2008). Mostly, this approach is widely used for sequencing projects targeting a small region in a large number of individuals. The new sequencing technologies that have replaced this method are based on the same principles (Gilbert and Maxam 1973). Automated sequencing has been developed so that more DNA can be sequenced in a shorter period of time. Despite dramatic changes in sequencing approaches, the primary data production for most genome sequencing since the Human Genome Project (HGP) has relied on the same type of capillary sequencing instruments as the HGP used. However, this situation is rapidly changing due to the invention and commercial introduction of several revolutionary approaches for DNA sequencing, the so-called "next-generation sequencing technologies".

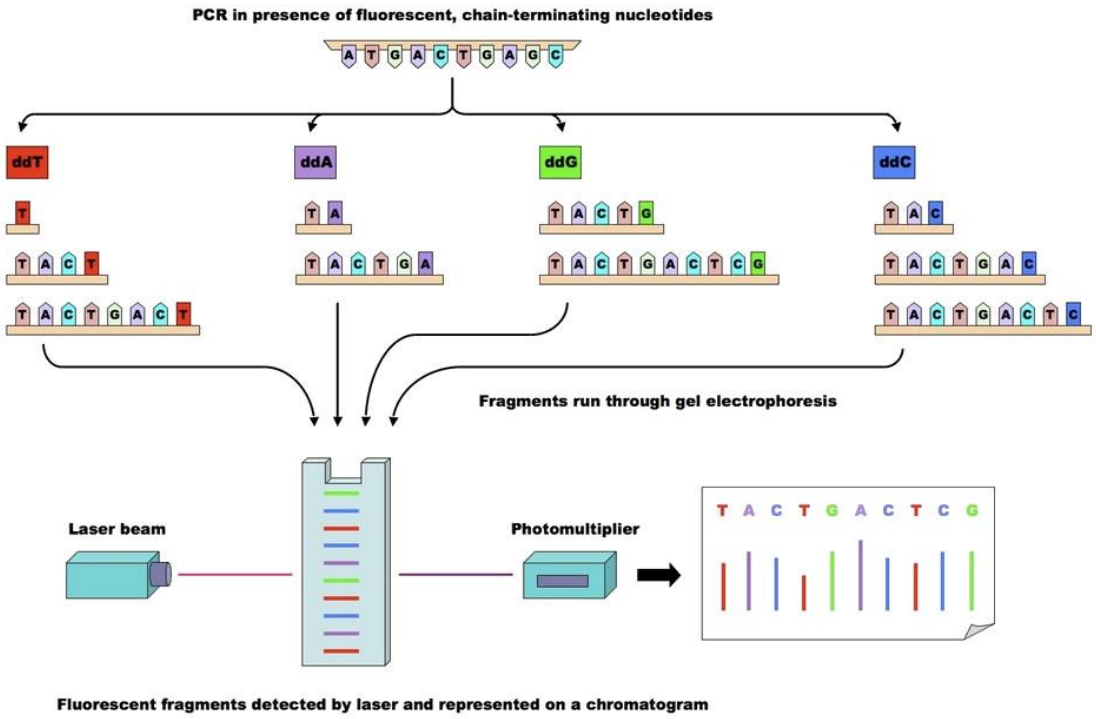


Figure 2.7 DNA sequencing via the Sanger method[13]

The sequencing machines produce large amount of sequenced base pairs or 'raw' sequence. These raw sequences are jumbled together, like the pieces of a jigsaw puzzle. Each nucleotide sequences is called a "read or short DNA sequences", which were used later to reconstruct the original sequence (Church and Gilbert 1984). All available genome sequencing platforms usually generate sequence data in the form of many independent reads. These reads are later assembled together using certain computational tools to form a complete sequence using pair-wise overlaps between the reads and other sophisticated assembly strategies[14]. For Sanger sequencing method these reads are routinely around 800-1000 base pairs long (Frederick Sanger et al. 1977). However, the next-generation sequencing methods produce comparatively much larger quantities of sequence, but in the form of much smaller reads. Illumina is the most commonly used platform, and here the read length is usually 100 to 150 base pair reads[15]. However, the lower-throughput platform can manage to produce read lengths of 400 base pairs[16] .

[13] http://www.vce.bioninja.com.au
[14] http://bioinformaticsonline.com/pages/view/22807/software-packages-for-next-gen-sequence-analysis
[15] http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote-nrc-exome-read-length.pdf
[16] http://www.hindawi.com/journals/bmri/2012/251364/tab1/

In 1988, Lander and Waterman first described the theoretical redundancy of fold-coverage (c) of a shotgun sequencing experiment as LN/G, where L is the read length, N is the number of reads and G is the haploid genome length (Lander and Waterman 1988). However, the empirical average "depth-of-coverage" of an assembly were calculated by LN/A, where N is the number of reads, L is read length and A represent assembly size (Lander and Waterman 1988, Sims et al. 2014). Therefore, "depth-of-coverage" or "fold-coverage" terms are not the same and might be different because of sequencing error and unclonable or unmappable regions of the genome. The term depth may also be used to describe how much of the complexity in a sequencing library has been sampled. In real-world sequencing approaches the read can contain sequence errors. Those errors are mostly indistinguishable from a sequence variant. Such sequencing errors can be identified or can be overcome by increasing the number of sequencing reads. Increasing the depth of coverage can resolve some errors but it does not cure all sequencing ills.

Demands for low cost sequencing have compelled the development of high-throughput sequencing technologies, which can produce millions of sequence reads at once. Several new methods have been introduced to decode the order of nucleotides in a genome. The three main platforms for massively parallel DNA sequencing read production are the following: i) Roche/454 FLX (Margulies *et al.* 2005), which uses a parallelised version of pyrosequencing, also known as the "single-nucleotide addition" (SNA) method (Hyman 1988); ii) Illumina/Solexa Genome Analyzer, which applies a reversible dye-terminator-based method (Bentley 2006, Mardis 2008); and iii) Applied Biosystems SOLiDTM System, which relies on sequencing with a ligation approach (Mardis 2008). In addition to that, two other massively parallel systems were recently announced: the Helicos Heliscope[17] and Pacific Biosciences Single Molecule Real Time[18] (SMRT). The important feature of both the Helicos and Pacific Biosystems instruments is that they do not require any amplification of DNA fragments prior to sequencing, as it as required by other sequencing approaches. Recently introduced nanopore sequencing methods, also known as "third generation sequencing" methods, use an approach that involves drawing individual strands of DNA through tiny nanoscopic holes, or pores (Clarke *et al.* 2009). This advance has the potential to sequence a mammalian genome within an

[17] www.helicosbio.com
[18] http://www.pacificbiosciences.com

hour with quality scores of Q40 (99.99% accuracy), read length of 1000 bp, coverage greater than 95% and, more importantly, at a total cost of less than $1,000. These technologies will lead genomics to an exciting stage where there will be a tremendous amount of data to allow the unlocking of biological questions.

Next-generation sequencers require long run times of between 8 hours to 10 days, depending upon the read type (single end or paired ends) and platform being used. The yield of sequence reads and total bases per instrument run is significantly higher than the 96 reads of up to 750 bp produced by a single capillary sequencer run, and can vary from several hundred thousand reads (Roche/454) to tens of millions of reads (Illumina and Applied Biosystems SOLiD) (Mardis 2008). The advantages of Roche/454 method are the following: first, it does not rely on cloning template DNA, and second, it does not skip uncloneable segments, such as heterochromatin, during sequencing. However, the major drawback to the pyrosequencing approach is the incomplete extension of homopolymers, or simple repeats of the same nucleotide (e.g., AAAAAAA). Each read is only about 250-400 base pairs long at this time, making it difficult to differentiate between repeated regions longer than this length. To compare, paired-end methods in Illumina sequencers enable paired-end sequencing of up to 2 x 100 bp for fragments ranging from 250 bp to 40 kb. In addition to that, pyrosequencing is also improving quickly, and new machines can generate 400-base pair sequence reads. Thus far, chromosomes cannot be sequenced by a single read; all sequencing methods produce a series of segments of DNA code, referred to as 'reads'. After sequencing occurs, genomes need to be reconstructed from millions of short reads, or "assembled". In order to reconstruct the original genome sequence from millions of reads, specialised computer programs called "assemblers" are used.

New techniques and algorithms for whole-genome sequencing (WGS) have made it possible to sequence a genome in a short period of time, but assembly of these genomic sequences is still a painstaking task. Genome maps, such as RH maps, linkage maps, FISH maps and optical maps, have become very important and necessary resources for the assembly of genome sequence and their validation. These maps provide markers for anchoring and guiding the placement and orientation of genomic contigs or scaffolds onto the chromosomes (E.S. Lander et al. 2001, Warren et al. 2007, Miller et al. 2010).

With the availability of genome sequences and comparative genomics modules, it is now possible to explore genomes and compare them at high resolution.

1.5 ASSEMBLY APPROACHES

Because of dropping costs and increases in sequencing efficiency, the whole-genome sequencing for 10,000 vertebrate species was recently proposed (Genome 2009). This genomic information will help us to understand genome evolution and gene structures of vertebrate species. However, after genome sequencing the most cumbersome task is to assemble millions of sequence reads, which are short in length and potentially contain sequencing errors (Metzker 2009, Alkan *et al.* 2010, Zhang *et al.* 2011). The paired-end (PE) sequencing method is used to generate reads from both ends can, and, to some extent, compensate for read length (Cahill *et al.* 2010). whereas the single molecule, real-time (SMRT) technology produces longer reads but has higher error rates (Cahill *et al.* 2010, Schadt *et al.* 2010).

Genome assembly, a bioinformatics technique to stitch sequence data into contigs, scaffolds and chromosomes, needs highly efficient algorithms to correctly merge the millions of reads within a limited period of time. In order to develop competitive software, programmers predominantly used non-primitive data structures that can be categorised into two types: i) string-based models and ii) graph-based models. Initially, contigs, a set of overlapping DNA segments derived from a single genetic source, were built using overlap-layout-consensus strategies (Myers 1995). The high-quality assemblies of human (E. S. Lander et al. 2001, Li et al. 2010) and mouse (Chinwalla *et al.* 2002) have been constructed with GigAssembler (Kent and Haussler 2001), Celera, ARACHNE (Batzoglou *et al.* 2002), and Phusion (Mullikin and Ning 2003) software. However, these programs compute a quadratic number of alignments and consequently are not efficient enough to handle the volume of sequences produced by next generation sequencing technologies, stimulating the development of a new generation of assembly software.

Several algorithms have been developed to correctly handle the genomic jigsaw puzzle, and assemble genome reads in correct order. Greedy-extension algorithm of string based model software such as Quality-value guided *de novo* Short Read Assembler

(QSRA) (Dohm *et al.* 2007, Bryant *et al.* 2009), SHARCGS (Jeck *et al.* 2007), and SSAKE (Warren *et al.* 2006) are efficient *de novo* assemblers for prokaryotic genomes (Bryant *et al.* 2009) because of less repetitive nature of their genomes than those of mammals. The graph based model and software are designed ABySS (Simpson *et al.* 2009), Velvet (Zerbino and Birney 2008, Zerbino *et al.* 2009), SOAPdenovo (Li *et al.* 2008, Li *et al.* 2010) with implementation of thread parallelization to reduce the time cost, and EULER-USR to cope up with the large genomes and exploit pair end (PE) sequencing information to reduce gaps from assembled contigs.

Some other genome-assembly software packages including Arachne (Batzoglou *et al.* 2002), Atlas (Havlak *et al.* 2004), Ray (Boisvert *et al.* 2010), Celera Assembler (Myers 2005), CAP3 (Huang and Madan 1999), Euler (Pevzner *et al.* 2001), Phrap (Bastide and McCombie 2007), RePS (Wang *et al.* 2002), Edena (Hernandez *et al.* 2008) implement OLC (Overlap-Layout-Consensus) approach that requires overlaps to be scored between all possible pairs of reads. This is computationally intensive and therefore is not widely used, whereas Taipan (Schmidt *et al.* 2009) uses a hybrid of string and graph based algorithmic approaches for assembly with a shorter period of run time.

Out of the above mentioned algorithms, de Bruijn graph and Eulerian path approaches (Pevzner *et al.* 2001) are predominantly used methods in current scenarios for assembly, but they are still not fully capable to correctly assemble complex and repetitive parts of genomes. In order to improve the computational methods of genome assembly, and decide the best algorithm and software to them, a collaborative effort have been taken by ASSEMBLATHON[19] to reassemble, compare and verify the genome assemblies with various assembly programmes. Comparative studies of de novo assemblies of individuals show that, assemblies were 16.2% shorter than the original genome sequence. It is speculated that *de novo* assembly algorithms collapse identical repeats(Green 2002), resulting into reduced or lost genomic complexity. The limitations of *de novo* assemblies were also confirmed by looking at missing 420.2 megabase pairs of common repeats and 99.1% of validated duplicated sequences from the assembled genome (Alkan et al. 2010, Hubisz et al. 2011, Zhang et al. 2011, Keith R Bradnam et al. 2013). The large size and high repetitive content of mammalian genome sequence still

[19] http://assemblathon.org/

requires new genome assemblers with highly memory-efficiency, reduced time cost, smart with repetitive and small sequences. The second collaborative meeting of ASSEMBLATHON uses varieties of sequenced data of three vertebrate species (a bird, a fish, and snake) and validated their assemblies. The ASSEMBLATHON team notice high degree of variability between assemblies, which invariably suggests certain possibilities of improvement in the field of genome assembly. Based on the findings of Assemblathon 2, they make broad practical considerations for *de novo* genome assembly and suggested that a single approach might not fit and work well in assembling of two different genomes (Keith R Bradnam et al. 2013). Several research groups are working in the direction to improve the accuracy level of genome assembly data using some new algorithmic approaches. For example, University of Washington is working on a new approach named as 'Sub-Assembly' (Young *et al.* 2010), with an idea of de-fragmentation of genomics DNA. Graph string algorithms for short reads are one of the prospects for the future development of assembly algorithms.

In the 20th century, genome sequencing was more expensive [20] (Figure 2.8) than constructing physical maps, but the development of new high-throughput and massively-parallel DNA sequencing technologies has radically changed the situation, reducing not only cost, but also the time required to sequence an entire genome (Metzker 2009, Mardis 2011). Currently, sequencing a mammalian genome at 30-fold coverage costs ~$10,000, which is comparable to the labour and reagents cost for physical mapping [21]. Though sequencing reads have been assembled by various algorithms, it is still difficult to validate resulting scaffolds and order them across chromosomes without having physical maps. However, by using computational and comparative genomics approaches, and with the aid of completely assembled genomes with reconstructed chromosome structures, such as human, mouse, rat, and cattle, it is possible to predict the order of scaffolds in newly sequenced genomes. Such approaches can even verify predicted chromosome structures using some chromosome features that can be identified from raw sequence reads because of their rarity (Kim *et al.* 2013).

[20] www.genome.gov/sequencingcosts Accessed: 07/06/14
[21] www.illumina.com

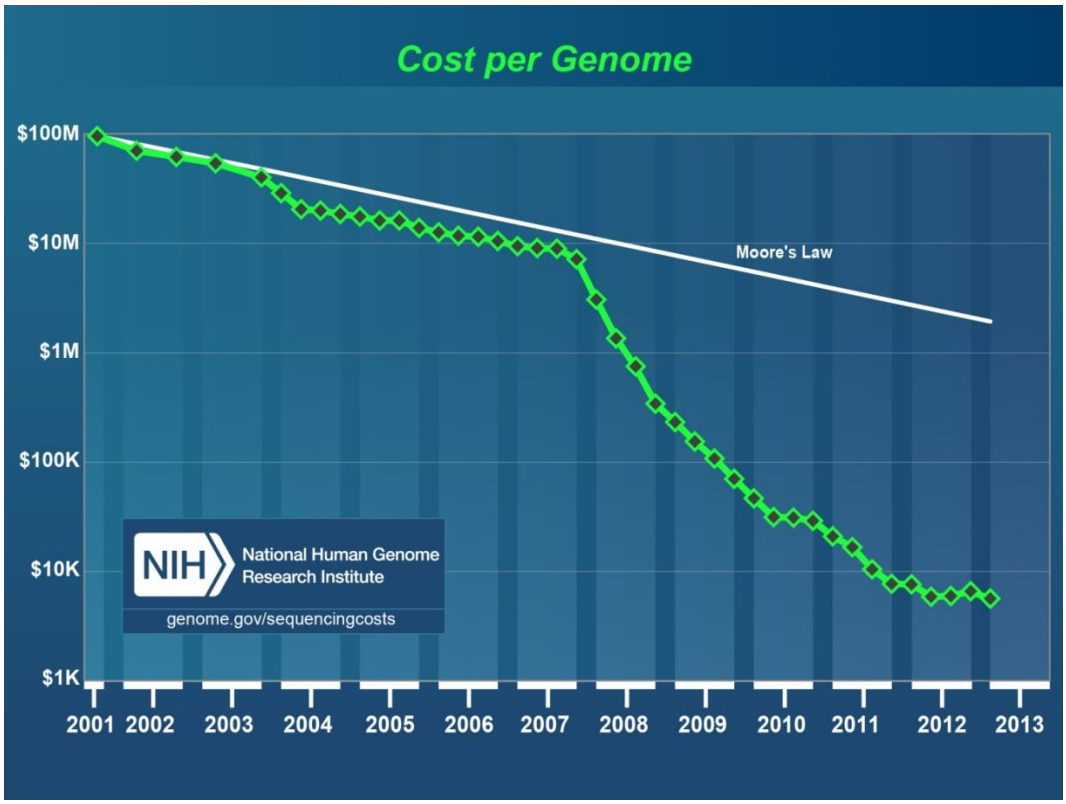


Figure 2.8 Genome sequencing versus cost statistics. Sequencing costs data from the NHGRI large-scale genome sequencing program[22]. The Gordon Moore observation is that over the history of computing hardware, the number of transistors in a dense integrated circuit has doubled approximately every two years. Moore predicted that this trend would continue for the foreseeable future (Brock and Moore 2006). In the above figure, it is clearly shown that the sequencing cost dramatically decreased even lower than the predicted line by Moore's law.

In this section, I first discussed the background information of amniotes biology and give an overview of genome, their organisation and various mapping techniques. I mainly focused on genome organisation and packing of the genetic material. The gene, genome, sequencing and their assembly, duplication and their impact on evolution were reviewed widely. This section also described how the computational complexities and approaches evolved over time. In addition to that, this section reviewed the impact of transposable elements and their role in shaping the genome.

[22] www.genome.gov/sequencingcosts Accessed: 07/06/14

In next section, I will discourse the research work related to chromosomal rearrangements and evolution. I will initiate with the basic concepts of evolutionary mechanisms and gave the detail description of complex terminologies in evolutionary biology. Apart from that, I will also review synteny, chromosomal rearrangements and their impact on amniote evolution.

SECTION 2: EVOLUTION AND CHROMOSOMAL REARRANGEMENTS

2.1 HISTORY OF EVOLUTIONARY CONCEPTS

Charles Darwin published "On the Origin of Species" in 1859 after decades of intense study of zoological and botanical specimens[23] (Darwin 1859). Darwin concluded that all living organisms on Earth are related and have descended from a common ancestor; in other words, all groups of organisms, including animals, plants, and microorganisms, originated from a single ancestral organism. This is now referred as the "theory of common descent". Another one of Darwin's theories, "descent with modification", postulates that organisms with complex features evolved from relatively simple organisms with many gradual modifications occurring over time. Darwin, in his theory of evolution, suggested that the organism with the best adaptive features for their environment would be more likely to survive and reproduce successfully.

In 1909 Wilhelm Johansen identified the fundamental units of heredity, which he called "genes" (Johannsen 1911). This discovery directed the scientific community to identify the entire set of genes in various species. Through these studies, scientists around the world hoped to discover which genes controlled traits of interest. In the early 1900s, the process of constructing genetic "maps" began, in an attempt to identify positions of chromosomal loci responsible for particular quantitative traits.

A major breakthrough in understanding the mechanisms of evolution resulted from the rediscovery of the work of Gregor J. Mendel. Mendel postulated several laws of inheritance and determined that a unit of inheritance exists. Flemming (1882) discovered the chromosomes in the nuclei of salamander cells and confirmed their hereditary nature (Sutton 1903). This discovery created an opportunity to study the biological mechanisms of inheritance and test hypotheses using genetic material. In addition, modern developments in techniques for chromosomal study have made it possible to obtain accurate comparisons of chromosomes in various species and to reconstruct how chromosomes evolved in different clades (Ferguson-Smith and Trifonov 2007).

[23] http://darwin-online.org.uk/specimens.html

2.2 SYNTENY

The genes in multicellular eukaryotes are distributed among a number of chromosomes. The chromosome number in a species is generally between 10 and 100, though in some species this number can be as low as 2, as in jack jumper ant *Myrmecia pilosula* (Crosland and Crozier 1986), or as high as 1440, as in adder's-tongue ferns *Ophioglossum reticulatum* (Khandelwal 1990, Grubben 2004). Each chromosome contains approximately 100 to 1000 genes. The term "synteny" was first introduced by Renwick (1971) to describe two or more genes located on the same chromosome (Renwick 1971, de Grouchy 1972). Whereas, "conserved synteny" is the presence of two or more genes in the same order on one chromosome in two or more species. The order of genes on a chromosome and synteny can be conserved across species (O'Brien and Nash 1982), and such genomic segments with identical gene content are called "Homologous Synteny Blocks" (HSBs) (W. J. Murphy et al. 2005). These synteny blocks have the same gene order without any disruption by rearrangements, which help in tracking the evolutionary histories of genomes (Delseny 2004, W. J. Murphy et al. 2005). The chromosomal rearrangements accumulated through the process of evolution lead to major differences in synteny organisation of different genomes. Therefore, the synteny maps provide insight into a large scale pattern of genetic divergence (Feuillet and Keller 2002, J. Lu *et al.* 2003, Delseny 2004). In addition, using gene order and cross-species synteny information, it is possible to predict the location of unknown genes in a poorly annotated genome from another well-annotated genome (Waterston *et al.* 2002, Gibbs *et al.* 2004, Lindblad-Toh *et al.* 2005). Taking in account synteny can also facilitate annotation and characterisation of a genome (as well as genome assembly) by identifying regions of homology between a genome currently being sequenced and another finished genome (Pop and Salzberg 2008, Kim *et al.* 2013).

Synteny and conserved synteny has been identified using cytogenetic as well as computational genomic techniques for many genomes. However, there have been disagreements amongst scientists as to how to correctly classify "conserved synteny". The work by Ovcharenko *et al.* (2005) on gene desert regions compelled researchers to rethink the definition of conserved synteny and then redefine it as "any conserved sequence block, regardless of whether it encompasses multiple genes, an area containing single genes, or areas devoid of known genes to be considered as synteny block as long

as there is conservation at the sequence level". Various algorithms that apply this new definition have been developed to detect and identify conserved HSB amongst species. (The list of tools which are commonly used for synteny detection and visualization are mentioned in table 2.3).

Most available synteny detection algorithms and tools (Table 2.3) use comparative genomic approaches that compare the genomes of both closely and distantly related species. Apart from computational synteny detection methods, the segments of conserved synteny can also be revealed by molecular–cytogenetic methodology such as ZOO–FISH (Chowdhary *et al.* 1996, Aleyasin and Barendse 1999). Both types of methods allow the characterisation of structural and functional differences in both conserved and divergent genomic regions. Almost every conserved synteny detection tool has some competitive advantage over others in terms of accuracy, algorithmic approaches, and computational complexities. The complexities include strandedness of genes, transpositions, gene insertions, gene inversions, gene duplications, and reciprocal translocations in genomes. Pevzner and Tesler (2003b) developed an algorithm called 'GRIMM-Synteny' to detect synteny blocks in sequenced genomes (Pevzner and Tesler 2003b). The genome complexities previously mentioned are efficiently handled by Ortho-Cluster, which accepts the annotated gene sets of candidate genomes and pairwise orthologous relationships as input and efficiently identifies the synteny blocks (Zeng *et al.* 2008). Similarly, Cinteny tool automatically compares multiple genomes and quantifies evolutionary relationships between species in terms of chromosomal rearrangements with computed reversal distances (Sinha and Meller 2007). Out of all available computational tools (Table 2.3) only AutoGRAPH was designed to provide an interactive display web server to detect preservation of synteny in large portions of a chromosome (macrosynteny), and for only a few genes at a time (microsynteny) (i.e., conserved segments [CS]) with high accuracy. This tool is particularly useful as it can handle not only genome sequences but also meiotic maps and RH maps for a single species (Derrien *et al.* 2007). Similarly, SyntenyTracker follows the set of rules defined by Murphy *et al.* (W. J. Murphy et al. 2005) and defines HSBs using pairwise high-resolution radiation–hybrid (RH) or gene-based comparative maps as inputs. Comparison of AutoGRAPH and SyntenyTracker outcomes showed some differences. The first major difference was detected on cattle chromosome 16 (BTA16), where the "out-of-place" markers were used to create two HSB blocks by AutoGRAPH but were combined into

one HSB block by SyntenyTracker (Donthu *et al.* 2009). The second major discrepancy was reported on cattle chromosome X (BTAX), where SyntenyTracker detected an inversion that was ignored by AutoGRAPH (Donthu *et al.* 2009). Therefore, the SyntenyTracker program has some competitive advantage and more accurate synteny detection when compared to AutoGRAPH (Donthu *et al.* 2009). Recently, Jean and Nikolski (2011) developed SyDiG, which outperforms several other tools (Table 2.3) in detecting synteny in distantly related genomes. Scalable and comprehensive algorithms for synteny detection are available not only for genomes with high degrees of inter- and intra-species chromosomal homology, but also for closely related microbial genomes (Minkin *et al.* 2013). Recently, SynChro was developed; it uses the Reciprocal Best-Hits (RBH) algorithm to reconstruct the backbone of synteny blocks between multiple genomes using their syntenic homologous genes and not DNA alignment. SynChro has an advantage over many other tools as it allows synteny blocks to be overlapping, which supports comparisons involving genomes that have undergone whole genome duplication events. SynChro also allows users to trace small rearrangements that may be responsible for small overlaps or inclusions between synteny blocks (Drillon *et al.* 2014). A newly-developed, user-friendly software package, PhylDiag, uses gene trees to identify statistically significant synteny blocks in pairwise comparisons of eukaryote genomes. PhylDiag takes into account gene orientations, allowed gaps between genes, blocks of tandem duplicates, and lineage specific *de novo* gene births during synteny block identification (Lucas *et al.* 2014).

Table 2.3 List of synteny detection and visualisation tools.

Tool name	References	Remarks
SyntenyTracker	(Donthu *et al.* 2009)	Efficient and accurate
Cinteny	(Sinha and Meller 2007)	Reversal distance measure
OrthoCluster	(Zeng *et al.* 2008)	Mining synteny blocks in multiple species
SyMAP	(Soderlund *et al.* 2006)	Synteny mapping and analysis program Consists of the algorithm to compute synteny blocks and visualise them
AutoGRAPH	(Derrien *et al.* 2007)	Display macrosynteny and microsynteny
SynChro	(Drillon *et al.* 2013, Drillon *et al.* 2014)	Defines conserved synteny blocks
SynBrowse	(Pan *et al.* 2005)	Synteny browser
Sibelia	(Minkin *et al.* 2013)	A scalable and comprehensive algorithm to detect synteny in closely related microbial genomes
GSV	(Revanna *et al.* 2011)	Genome synteny viewer
SyDiG	(Jean and Nikolski 2011)	Uncover synteny in distant genomes

The study of synteny relationships and chromosome rearrangements between the genomes of closely- or distantly-related species yields significant insight into the processes of evolution, development, and gene regulation (W. J. Murphy et al. 2005, Lemaitre et al. 2009). In other words, chromosome rearrangements often play an important role in the evolution of a genome through changes in DNA sequence and organisation. In the next sections, emphasis will be given to discuss chromosomal and genome rearrangements in various species.

2.3 CHROMOSOMAL REARRANGEMENTS

Chromosomal rearrangements are a common type of mutation that occurs in eukaryotic genomes. These rearrangement events occur when a substantial track of DNA is inverted or repositioned on chromosomes (Lysák and Schubert 2013). The repositioning of chromosomal segments results in different classes of events: inversions (Sturtevant 1926, Eisen *et al.* 2000), duplications, fissions, fusions, and translocations. During an inversion, the segment of a chromosome between two DNA breaks becomes inverted and as a result the gene order and nucleotide sequence for the segment is reversed relative to its original order. This mechanism is further classified as either a "pericentric" or "paracentric" inversion. If inversion does not include the centromere, then the inversion is called "paracentric", whereas an inversion spanning the centromere region it is called "pericentric" (Figure 2.9). A translocation occurs when a piece of chromosome breaks off and attaches elsewhere in the genome. There are of two types of translocations: reciprocal and non-reciprocal. Non-reciprocal translocations are one-way transfers of a given chromosomal segment to another chromosome, whereas reciprocal translocations occur when chromosomal segments are exchanged between two non-homologous chromosomes (Griffiths *et al.* 2000). A Robertsonian translocation (ROB), first reported in grasshoppers (Robertson and Rees 1916), is a type of nonreciprocal translocation in which two acrocentric chromosomes break at the centromere and fuse whole long (q) arms to form a single chromosome with a single centromere. During a reciprocal translocation, chromosomes break and exchange fragments (Lysák and Schubert 2013).

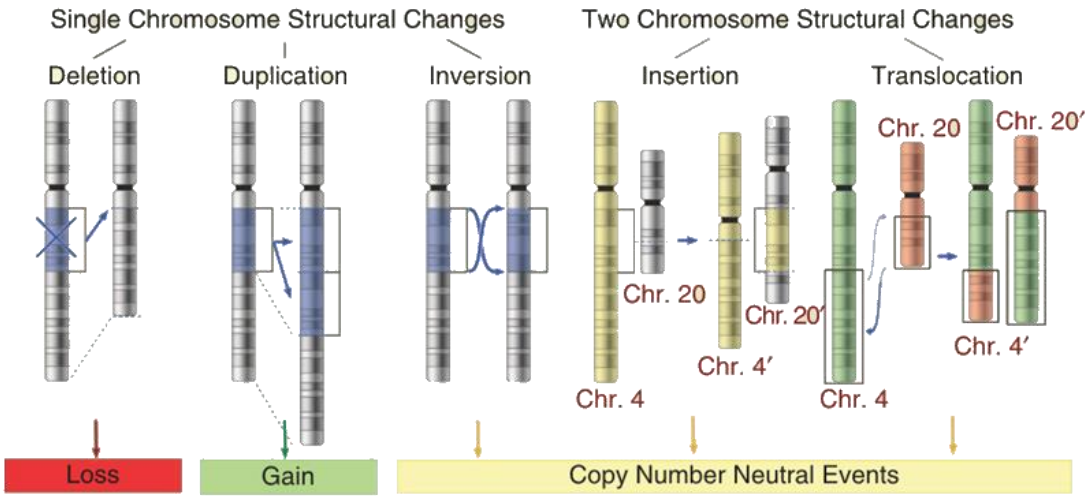


Figure 2.9 A schematic representation of different types of chromosomal rearrangements. The chromosome changes involving a single chromosome or multiple chromosomes are depicted above (Schwab and Amler 1990).

The repositioning of chromosomal segments is known to play an important role in genome evolution. For instance, it was reported that in *Candida albicans* and *C. tropicalis* chromosomal aberrations caused morphological changes (Suzuki *et al.* 1989, Barton and Scherer 1994) and in *Aspergillus nidulans,* rearrangements lead to sterility and negative fitness (Geiser *et al.* 1996). Similarly, chromosomal doubling of *Drosophila melanogaster* chromosomes fails to restore pairing and thus fertility (Dobzhansky 1936). Contrary to the situation observed in insects, rearrangements and doubling of the chromosomal complement in plants does not dramatically reduce fertility (Stebbins 1958). For example, several interchromosomal translocations have been observed in *Helianthus annuus* and *H. petiolaris* genomes (Rieseberg *et al.* 1995, L.H. Rieseberg 2001) that led to lower recombination frequency, but did not affect fertility. The deletion of chromosomal segments causes a loss of genes, while duplication expands gene families (Hannenhalli and Pevzner 1995, Kececioglu and Sankoff 1995, Tesler 2002). Similarly, inversions in higher eukaryotes are associated with reproductive isolation (Noor *et al.* 2001, Iriarte *et al.* 2003), and may therefore contribute to speciation (L.H. Rieseberg 2001). The in depth analysis of chromosomal rearrangements also shows their role in disrupting gene expression and regulation, which can exert genome-wide effects on expression (Harewood and Fraser 2014).

2.3.1 Genome rearrangements in non-mammlian species

Genome rearrangements have been identified both in prokaryotic and eukaryotic organisms (Suyama and Bork 2001). Yeast is an important model in molecular and cellular biology that has helped to decipher the molecular functioning of eukaryotic cells. Because of its small genome size compared to mammals and the phylogenetic diversity of yeast, it is also an ideal model organism for genome rearrangements studies. Prior to the determination of chromosomal rearrangements throughout mammalian genome evolution, extensive studies were conducted with yeast genomes to understand the chromosomal organization and effects of genome rearrangements.

Yeast species have undergone extensive genomic rearrangements, which include chromosome aberration and gene order changes (Langkjær *et al.* 2000, Llorente *et al.* 2000, Fischer *et al.* 2001, Delneri *et al.* 2003, Špírek *et al.* 2003, Fischer *et al.* 2006). Chromosomal translocations have been characterized within the genomes of six closely related *Saccharomyces sensu stricto* species of yeast that mate with one another, but produce sterile hybrids on interspecific pairing (Fischer *et al.* 2000). Fischer and colleagues observed that distantly related genomes can be collinear whilst closely related species may be rearranged. Based on this finding they concluded that rearrangements are not required for speciation in yeast. Studies using genomic comparison of two yeasts (Saccharomyces *bayanus* and *S. cerevisiae)* identified rearrangements between distantly related species, which contradict the Fischer *et al.* (2001) conclusion. Comparative genomic studies of three species, *S. paradoxus, S. mikatae, and S. bayanus*, revealed 20 unique inversions, of which 13 were found only in *S. mikatae,* indicating their relative genome instability (Liti *et al.* 2005). In the above comparisons, the order of genes in the inverted segment was also found to be conserved. Chromosomal rearrangements analysed in *Saccharomyces cerevisiae* strains that were raised for 500 generations by Dunham *et al.* (2002), showed a common translocation point supporting the previous finding that rearrangements can reoccur at the same point in evolution. It also suggests that rearrangements may be adaptive and increase the fitness of the strain (Dunham *et al.* 2002). Similarly, reciprocal translocation between chromosomes VII and XVI appears to cause overexpression of the SSU1 gene in yeast, which is associated with resistance to sulfite concentrations. This rearrangement was shown to be adaptive (Pérez-Ortín *et al.* 2002). Chromosomal rearrangements and their contribution to yeast's copper tolerance have been reported, including one that showed the copy number of the crucial

transcriptional activator CUP2 to be correlated with the level of copper tolerance. The copper-tolerant phenotype correlates with chromosomal rearrangements of genes involved in the response to copper ions (CUP1, CUP2 and COX23); these regions were found to be highly significantly enriched for these genes (Chang *et al.* 2013). Moreover, the impact of environment to fix genome rearrangements has been widely demonstrated in yeast, in which adaptive phenotypes formed due to chromosomal rearrangements in natural populations. Later, it was reported that chromosomes could revert back to the wild-type-like organisation once suitable environment was provided in laboratory experiments (Chang *et al.* 2013).

Comparative analysis of *Caenorhabditis elegans* and *C. briggsae* genomes identified 252 conserved segments and 517 chromosomal rearrangements, with a high amount of transpositions in these two genomes. In addition, it has also been observed that the rates of rearrangements in nematodes is the highest among all eukaryotic species (Coghlan and Wolfe 2002). Comparative studies of *Drosophila pseudoobscura*, its close relative *D. miranda*, and its distant out-group species *D. melanogaster* showed that the rates of rearrangement in these species were even higher than those found in *C. elegans* (Bartolomé and Charlesworth 2006). In addition, it was noticed that the *D. pseudoobscura* chromosomes with the highest level of inversion polymorphisms does not show an unusually fast rate of evolution with respect to their chromosome structure. This suggests that this classic case of inversion polymorphism reflects selection rather than a random mutational process (Bartolomé and Charlesworth 2006).

2.3.2 Genome rearrangements in mammals

In 1970 Susumu Ohno proposed a Random Breakage Model (RBM) of chromosome evolution, which postulated that evolutionary breakpoints occur at random chromosome positions and thus there are no rearrangement hotspots in mammalian genomes (Ohno 1970, Ohno 1973). Nadeau and Taylor (1984) did a comparative analysis between the human and mouse autosomes among 83 homologous loci. They observed that the distribution of lengths of 13 conserved segments in human and mouse genomes fits the distribution expected from a Poisson process and concluded that the evolutionary breakpoints were independently and uniformly distributed across human and mice genomes (Nadeau and Taylor 1984). The RBM has been confirmed by

many studies based on relatively low resolution comparative maps (Alekseyev and Pevzner 2010). Later, with the advancement of comparative genomics, data visualization, and DNA sequencing (see Chapter 2 section 1.4), it became possible to decode various genomes and trace their evolution (W. J. Murphy et al. 2005, Ma et al. 2006). These technological advancements and improved resolution allowed us to observe that the number of small conserved segments appears to be larger than predicted by the RBM (Eichler and Sankoff 2003, Kent *et al.* 2003).

After the completion of the human and mouse genome sequence assemblies, Pevzner and Tesler in 2003 did a detailed comparative analysis of the human and mouse chromosome organisations and identified 281 synteny blocks (Pevzner and Tesler 2003a). Using the Hannenhalli and Pevzner algorithm (2003), they determined that at least 190 "reuse" evolutionary breakpoints were required to transform the mouse genome into the human genome in the most parsimonious scenario (Pevzner and Tesler 2003b). The finding of reuse evolutionary breakpoints in mammals suggests the presence of evolutionary breakage hotspots in chromosomes and contradicts the RBM (Sankoff and Trinh 2004, Sankoff and Trinh 2005). Later, Pevzner and Tesler (2003b) suggested a new model of chromosome evolution that is known as the Fragile Breakage Model (FBM), suggesting that chromosome breakage occurs in fragile regions of the genome (Becker and Lenhard 2007). Trinh *et al.* (2004) investigated the breakpoint regions between the syntenic blocks in humans and mice and discovered that evolutionary breakpoints are not randomly distributed across the genome, supporting the FBM model (Trinh *et al.* 2004, Alekseyev and Pevzner 2011). Based on the comparative study of the human, mouse, and cattle genomes, Larkin *et al.* (2003) independently proposed the idea of breakpoint reuse (Larkin *et al.* 2003). Larkin *et al.* (2003) used direct experiential evidence and counted overlapping EBRs in multi-genome synteny-based comparisons to detect reuse breakpoints. In contrast, the algorithmic approach used by Pevzner and Tesler (2003) identified an excess of small synteny blocks that could be explained only by breakpoint reuse (Larkin *et al.* 2003). While several models like RBM postulate that chromosomal rearrangements are "random" in nature (Ohno 1970), the Fragile Breakage Model (FBM) suggests that there are some specific fragile regions or hotspot in genomes which are prone to break and reorganize throughout evolution (Pevzner and Tesler 2003). Alternatively, the Turnover Fragile Breakage model (TFBM) postulates that fragile regions have a limited lifespan

and they are subjected to undergo birth and death processes, which implies that they can migrate between different genomic locations over evolutionary time (Alekseyev and Pevzner 2010).

Evolutionary breakpoint analysis indicates that the breakpoint regions are gene-dense (Everts-van der Wind *et al.* 2004, Wind *et al.* 2005) and contain an elevated number of repeats (W. J. Murphy et al. 2005, Ma et al. 2006). In a multi-species comparative genome study, Larkin *et al.* (2009) also detected that evolutionary breakpoint regions have higher densities of structural variants, single nucleotide polymorphisms (SNPs), exoniphy, zinc-finger transcription factor genes, retrotransposed genes, and lower densities of highly conserved sequences and meiotic recombination hotspots compared to the rest of the human genome. The genes found in primate EBRs are associated with immune responses, and their enrichment in EBRs suggests that rearrangements may contribute to the development of adaptive phenotypes (Larkin *et al.* 2009). Recently, additional support for the role of EBRs in lineage-specific adaptation has come from analysis of the cattle genome (Elsik *et al.* 2009, Womack 2012). This cattle-based analysis found that gene families encoding proteins present in milk, such as *HSTN,* were affected due to substantial reorganization of cattle chromosome 6 (BTA6) which lead to juxtaposition of *HSTN* next to the regulatory element (*BCE*) important for β-casein (*CSN2*) expression. These events subsequently provided additional immune protection in cattle milk (Elsik et al. 2009, Danielle G. Lemay et al. 2009). Similarly, the β-defencin antimicrobial peptide genes were found within an artiodactyl-specific EBR and expanded in cattle chromosome 27. This might have contributed to the adaptive immune response in rumen evolution, suggesting that these adaptive changes are connected to the increased amounts of microorganisms present in rumens (Elsik *et al.* 2009, Larkin 2012).

In summary, genomes contain prolonged regions that are evolutionary stable for hundreds of millions of years of evolution. In contrast, the fragile or hotspot regions of the genome are prone to breaking and are involved in chromosomal rearrangements because of their underlying genomic sequence features, like segmental duplications, copy number variants, and retrotransposed genes. These sequence features are a resource for producing adaptive phenotypes. Several research findings suggest that evolutionary chromosome rearrangements may have adaptive value and thus are subject

to selection (Ayala and Coluzzi 2005). With the advancement of new genome sequencing technologies and methods of genome assembly, newly sequenced genomes are a great resource for understanding molecular evolution. Along with chromosome organisation as well as gene expression, new full genome sequences will clarify the role of evolutionary chromosomal rearrangements in adaptation and speciation.

Despite experimental difficulties, many speciation and adaptation theories have been proposed to explain evolutionary mechanisms, but the physical as well as genetic evidence has proved to be elusive. Till now breakpoint discoveries derived from precision physical mapping as well as genetic mapping of amniote genomes indicates that these fragile regions are reused in evolution (Pevzner and Tesler 2003b, W.J. Murphy et al. 2005), and enriched with genes and segmental duplications (Bailey et al. 2004, Everts-van der Wind et al. 2004, W.J. Murphy et al. 2005). In addition to that, functional differences of genes in EBRs and HSBs has also been reported (Larkin et al. 2009). The role of repeat sequences in chromosomal rearrangements as well as uneven rates of chromosome evolution in different lineages has been widely explored and well accepted in evolutionary biology (W. J. Murphy et al. 2005). Despite exhaustive studies, no positive relationship between EBRs and their impact in adaptive evolution has ever been made. This has proved to be the most difficult problem of all. In spite of the fact that enormous progress has been made by scientists in recent years towards (see chapter 1 and 2 for more detail review) understanding and determining the relationships between EBRs and various sequence features and their association with probable mechanisms of chromosome breakage in evolution, the role of EBRs in adaptation to the environment is still unclear. The following subsequent chapters will explore the evidence in more detail.

3. DETECTION OF CONSERVED SYNTENY AND ANALYSIS OF EVOLUTIONARY BREAKPOINT REGIONS IN THE PIG GENOME

3.1 INTRODUCTION

Domestic pig (*Sus scrofa domestica*) belongs to genus *Sus* and is a part of the family *Suidae.* According to pig taxonomy review, there are seven species of pigs and 22 subspecies living in different parts of the World (Groves and Grubb 1993). The domestic pig, *Sus scrofa domestica,* is an even-toed ungulates livestock animal, a member of the order Artiodactyla (Figure 3.1). The Artiodactyla order is a distinct clade from rodents and primates that last shared a common ancestor with the human lineage between 79 and 97 million years ago (Mya) (Kumar and Hedges 1998, Hedges and Dudley 2006). Artiodactyls include such animals as sheep, goats, camels, pigs, cows, deer, giraffes, and antelopes. Multiple artiodactyls have evolved features that are adaptive for life on open grasslands. As beasts of burden and/or as sources of meat, milk, hair, and leather, many artiodactyls have assumed important roles in many cultures and are important livestock species.

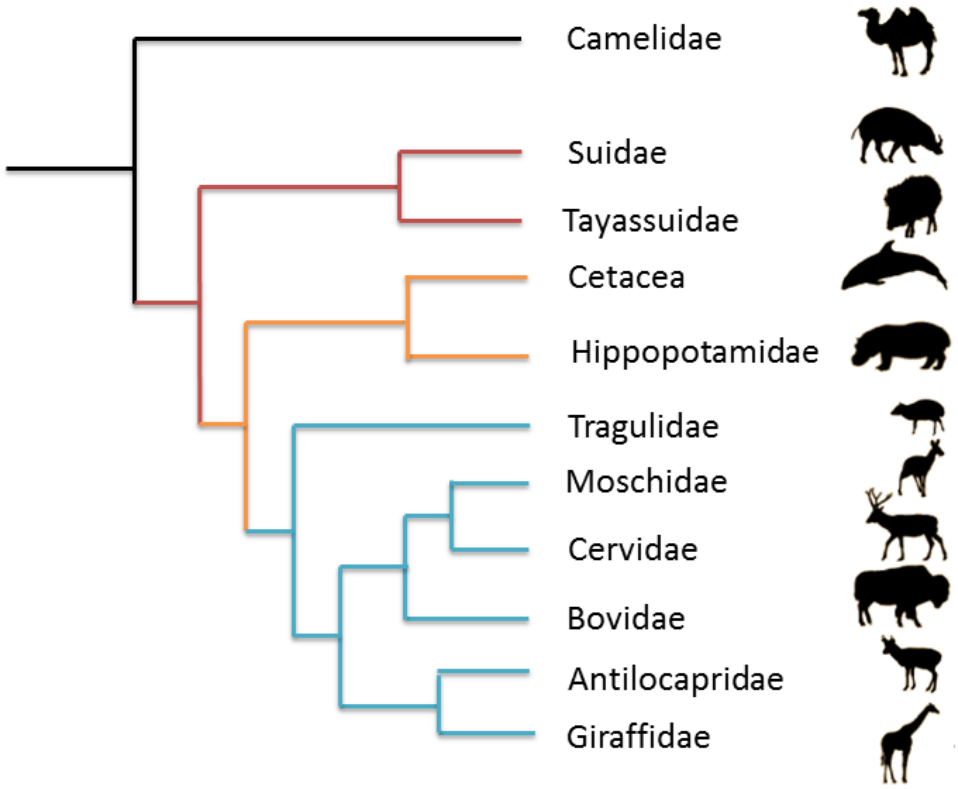


Figure 3.1 Phylogenetic tree of the order Artiodactyla. Some classifications tend to group Cetacea and Artiodactyla into order Cetartiodactyla. The blue colour branches represent the largest suborder Ruminantia in the Artiodactyla which contains 66 living genera and 164 species (Price *et al.* 2005). The branch in orange denotes Cetancodonta suborder, which includes hippos and cetaceans (baleen and toothed whales). The red colour indicates Suina (also known as Suiformes) suborder, which includes Suidae (pig family) and Tayassuidae (peccary family). Camelidae branch in dark black colour highlights Tylopoda suborder, which includes camels. Branch lengths are not proportional to species divergence time. Adapted from (Price *et al.* 2005).

The theories about the origin of domestic pigs were controversial until recently. However, recent genetic and domestication studies suggest that Island South East Asia (ISEA) was the origin of pig-like animals later spread in trajectories by both hunter-gatherers and farmers (Gosden 1995, Latinis 2000, Groves 2007). Moreover, the mitochondrial DNA (mtDNA), and available dental *Sus* fossil-based analysis of wild boars support the theory that pigs originated in the ISEA, later dispersed across Eurasia, and were domesticated approximately 9,000 years ago in several regions of the World (Epstein 1969, Oppenheimer and Richards 2001, Larson *et al.* 2007). Over the centuries,

pig farming in different geographical territories and environmental conditions ranging from extreme hot to cold climates has resulted in formation of breeds with distinct biological traits such as heat or cold tolerance, food adaptations, and disease resistance, which invariably favour their survival under environmental stresses. Pigs have also long undergone a breeding process by farmers for a variety of attributes with a major focus on productivity traits such as meat yields and fertility. To date, there are likely to be over 700 pig breeds worldwide of which two thirds reside in China and Europe (Epstein 1969, Oppenheimer and Richards 2001, Larson *et al.* 2007). There are five international trans-boundary (found in more than one country) pig breeds from the United States (US) or Europe [24] that dominate in the world. Pig breeds vary greatly in size, skin colour, body shape, ear carriage, behaviour, profligacy, and other traits. Nowadays, according to the food and agriculture organization (FAO) pigs are one of the most important nutritional sources of animal protein in the world[25]. A recent World health organization (WHO) report predicts a growing increase of meat production from 218 million tonnes in 1997-1999 to 376 million tonnes by 2030 [26] (Pilling and Rischkowsky 2007). Similarly, a study of human food chains by Bonhommeau *et al* shows a global trend toward the incensement of diets richer in meat from 1961 to 2009 by 3% (Bonhommeau *et al.* 2013). These reports indicate a high demand of meat including pork around the world. It is expected that world population of domestic pigs will reach 1 billion by 2015 to fulfil the demands of growing human population[27].

Pigs are of particular interest for scientific studies not only because of existing breeds that show great phenotypic varieties for morphological, physiological and behaviour traits but also because of their similarities with humans anatomically, physiologically, and genetically (Rothschild and Ruvinsky 2011). Therefore, the utility of pigs in biomedical research promises many advantages compared with other animals such as mice and rats (Prather 2013). Due to physiological and biochemical advantage of pigs over other counterpart biomedical model organisms, pigs are treated as a model organism for humans to understand complex traits such as obesity (Kogelman *et al.* 2013), arthritis, Parkinson, Alzheimer (Martien AM Groenen et al. 2012), cancer (Flisikowska *et al.* 2013) and cardiovascular disease (Tumbleson and Schook 1996). Pigs

[24] http://dad.fao.org/
[25] http://www.fao.org/docrep/007/y5019e/y5019e03.htm Accessed: 14/06/2012
[26] http://www.who.int/nutrition/topics/3_foodconsumption/en/index4.html Accessed: 14/06/2012
[27] http://www.fao.org/ag/againfo/themes/en/pigs/home.html Accessed: 14/06/2012

are also proven to be the most successful non-primate animal for xenotransplantation in humans (Lunney 2007). The recent comparative anatomical analysis indicates differences between porcine and human organs, but still pigs are currently the only animal being considered as a source of organs for transplantation to humans (Schmoeckel *et al.* 1998, Goddard *et al.* 2000). For example, the xenotransplantation from non-human primates to humans were initially found more clinically suitable but later it was discovered that there is a higher risk of disease transmission from primate organs to humans than from pig organs to humans (Michler 1996). The xenotransplantation may transmit potentially lethal viruses from non-human primates to humans, including Ebola, Marburg, hepatitis A and B, herpes B, SV40, and SIV, and hence it is considered not safe to use non-human primates for this purpose (Vanderpool 2002, Matoušková *et al.* 2013).

Pigs also exhibit multiple adaptations. They have a strong sense of smell, providing a reason why they are used to sniff out truffles — edible fungi found underground[28]. The sensing ability of pigs is confirmed by the large number of the olfactory receptor (OR) genes present in the pig genome. Recently it has been found that the number of OR genes in the pig genome is larger than in the human, mouse and even dog genomes, which corroborates the pig's physical sensing ability and reflects the strong reliance of pigs on their sense of smell while scavenging for food (M. A. Groenen et al. 2012). Additionally, pigs are omnivorous animals feeding on a variety of food of both plant and animal origin, and are indiscriminative in feeding. This unique ability probably made pigs able to survive in harsh environments and also an attractive target for domestication.

The pig genome consists of 18 pairs of autosomes and X/Y sex chromosomes. The high quality pig whole genome RH maps (Hawken *et al.* 1999), linkage maps and bacterial artificial chromosome (BAC) clone libraries (Anderson *et al.* 2000) have been constructed to discover the small genomic regions of particular interest (e.g., loci controlling economically important quantitative traits; quantitative trait loci (QTL) (Sean J Humphray *et al.* 2007). The fatness and muscle traits linked to chromosome X, were initially investigated with linkage and RH mapping of 10 pig genes (Čepica *et al.* 2006). In other work, 21 genetic markers were mapped to a QTL region controlling for meat

[28] http://ori.hhs.gov/education/products/ncstate/pig.htm Accessed: 14/06/2012

quality on pig chromosome 17(Ramos *et al.* 2006). The QTL related to muscle mass and fat deposition (backfat thickness) were reported and confirmed on pig chromosomes 7 and 2 (de Koning *et al.* 1999, Rattink *et al.* 2000, Tanaka *et al.* 2006). A comprehensive list of economically important pig QTLs with their genomic locations are available from the PigQTL database for further exploration and analysis[29] (Hu *et al.* 2013).

The chromosome rearrangement studies have identified a number of evolutionary events including duplications, inversions, translocations, fissions and fusions in many pig chromosomes once compared with human, mouse, rat, dog (Jiang *et al.* 2005) and cattle (Pinton *et al.* 2003) chromosomes. For example, the porcine-human whole-genome RH comparative map constructed with 2,274 loci, including 206 ESTs and 2,068 BAC-end sequences, identified a total of 51 conserved synteny groups that include 173 conserved segments between the human and the porcine genomes (Johansson *et al.* 1995, Meyers *et al.* 2005). Similarly, Rink *et al.* were also able to reveal a high degree of gene order conservation in porcine-human comparative RH map, with at least 60 large scale genome rearrangements and an additional 90 micro-rearrangements (Rink *et al.* 2002). Furthermore, Sun *et al.* (1999) have validated the extensive synteny and gene order conservation between the human chromosome 13 and pig chromosome 11 using FISH mapping technique(Sun *et al.* 1999). A high-resolution comparative RH map constructed for porcine chromosome 2 (SSC2) showed four conserved segments between the SSC2 and human chromosomes 11 (HSA11), 19, and 5 (Rattink *et al.* 2001). Later, the rearrangement of gene order in the segment HSA11p15.4-q13 was observed and confirmed to be inverted on the SSC2 (Rattink *et al.* 2001). Additionally, 29 evolutionary breakpoints were reported though a high resolution comparative mapping between human and pig chromosomes 2 and 16 (Lahbib-Mansais *et al.* 2006). The high resolution, bacterial artificial chromosome-based physically anchored, human-pig comparative maps were used in the pig genome sequencing project (Meyers *et al.* 2005, S.J. Humphray *et al.* 2007). The physical maps enabled coverage of over 98% of the 18 pig autosomes (S.J. Humphray *et al.* 2007) and provided a template for genome sequencing and assembly of physically-anchored sequences across the genome (McPherson *et al.* 2001, Warren *et al.* 2006, Lewin *et al.* 2009).

[29] http://www.animalgenome.org/cgi-bin/QTLdb/SS/index Accessed: 14/06/2012

The recent advancement and developments in the next generation sequencing techniques and reduction in sequencing costs (Shendure and Ji 2008) and henceforth an increase in the genomic data, empower evolutionary biologists to peruse, interpret and understand the evolutionary mechanisms at genomic level. The whole genome sequencing (WGS) of pigs has been initiated by the Swine Genome Sequencing Consortium (SGSC). The pig WGS sequence was performed using DNA isolated from a single Duroc sow (Schook *et al.* 2005, Archibald *et al.* 2010). The capillary sequencing was done at the Korean Livestock Research Institute, whereas the Illumina/Solexa sequencing was completed by the Wellcome Trust Sanger Institute and Beijing Genomics Institute (BGI) (~40X coverage) through funding provided by Cooperative State Research, Education and Extension Service at the United States Department of Agriculture (CSREES-USDA)(Schook *et al.* 2005, Chen *et al.* 2007, Archibald *et al.* 2010). The current pig genome assembly (Sscrofa build 10.2) comprises 2.60 Gbp of DNA sequence assigned to chromosomes and 212 Mbp in unplaced scaffolds. This recently accomplished pig genome sequencing and annotation empowers us to study the chromosomal evolution in mammals, and connect chromosomal rearrangement events to changes gained by species during adaptation. Also, the genomic data facilitate the understanding of genetic complexity and assist in elucidating genetic variations that contribute to economically important traits and animal diseases(Jiang and Rothschild 2007).

Therefore our study aimed to investigate the chromosomal rearrangement events in the pig genome and their contribution to adaptive changes occurring during pig genome evolution. The first objective was to detect pig and artiodactyl EBRs with high accuracy. The second objective was to determine the probable impact of chromosome rearrangements on gene networks in pigs using gene enrichment analysis. In addition, the distribution of TEs families in and around pig and artiodactyl EBRs were compared to explore the role of TEs in the pig chromosome evolution. These studies were carried out using the pig whole-genome sequence assembly.

The following lists the work performed by me in this chapter:

- Identified the homologous synteny blocks amongst seven mammalian genomes.
- Discovered evolutionary chromosomal breakpoints and analysed them.
- Detected novel porcine bitter taste receptor genes, and connected these to the EBRs.

- ❖ Detected transposable elements in the pig genome and performed enrichment analysis in EBRs.
- ❖ Enrichment analysis of genes present within and around EBRs.

The validation of certain dubious EBRs (detected by me) was done with FISH techniques by Dr. Katie Fowler at University of Kent.

3.2 METHODOLOGY

3.2.1 Identification of homologous synteny blocks

Seven sequenced mammalian genomes assembled to the chromosomal level were compared: cattle (UMD 3.0), dog (Cfam 2.0), horse (equcab 1.0), macaque (mmu 2.0), rat (rn 4.0), orang-utan (ponAbe 2.0) using the pig (build 10.2) and human genomes (hsg37) as references. All the genomes were separately aligned against the pig and also human genomes using the SatsumaSynteny program (M.G. Grabherr *et al.* 2010). In order to define pairwise HSBs between each of the genomes and the human or pig genomes the SyntenyTracker program was used (Donthu *et al.* 2009). The SyntenyTracker program settings allowed detection of HSBs >500 Kbp, >300 Kbp, >100 Kbp in the reference genome. Furthermore, a Perl script was written to split overlapping HSBs found the SynteyTracker output. The script finds the HSBs overlapping EBRs in at least one other target species and checks for probable breakpoints across all species studied at that position. If there were any small rearrangements detected in any target species then the corresponding HSBs were split to reveal missed EBRs (Figure 3.2). The visualization of HSBs using the pig or human chromosomes as references was performed in the Evolution Highway (EH) comparative genome browser[30].

[30] http://evolutionhighway.ncsa.uiuc.edu

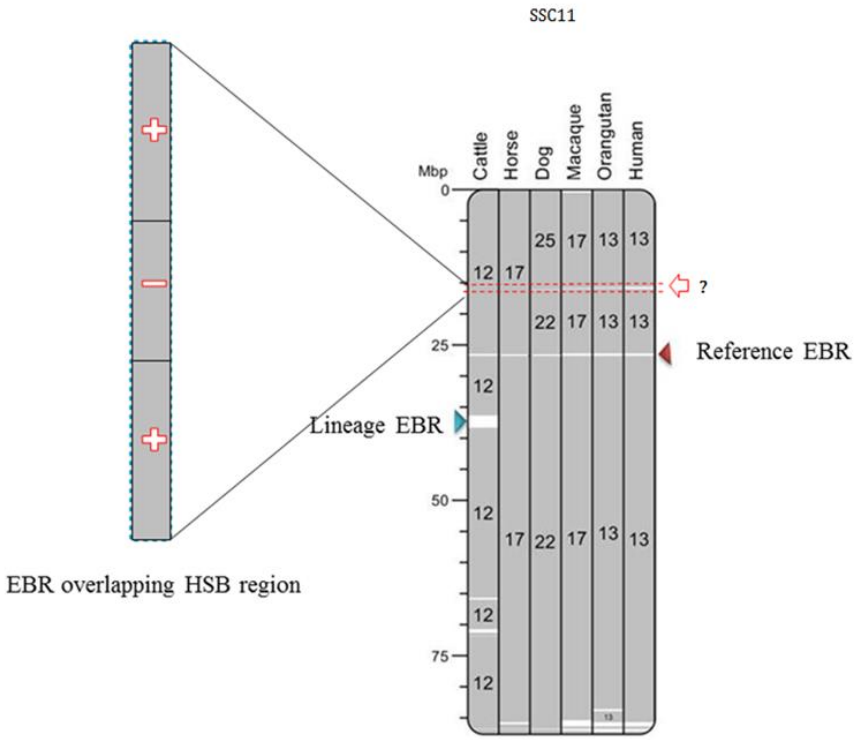


Figure 3.2 Detection of missed rearrangement events in HSBs. The visualisation shows a comparision of 6 mammalian species using the pig chromosome 11 (SSC11) as a reference. The grey blocks indicate HSBs; with the target species chromosome numbers inside the blocks. The white colour indicates the EBRs or gap regions between HSBs. All HSBs in the target species were further checked for small rearrangement events overlapping with EBRs detected in at least one pairwise comparison. If a small rearrangement was identified within an HSB region then the original HSB was split to reveal missed EBRs.Plus (+) and minus (-) in figure indicate the orientation of the HSBs compared to the reference chromosome.

3.2.2 Identification and analysis of evolutionary breakpoints regions (EBRs)

The EBRs were identified as intervals demarked by two adjacent HSB boundaries on the same reference chromosome. EBRs were assigned to phylogenetic lineages using the following species topology: ((pig, cattle), (dog, horse)), (rat, ((human, orang-utan), macaque)).

To perform phylogenetic classification of EBRs, a custom algorithm was developed to define and classify different types of EBRs in genomes: lineage-specific (EBR that are present in one species), ordinal (EBRs that occur in all species from the same order),

and superordinal (EBRs present in species from the same super-order) (Figure 3.3 and 3.4) by setting a score based on the probability of an EBR to belong to different phylogenetical nodes. As an input this algorithm uses a tab-delimited table containing coordinates of pairwise HSBs for all species compared to a single reference genome. Then it defines EBRs as intervals in-between two adjacent HSBs that belong to the same reference chromosome. Once the coordinates of probable EBRs are extracted, the algorithm checks the EBRs and classifies them in accordance with phylogenetic relationships of the species involved in the analysis. For a reliable classification of EBRs two scores were calculated for each EBR– a phylogenetic score and a gap score.

- The *phylogenetic score* shows if an EBR is present in all species from the expected clade. For example, if an EBR is "pig-specific" and the pig genome was used as a reference for the chromosome comparison, then the highest quality EBR is expected to be present in all target species at the same reference genome position (phylogenetic (expected) score = 1, means expected clade EBR is classified with 100% accuracy).

  If the EBR is not detected in one of the species-**[**Clade: Break*(species1, species2, species3, species4),* NoBreak*(species5),*Break *(species6, species7)***],** then the score will be (*ExpectedPhyloScore-(NoBreakNum/TotalSpeciesNum)*) given that seven species were aligned with the pig genome sequence. Using above clade as an example, the phylogenetic score will be ~0.86 (1-(1/7)).
- The *gap score* is affected by the number of species in which the EBR is present and whether the EBR detected in one of the genomes overlaps with more than one non-overlapping EBRs in other genomes. For example, the phylogenetic score equals one and the gap score <7, implies that the EBR present in one genome overlaps with intersecting EBRs in other genomes.

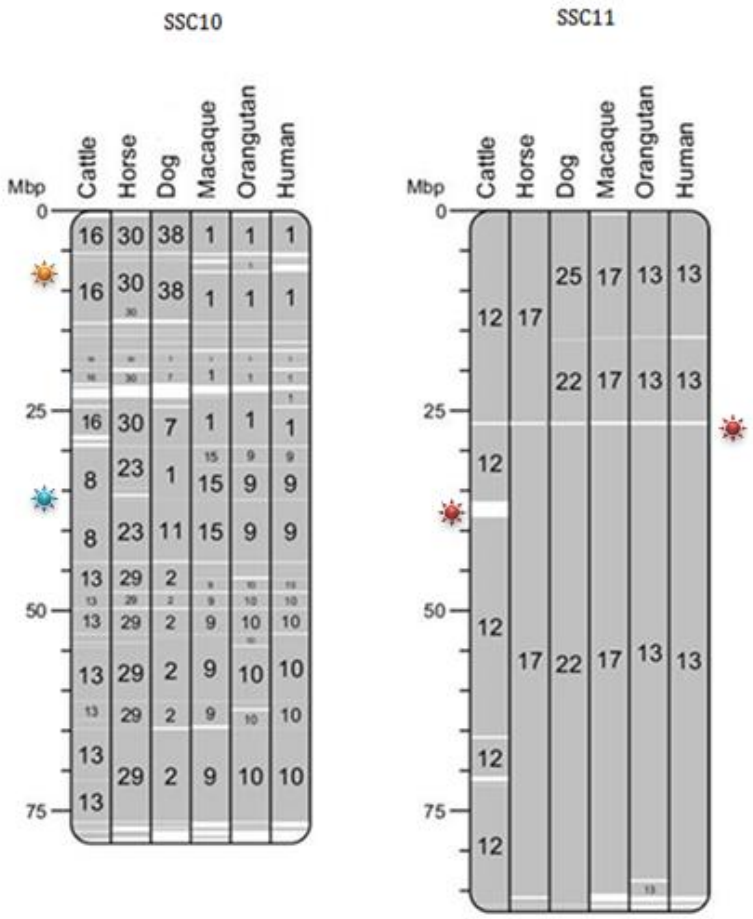


Figure 3.3 Examples of HSBs and visualisation of EBRs using cattle, horse, dog, macaque, orang-utan, and human genomes on SSC10 and SSC11. In all chromosome images, the grey blocks indicate HSBs, with the target species chromosome numbers indicated inside the blocks and the white regions indicating EBRs or gaps. The orange arrow indicates the position of a gap region. Any breakpoint is called a "gap" if it overlaps with more than one EBR that does not overlap with each other in different target species or it overlaps with more than one EBR in the same target species. The artiodactyl (order-specific) EBR is indicated with a blue star on SSC10. This EBR is present in the cattle and pig genomes (pig genome is used as a reference) suggesting that the cattle and pig genomes have a chromosome organisation different from all other mammals in this region. The breakpoint present across all the species is a pig-specific EBR which is highlighted with a red star in this example. There is one additional lineage specific breakpoint which was highlighted with a red star on SSC11, in which there is only one chromosomal break detected in the cattle lineage.

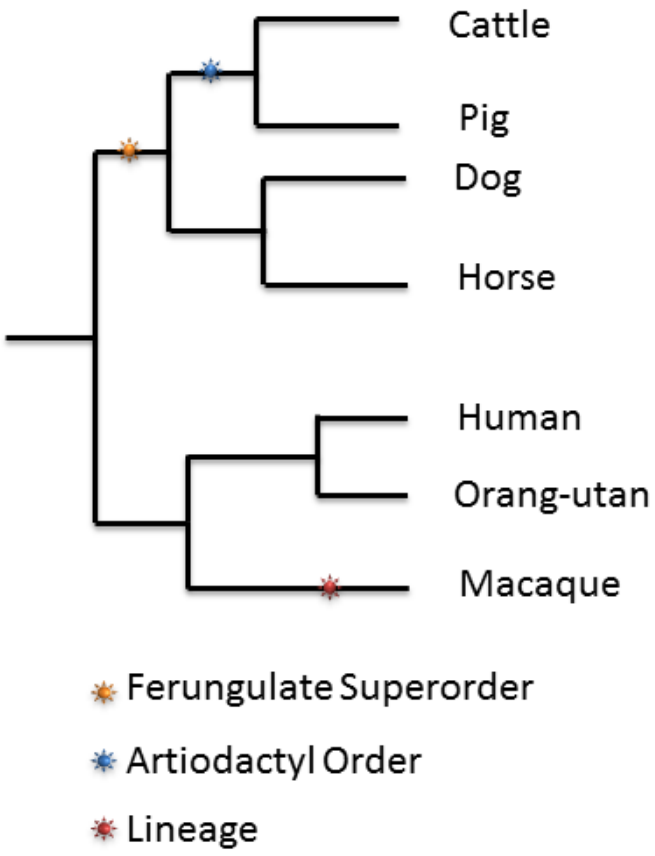


Figure 3.4 The phylogenetic origin of EBRs. The EBRs phylogenetic relationships are denoted by stars in this tree. The blue colour star highlights an artiodactyl EBR which occurred in the cattle and pig ancestral lineage. The yellow colour star is used to represent a ferungulate EBR which occurred in the common ancestor of artiodactyls, dogs, and horses. The lineage-specific breakpoint found in a single species is represented with a red star. The branch lengths are not proportional to divergence time.

If lineage–specific EBRs are identified using an out-group genome as a reference, (e.g. pig-specific EBRs are detected in the human genome) then a phylogenetic score of 1 would imply that the EBR is present in only one species. The score would be decreased if the EBR was present in another genome as well, e.g., an overlapping EBR in the pig and mouse genomes has the phylogenetic score of 0.5 implying that it is present in two lineages. Moreover, the gap score in such cases will increase because the number of genomes sharing the EBR increases. The algorithm for the EBR classification was implemented as a custom *Perl* script (Figure 3.5).

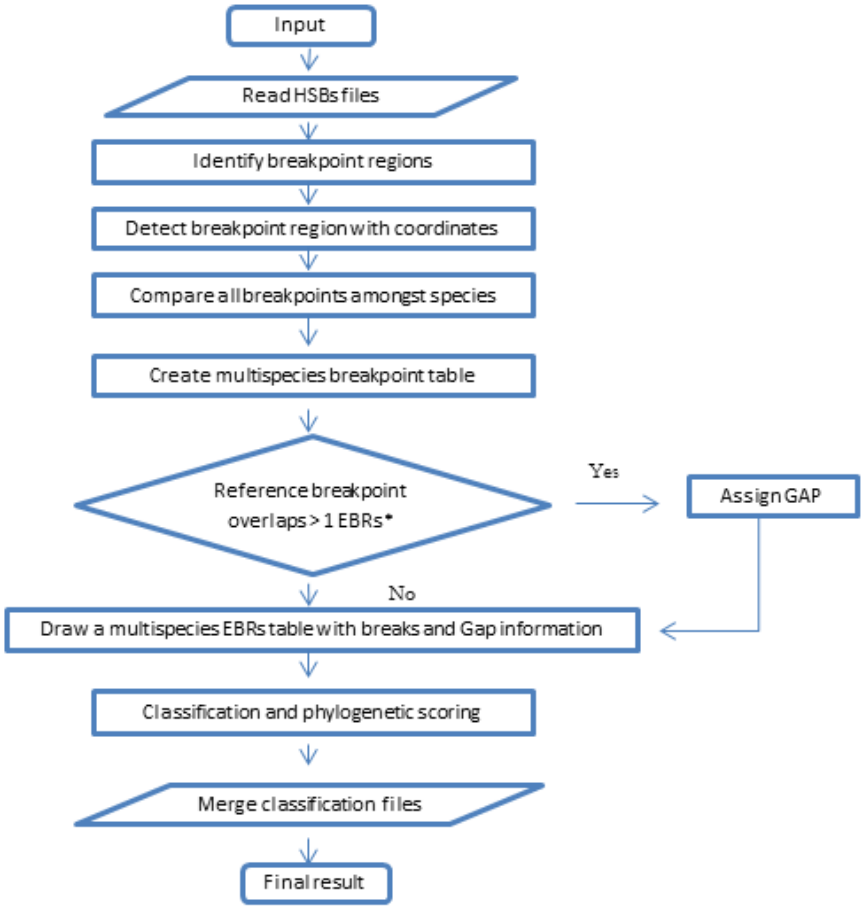


Figure 3.5 Schematic representation of the EBR identification and classification process. The star (*) indicates that an EBR overlaps with more than one EBR that do not overlap with each other in different target species or it overlaps with more than one EBR in the same target species.

3.2.3 Detection of novel porcine bitter taste receptor genes

A total of 105 sequences from taste receptor, type 2 (TAS2R) gene family from cattle, dog, chimp, mouse, human, and pig genomes were collected. A tBLASTn comparison of the genes was performed against the pig chromosomes and unassigned contigs using E-value of e-10 as the threshold. All non-overlapping pig sequences that had matches >100 aa with known TAS2R genes were extracted. I added 1,000 bp to the 5' and 3' ends of the extracted sequences. Then I translated all six frames from all the DNA sequences into protein sequences and performed a BLASTp analysis against the NCBI nr database to identify orthologs of putative TAS2R genes. After detection of the matches I searched the pig sequence for the closest start and stop codons near the longest match from a known TAS2R gene. I considered an identified pig TAS2R gene

'intact' if it encodes for >290 aa, and has no frame-shift mutations or premature stop codons.

3.2.4 Transposable elements enrichment in EBRs

The distribution of TEs and other repetitive sequence families were studied in and around pig and artiodactyl EBRs. Detection of repetitive elements in the reference genome was performed by RepeatMasker (version 3.3.0)[31] (Smit *et al.* 2004) using Repeat library v.20120124. An in-house pipeline was used to calculate the densities of TEs in each EBRs and non-EBRs regions of the pig genome. The pipeline divides chromosomes into 10 Kbp segments (bins) and calculates the number of bases from each TE family within each bin. The distribution of TE families was compared with the average number of bases (>100) in all genome bins between the EBR regions and other parts of the genome. A Student's t-test with unequal variances was used to identify repeat families that were unequally distributed in EBRs when compared to the rest of the genome. FDR (Benjamini-Hochberg) and lfdr (Efron-Bradley) algorithms were used using FDRTool to calculate critical values and control for a false positive discovery rate (Strimmer 2008).

Apart from analysing the overall density of TE elements in pig EBRs, a potential influence of lineage-specific TE insertions was searched for on genes involved in the taste transduction pathway. Henceforth, an attempt was made to look for TE that were inserted into the taste transduction genes (focusing on exons, 5' and 3' untranslated regions (UTRs)) and found in/near the pig-specific EBRs. The taste transduction pathway-related genes were extracted from the Kyoto Encyclopaedia of Genes and Genomes (KEGG)[32] database and were cross-verified using the pig EnsEMBL gene set. Later, the genome coordinates of 29 taste transduction genes and their corresponding transcript sequences were extracted from the Sanger pig transcript dataset[33] for the identification of exons, 5' and 3' UTR sequences.

3.2.5 FISH Analysis

The cytogenetic technique, FISH was used to check the chromosomal rearrangement in SSC3. The FISH analysis was performed by Dr. Katie Fowler (University of Kent). In this case, specific BAC probes CH242-207N16 and CH242-191E23 from the CHORI-

[31] http://www.repeatmasker.org/
[32] http://www.genome.jp/kegg/
[33] ftp://ftp.sanger.ac.uk/pub/sf7/sscrofa10_2/e67_final_names/ Accessed: 14/06/2012

242 BAC library were used. Based on the pig genome assembly (build 10.2) BAC clone CH242-207N16 was assigned to chromosome 10 and clone CH242-191E23 to chromosome 3 forming boundaries of two EBRs detected in our analysis.

The ordering of BACs from CHORI[34] and FISH analysis were performed at the University of Kent. A sterile technique was used to streak an agar plate and the plate was placed at 37 ^{0}C overnight. The following day, the colonies were removed from the plate using a Pasteur pipette and sterile PBS. Subsequently, the QIAprep Spin Miniprep kit (Qiagen) was used (following the manufacturer's instructions) to purify the plasmid DNA. Later, the BACs were amplified using the Illustra GenomiPhi V2 DNA Amplification Kit, following the manufacturers' instructions. The nick translation was subsequently performed to directly label the BACs with fluorophores (FITC=green, Texas Red=Red), Agarose gel electrophoresis was performed to ensure the probes were of the correct size for downstream FISH analysis. The probes were purified using the QIAQuick Nucleotide Removal Kit (Qiagen), following manufacturers' instructions. Metaphase preparations were dropped onto clean microscopy slides and observed under phase contrast microscopy to check for density of metaphases and presence of cytoplasm. Same species FISH and fluorescence microscopy was subsequently performed in house. Thereafter, FLPter analysis was performed using ImageJ software to ensure the probes hybridised to the expected chromosomal locations.

3.2.6 Enrichment analysis of genes present within and around EBRs

Gene enrichments were searched for in and around pig and artiodactyl EBRs, using the human and pig genomes and gene sets as references. The following materials were used for the analysis:

3.2.6.1 Using human genome as a reference

The human gene data set was downloaded from the NCBI ftp server[35]. The total number of genes annotated by NCBI was 45,542 which included unplaced, mitochondrial and pseudo-genes. The set was filtered to remove 74 mitochondrial genes, 511 genes on chromosome Y and 2,288 genes located on unplaced scaffolds. Additionally, all gene annotation files for human genomic contigs (GRCh37.p2) were

[34] http://bacpac.chori.org/libraries.php Accessed: 14/06/2012
[35] ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/ Accessed: 14/06/2012

downloaded from the NCBI human genome database[36] and gene coordinates were extracted. Later, the chromosome coordinates were added to each of the genes using Entrez gene ID as a matching criteria between the gene and contig annotations. Finally, a filtered set of 37,299 genes (including putative or hypothetical genes) was obtained which was used during the pig EBRs gene enrichment analysis using human genome as a reference.

3.2.6.1.1 Gene network analysis within pig EBRs

The human genes were checked for an overlap with 189 pig-specific EBRs identified in the human genome (as a reference). A master file was created that contained information about the human genes with sequence coordinates overlapping with pig EBRs defined in the previous step (see methodology section 3.2.2). The genes that were located within +/-500 Kbp from the pig-specific EBRs were also identified. The total number of human genes that were found in or near 189 pig EBRs was 2,848. The genes were submitted to the DAVID v6.7 and separately to the GeneGo MetaCore database[37] (MetaCore™ v.6.9 build 30881) for the gene network enrichment analysis using the human filtered set of 37,299 genes as a reference. The false discovery rate (FDR) of 5% was used as a significance threshold for the analysis.

3.2.6.2 Using pig genome as a reference

The latest pig gene annotation files were downloaded from NCBI[38] and EnsEMBL[39] and coordinates of all pig genes annotated in these databases were extracted. A total of 25,827 genes were found which were predicted by the NCBI in the pig genome (including unplaced scaffold, Y, and MT), out of which 8,051 genes were assigned a gene name. Similarly, in the EnsEMBL dataset a total of 25,009 genes were predicted (with all UN, MT, and Y), out of which 15,554 genes were assigned a unique name.

The set of homologs between the pig and human genomes were downloaded from EnsEMBL[40]. In addition, pig and human EnsEMBL gene annotation files were downloaded from the EnsEMBL server, which contain gene location and structure information of the respective genomes. Later, the unplaced scaffold (24), chromosome

[36] ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ Accessed: 14/06/2012
[37] http://www.genego.com
[38] ftp://ftp.ncbi.nlm.nih.gov/genomes/Sus_scrofa/GFF/ Accessed: 14/06/2012
[39] ftp://ftp.sanger.ac.uk/pub/sf7/sscrofa10_2/e67_final_names Accessed: 14/06/2012
[40] ftp://ftp.ensembl.org/pub/release-67/mysql/ensembl_mart_67/ Accessed: 14/06/2012

Y (23) and mitochondrial (14) genes were filtered out from the pig EnsEMBL gene list. These genes were filtered out from the list because mitochondrial genes and genes on chromosome Y were not analysed for overlaps with chromosome EBRs. The gene set at this stage contained 19,094 annotated genes in the pig genome (Table 3.1).

Moreover, the filtered file with 19,094 annotated pig genes was further filtered for genes that had more than one known ortholog in the human or pig genomes. Depending on the number of genes found in each species, EnsEMBL differentiates among *one2one*, *one2many* and *many2many* gene relationships. These relationships and their potential influence on gene annotations are discussed in Chapter 2 section 1.2.2.1.

Table 3.1 Number of pig and human homologous genes in the pig genome.

All homologs	Not placed to chromosomes	Mitochondrial	ChrY	Filtered homolog set*	one2one orthologs	one2many orthologs	many2many orthologs
21,099	1,976	13	16	19,094	**12,660**	4,799	1,634

* Final filtered set of pig genes which does not include unplaced, chromosome Y and mitochondrial regions.

3.2.6.2.1 Orthologous gene set

A total of 12,660 pig genes annotated by EnsEMBL (build 67) were extracted, mapped to known chromosome positions in the pig genome and with a single known ortholog in human chromosomes. This set was further filtered by excluding those genes which were located in the non-orthologous positions of the pig and human chromosomes identified from the whole-genome pig-human SatsumaSynteny alignment dataset used to build pairwise HSBs between the human and pig genomes. The orthologous positions were identified either by a direct overlap with the pig-to-human sequences alignments, or predicted if a gene was located in between two homologous positions within an HSB as defined by the sequence alignment. Those genes that had a single ortholog in the human and pig genomes and were located in an EBR in the pig genome were also kept. As the result of this filtering step 613 genes were removed. To produce a comprehensive set of genes with well-defined orthologous relations between the pig and

human genomes, 127 genes were added to the dataset that were found in the independent pig genome annotation from NCBI, had no coordinate or name overlap with the annotated pig EnsEMBL gene set, had human orthologs located in the homologous positions in the pig and human chromosomes as defined by sequence alignments (see above). A further 209 genes that had assigned gene names by NCBI only, were found in homologous positions in human chromosomes confirmed by the whole-genome sequence alignment and had >30% overlap with unnamed pig genes in the EnsEMBL gene set were added.

The resulting set of 12,383 orthologs between the pig and human genomes was used to build human-pig HSBs with SyntenyTracker program (Donthu *et al.* 2009). This led to the detection of 109 genes that were located in unexpected positions within HSBs ("out-of-place") or represented a single gene HSB ("singleton"). These genes were excluded because they are likely to be located in misassembled pig genome intervals and could affect our gene network analysis. At the end, there was a set of 12,274 genes that were used for the gene network analysis.

3.2.6.2.2 Gene network analysis within pig EBRs

The 12,274 pig genes with defined orthologs in the human genome were checked for an overlap with 192 pig-specific EBRs found in pig chromosomes. In total 1,329 genes were detected that are located within the EBRs or in ±500Kbp intervals adjacent to the EBR boundaries. To find gene ontology (GO) categories overrepresented in the genes present in pig EBRs, the human EnsEMBL gene IDs were used.

MetaCore GeneGo v.6.9 build 30881 online database[41] and DAVID v6.7 were used to identify GO categories overrepresented within the gene set found in/near pig EBRs. The complete set of 12,274 orthologous genes were used as a background for this analysis out of which 12,249 EnsEMBL gene IDs were recognised by MetaCore. Out of 1,329 genes in/near the EBR regions 1,320 were recognized. The KEGG[42](Ogata *et al.* 1999) were later used to look into the pathways that were found significantly enriched in the pig EBRs gene set.

[41] http://www.genego.com
[42] http://www.genome.jp/kegg/

3.3 RESULTS AND DISCUSSION

3.3.1 EBRs

Using the pig-based HSB sets and stringent filtering criteria (see methodology section 3.2) 192 "consensus" pig-specific EBRs were detected. These EBR were consistently present in all three HSB datasets or in the 300 Kbp and 100 Kbp sets (missed in the 500 Kbp set because of a lower resolution of this set). Similarly, when the human genome was used as reference, 189 pig-specific EBRs were detected. In addition to pig EBRs, the EBRs present in the artiodactyl ancestral genome (common ancestor of pigs and cattle in our dataset) were identified. A total of 20 and 18 artiodactyl EBRs were identified using the pig and human genomes as references, respectively. The number of lineage-specific EBRs in the cattle genome detected at the 500 Kbp resolution set (Elsik *et al.* 2009, Larkin *et al.* 2009) is comparable to the number of EBRs detected in the pig genome at the same resolution (100 in the cattle lineage compared to 146 EBRs in the pig lineage, Table 3.2) suggesting that both lineages evolved with the rate of ~1.7 - 2.4 large-scale rearrangement per million years after the divergence from a common artiodactyl ancestor ~60 Mya (W.J. Murphy et al. 2005). This compares to ~1.9 (127/65Mya) rearrangements per million years of evolution within the primate lineage (Table 3.2).

The comparison of the number of genomic rearrangements between the 500 Kbp and 100 Kbp resolution sets in the primate and pig genomes indicates that there is ~689% increase in the number of rearrangements in the pig lineage while in the primate genomes there is only ~158% increase (Table 3.2). This suggests either an extremely high level of small-scale genomic rearrangements in the pig lineage or (more likely) assembly issues present at <300 Kbp resolution level in the current pig genome assembly. Both scenarios should be evaluated during further efforts on the improvement of pig genome assembly.

Table 3.2 Pig and primate EBRs at 500Kbp, 300Kbp, and 100Kbp resolutions of HSB detection.

Resolutions	Pig as reference		Human as reference	
	Pig EBRs	Pig filtered EBRs*	Primate EBRs	Primate filtered EBRs*
500	198 (100%)	146 (100%)	151 (100%)	107 (100%)
300	270 (136%)	193 (132%)	175 (115%)	127 (119%)
100	1,495 (755%)	1,006 (689%)	231 (132%)	169 (158%)
Consensus**	NA	192	NA	NA

*Indicates the number of EBRs present in the porcine and primate lineages that passed stringent thresholds (gap score >2, phylogenetic score >0.86). Percentages indicate fractions of EBRs identified at the 300Kbp resolution sets compared to 500 Kbp resolution (100%). There is an increase in numbers of EBRs observed due to higher resolution of the 300 Kbp set.

**Consensus EBRs were defined in the pig lineage as those that are consistently present in the sets of 500 Kbp, 300 Kbp and 100 Kbp, or missed only in the 500 Kbp set because of a lower resolution of this set. The consensus EBR set was used for the gene and TE enrichment analyses.

3.3.2 Transposable enrichment in EBRs

Transposable elements (TEs) comprise a large fraction of mammalian genomes and influence the structure of the genomes they have invaded. These mobile elements play an important role in shaping the genomes during evolution (Lowe and Haussler 2012). The genome analyses indicate that TEs are not uniformly distributed in genomes, but are clustered at certain regions of chromosomes (Duret *et al.* 2000, Caspi and Pachter 2006, Fontanillas *et al.* 2007, Elsik *et al.* 2009). Moreover, a significant enrichment for LINE-L1s and ERVs have been reported in tammar wallaby EBRs (Longo and Carone 2009) and Alu repeats with AAAT motif in Great Apes (Farré *et al.* 2011). These findings suggest that TEs might play an important role in chromosomal rearrangements and genome evolution by altering the state of the chromatin conformation or by

stimulating the insertion of other TEs (Lim and Simmons 1994, Craig 1996). Similarly, the cattle genome studies show the tRNAGlu–derived and LTR-ERVL repeat densities were significantly higher in artiodactyl EBRs compared to the rest of the cattle genome suggesting their contribution to formation of ancestral artiodactyl chromosome rearrangements (Elsik *et al.* 2009).

A comparative examination of densities of TEs and other repetitive sequences in the pig and artiodactyl EBRs has revealed a significant enrichment for LTR-ERV1 TEs and satellite repeats in the pig-specific EBRs compared to other intervals of the pig genome (Table 3.3). This suggests that these two families contributed to chromosomal evolution in the pig lineage. However, the current work failed to detect enrichment for the LINE-L1 elements (ancestral TEs which were shown replicating in many mammals since ~170 Mya (W.J. Murphy et al. 2005) in the porcine EBRs contrary to previous observations in the cattle and other mammalian genomes (Larkin *et al.* 2003, Larkin 2012) (Figure 3.6) where lineage-specific EBRs were found enriched for the LINE-L1 elements (Table 3.3). This suggests that LINE-L1 transposons could not be as active in the pig lineage as in other mammals and did not contribute to the genomic rearrangements in the pig genome. A recent analysis of TE activity in the pig genome indicated that indeed LINE-L1 were not active in the in the pig lineage. The fact that LINE-L1 elements were found enriched in artiodactyl EBRs in both the pig (this study) and cattle genomes (Elsik *et al.* 2009) indicates that this group of mobile elements was active in the artiodactyl ancestor and promoted at least some of artiodactyl rearrangements (Table 3.3).

Another group of mobile elements that could have promoted artiodactyl chromosomal rearrangements is SINE-tRNA-Glu. This group of elements has originated in the common ancestor or all cetartiodactyls (Shimamura *et al.* 1999) and was found overrepresented in artiodactyl EBRs in the cattle genome (Elsik *et al.* 2009, Larkin 2012) (Figure 3.7). The fact that this group of transposons was also found enriched in artiodactyl EBRs detected in the pig genome in the current study strongly supports the hypothesis that active TEs promote lineage-specific genomic rearrangements.

Table 3.3 Densities of repetitive element families found to differ significantly in pig or artiodactyl-specific EBRs compared to other parts of the pig genome. Repetitive element content is expressed as bp/10Kbp.

Repeats	Pig EBRs	Other Intervals	Artiodactyl EBRs	Other Intervals
Number of 10 Kbp intervals	2,156	257,329	210	259,275
LINE-L1	1,429	1,332	1,813*	1,332
SINE-tRNA-Glu	944*	1,050	1,239*	1,049
LTR-ERV1	210*	145	270*	145
LINE-L2	131*	256	145*	255
SINE-MIR	116*	227	102*	226
LTR-ERVL-MaLR	105*	160	122*	159
DNA-hAT-Charlie	65*	111	70*	111
Satellite	300*	229	368	229

*Found significant at FDR < 0.05

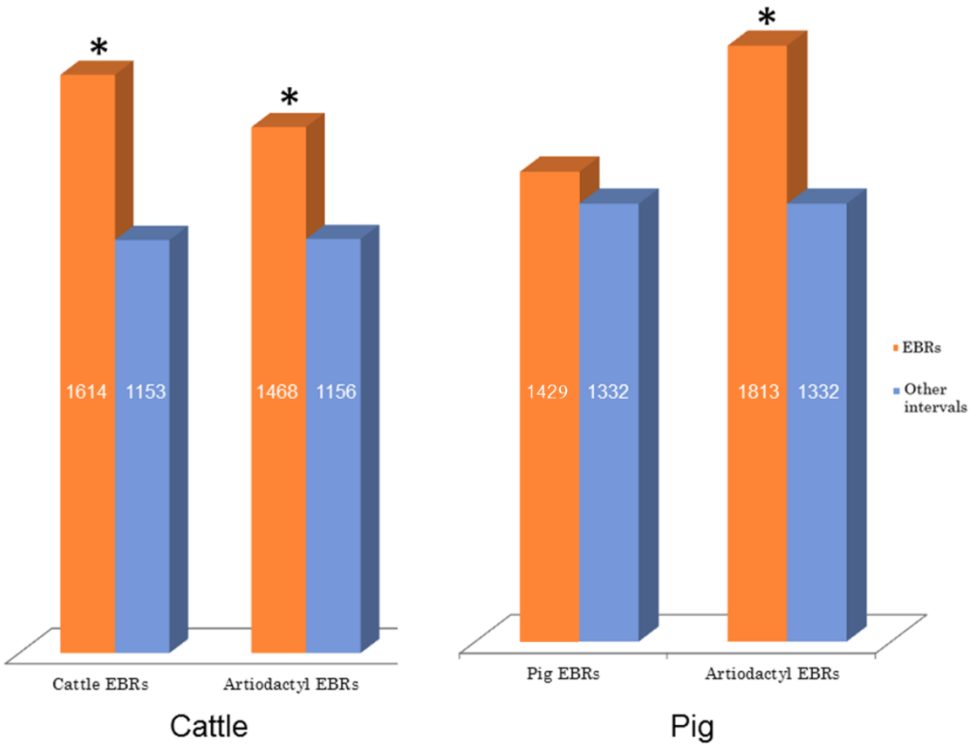


Figure 3.6 Density of LINE-L1 elements in cattle, and artiodactyl EBRs. The enrichment analysis for LINE-L1 elements in pig (this study; right) compared with cattle published data (Elsik *et al.* 2009; left), shows significant enrichment of LINE-L1 in artiodactyl EBRs. These finding suggest that this group of mobile elements was active in the artiodactyl ancestor and promoted at least some of artiodactyl rearrangements. The star (*) indicate the statistically significant result at FDR<0.05.

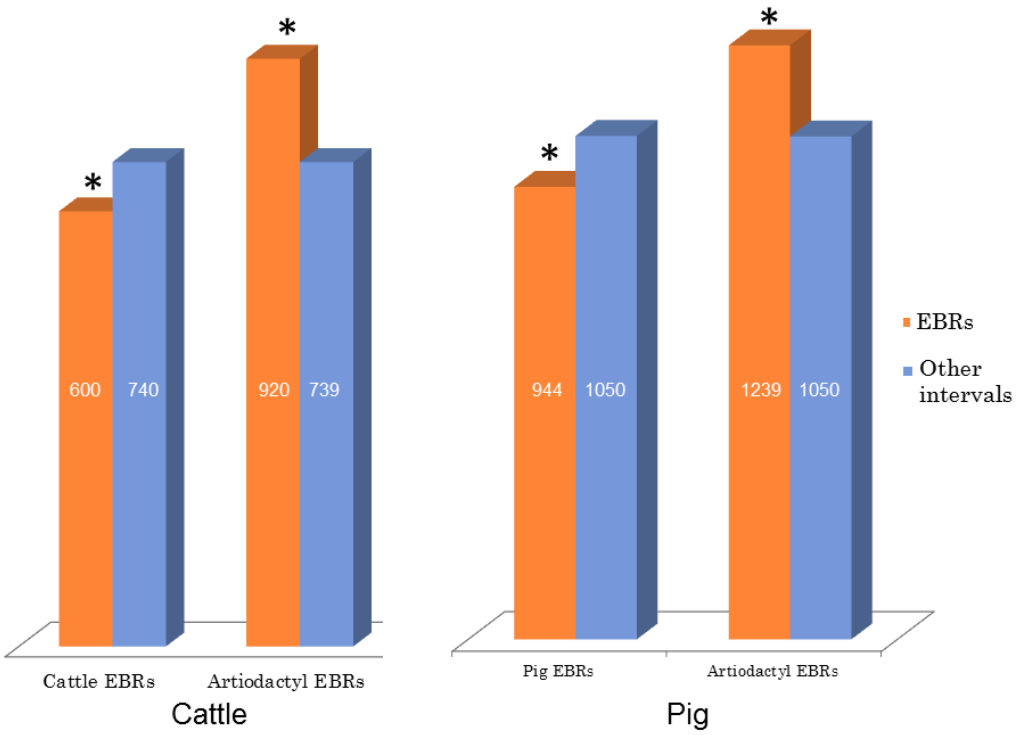


Figure 3.7 Density of SINE-tRNA-GLU elements in cattle, pig, and artiodactyl EBRs. The enrichment analysis for SINE-tRNA-GLU elements in pig (right) compared with cattle published data (Elsik *et al.* 2009) (left) shows a significantly enriched for tRNAGlu-derived SINEs elements in artiodactyl EBRs. These results suggest an active role of tRNAGlu-derived SINEs in formation of at least some of artiodactyl rearrangements. The star (*) indicates the statistically significant result at FDR<0.05.

3.3.3 Gene networks affected by chromosome rearrangements in the pig genome

The gene network enrichment within and around pig-specific EBRs was analysed to determine if genes from specific functional pathways are found preferentially in the EBRs. For this analysis, the enrichment for specific gene functions within and +/-500 Kbp from the pig-specific EBRs was analysed. The pig-specific EBRs for the pig and human reference datasets were analysed independently.

3.3.3.1 Human genome as reference

The Gene Ontology (GO) analysis of *cellular processes* categories enriched in the pig EBRs using the human genome as a reference was carried out with the Metacore [43] (MetaCore™ v.6.9 build 30881) and DAVID software (Dennis Jr *et al.* 2003). The GO analysis using Metacore demonstrates a significant enrichment for the genes involved in

[43] http://www.genego.com

sensory perception of taste, *keratinisation* and *epidermal cell differentiation* processes (FDR < 0.05; P < 0.05). The results suggest that genes involved in skin- and taste-related biological processes were likely affected by chromosomal rearrangements in the pig evolution (Table 3.4). Moreover, by looking at the KEGG taste transduction (TT) pathway it was observed that certain network signalling nodes (substrates and reactions) related to sensory perception of taste were affected (denoted with yellow stars) and underwent evolutionary changes during the course of genome rearrangements in the pig genome (Figure 3.8). Similarly, genes involved in keratinisation, epidermal cell differentiation, and keratinocyte differentiation process were found significantly affected by genomic rearrangements. All these three processes are directly connected to the keratinisation mechanism in which lower layers of the dermis become tough, insoluble and subsequently skin becomes almost waterproof; which helps to maintain water balance in the body and afford a degree of protection. A further look into the genes related to the keratinisation process and related pathways led to the identification of seven genes: GNB, IVL, LOR, SHARPIN, SPRR2G, SPRR3, and TGM1 which were located very close to the positions of chromosome rearrangement events in the pig genome. These findings suggest that certain keratinization pathway genes were affected by genome rearrangements during pig evolution, which could be connected to change of gene regulation leading to adaptations required to develop thick skin. The proximity of pig EBRs to genes involved in important metabolic pathways and processes supports previous findings of Larkin *et al.* (2009) suggesting that the EBRs are associated with genes having adaptive functions(Larkin *et al.* 2009).

Table 3.4 Gene Ontology cellular processes enrichment in pig EBRs using human genome as a reference.

No.	GO Process	P-value	Ratio
1	Sensory perception of taste	$1.9e^{-10}$*	21/49
2	Keratinisation	$9.7e^{-10}$*	20/48
	a) Epidermal cell differentiation	$1.6e^{-5}$*	23/101
	b) Keratinocyte differentiation	$2.5e^{-5}$*	21/90
5	Detection of chemical stimulus involved in sensory perception of bitter taste	$1.6e^{-4}$	8/20

Note: **Sensory perception of taste, keratinisation, epidermal cell differentiation, and keratinocyte differentiation* were found significantly enriched in pig EBRs at FDR < 0.05. Certain GO processes, such as epidermal cell differentiation and keratinocyte differentiation were linked to a wider keratinisation category, therefore were sub-grouped (a, and b) under one process.

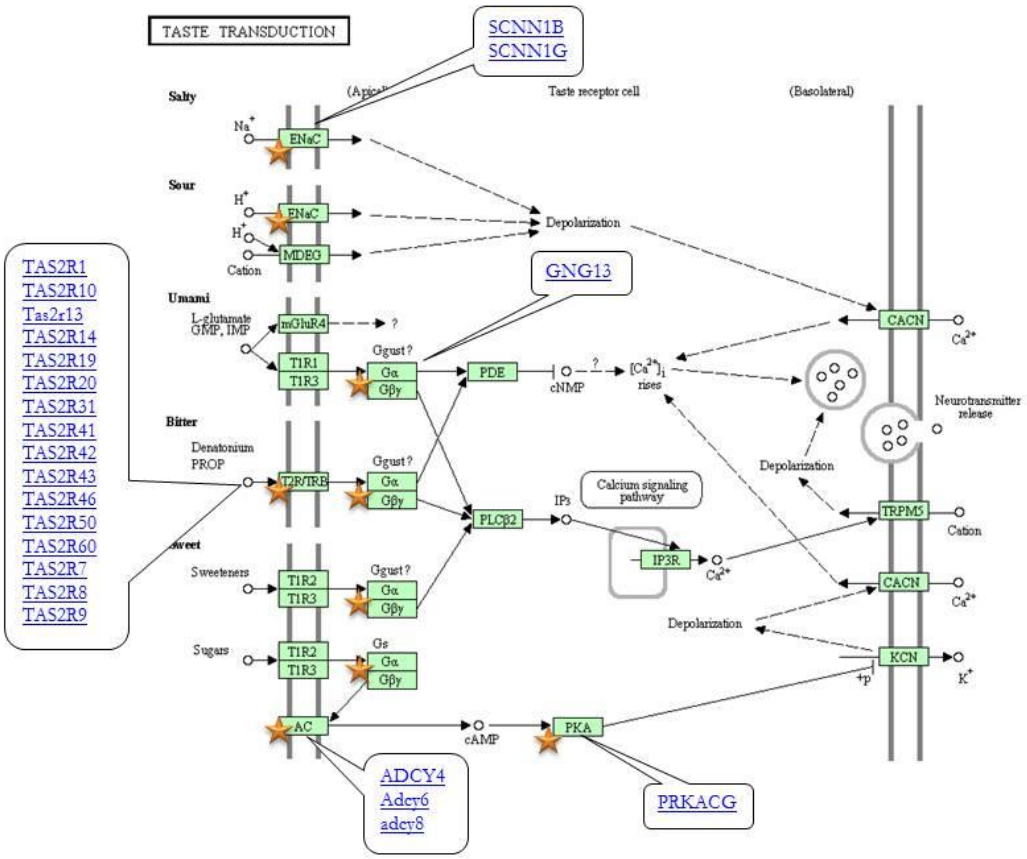


Figure 3.8 Human taste transduction pathway and gene nodes affected by pig genome rearrangements. The KEGG nodes marked with red stars are those affected by genome rearrangements. The names for the node genes found near/in the EBRs are shown in boxes.

3.3.3.2 Pig genome as reference

3.3.3.2.1 GO cellular process analysis:

The enriched functional annotations of porcine one-to-one orthologs of human genes based on the "cellular process" tree of the Gene Ontology were analysed. The GO analysis of a filtered set of orthologous genes using the MetaCore database shows that porcine EBRs and adjacent intervals are enriched for the genes involved in *sensory perception of taste (P<8.9e^{-6}; FDR<0.05)* (Table 3.6) suggesting that taste phenotypes may be affected by the events associated with genomic rearrangements in pigs. These *sensory perceptions of taste* were further studied to get a better sense of affected nodes and genes.

3.3.3.2.1 .1 Salty taste perception

Among the thirteen taste-perception-related genes present in/near the porcine EBRs (Table 3.7), the *SCNN1B* (a gene encoding a sodium channel involved in the perception of *salty* tastes) was found translocated from its adjacent paralog *SCNN1G* (an association found in the human genome in HSA16: 23.19 Mbp and other mammalian genomes) to the telomeric region of SSC10 in the current pig assembly build 10.2. However, there was a doubt that a large genome block of homology in the SSC3 would break down and recombine without one small fragment translocated to the telomeric region of SSC10 (Figure 3.9). This process could not be explained by known chromosome rearrangement mechanisms in mammals. Therefore, the translocation was further tested using the FISH technique by Dr. Katie Fowler at the University of Kent.

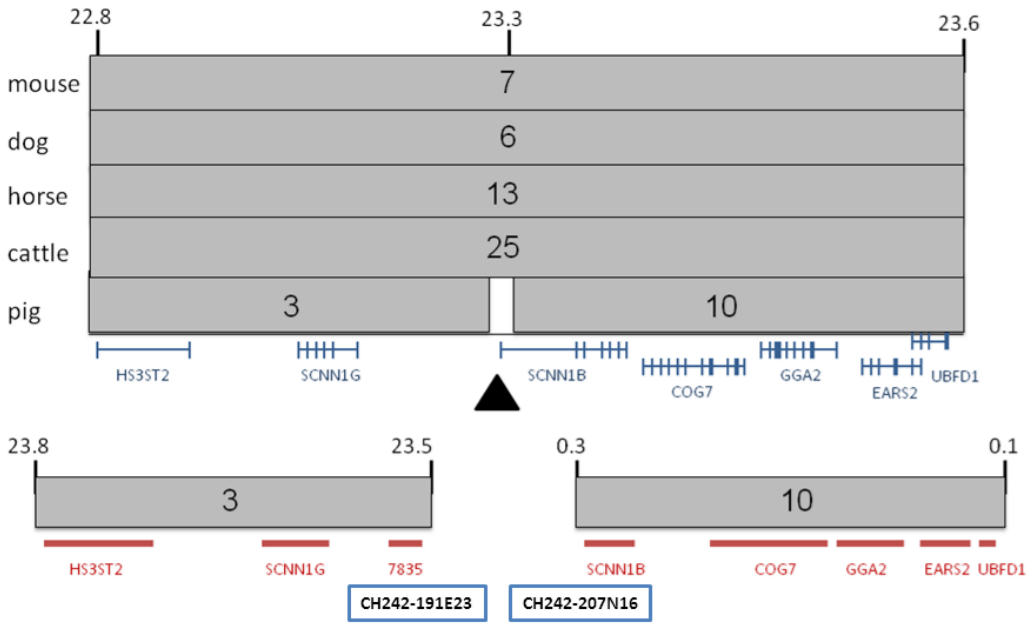


Figure 3.9 A putative pig genome rearrangement affects the *SCNN1B* gene. The upper panel shows HSA16:22.8-23.6 Mb with aligned sequences from the mouse, dog, horse, cattle, and pig chromosomes. The blue gene track shows the order of human genes that have defined orthologs in the pig genome. The black arrow indicates the position of a putative pig EBR that results in translocation of a 307Kb interval homologous to HSA16 to SSC10. This event leads to breakage of synteny in between SCNN1G and SCNN1B genes in pig. The pig SCNN1G is located in SSC10 with a partial copy (ENSSSCG00000007835) of SCNN1B found next to it. The red gene track shows the order of genes in the pig genome. The BAC clones CH242-207N16 and CH242-191E23 from the CHORI-242 BAC library assigned to chromosome 10 and chromosome 3, respectively in the pig genome assembly were used for a FISH experiment to verify an accuracy of the genome assembly in this region.

Table 3.5 Positions of the *SCNN1B* gene in genome assembly and in the pig genome (based on the FISH data).

Gene Name	Assembly position	FISH mapping results
SCNN1B	SSC10:309,239-337,906	BAC clone CH242-207N16 containing SCNN1B was assigned to SSC3, p-arm

Both porcine BAC clones (CH242-207N16 and CH242-191E23) flanking a potential genomic rearrangement between SSC3 and SSC10 were unambiguously mapped to SSC3 (Figure 3.10) by FISH. The clone CH242-207N16 contains the gene SCNN1B. These results suggest an assembly error involving SSC3 and SSC10. It is likely that the SCNN1B gene is still involved in some kind of rearrangement or duplication events in the pig genome that have complicated assembly of this region, confirming a previous studies that report that pigs have a low ability to taste salty compounds (Hellekant and Danilova 1999).

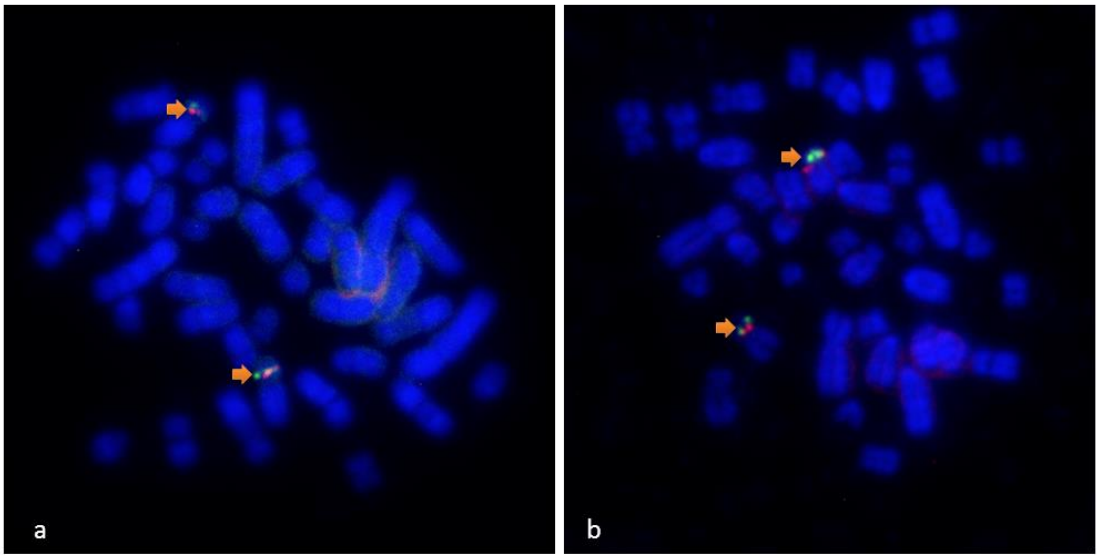


Figure 3.10 Fluorescence *in situ* hybridisation of probes CH242-207N16 and CH242-191E23 with porcine metaphase chromosomes. The partial metaphase plate shown above after FISH with CH242-207N16 and CH242-191E23 probes named 'a' and 'b' respectively. The pig chromosomes can be seen in blue. The fluorescent signal was observed only on SSC3 (highlighted with orange arrows where the sequenced-tagged BAC were hybridized (red) and clearly did not map to SSC10. The probes were re-run with a confirmed SSC3 probe labelled with green fluorescein isothiocyanate (FITC) to confirm that they map to the p-arm of SSC3.

3.3.3.2.1.2 Umami and sweet taste perception

A gene, *ITPR3,* a receptor for *inositol triphosphate* and a calcium channel involved in the perception of *umami* and *sweet* tastes was affected by the insertion of several copies of porcine-specific SINE mobile elements into its 3'UTR region, consistent with the observation of a higher density of some TEs in EBRs. The 3' Untranslated Region (3'-UTR) may contain sequences that regulate translation efficiency, regulatory regions, mRNA stability, and polyadenylation signals and influence post-transcriptional gene expression. Therefore, the insertion of TEs in 3'-UTR can directly influence the gene regulation and expression, both at the transcriptional and post-transcriptional levels(Smit 1999).

Table 3.6 Gene Ontology cellular processes enrichment in pig EBRs with pig a reference dataset.

Processes	P-values	Ratio
Sensory perception of taste	$8.9e^{-6}$*	11/23
Glutathione metabolic process	$8.0e^{-4}$	9/25
Sensory perception of bitter taste	$1.3e^{-3}$	5/9
Midbrain-hindbrain boundary development	$1.3e^{-3}$	5/9
Regulation of protein ubiquitination involved in ubiquitin-dependent protein catabolic process	$1.3e^{-3}$	5/9

*_Sensory perception of taste_ was found significant at FDR < 0.05.

Table 3.7 Genes from taste transduction pathways (KEGG) and taste transduction processes (MetaCore) found in/near pig EBRs.

Gene name	Gene functions[44]	Pig EBR coordinates	Database
DBH	Dopamine beta-hydroxylase/ monooxygenase (DBH) is a protein-coding gene mostly associated dopamine beta-hydroxylase deficiency.	1:306,934,651 - 306,985,541	MetaCore
GNG13	GNG13 (guanine nucleotide binding protein (G protein), gamma 13) is a protein-coding gene. Its function includes a signal transducer activity.	3:41,571,689 - 41,622,736	MetaCore, KEGG
ADCY6	This ADCY6 gene encodes adenylate cyclase 6, which is a membrane-associated enzyme and catalyses the formation of the secondary messenger cyclic adenosine monophosphate (cAMP). This gene prominent role in adenylate cyclase activity and protein kinase binding.	5:15,059,839 - 15,062,939	KEGG
WNT10B	WNT10B (wingless-type MMTV integration site family, member 10B) is a protein-coding gene which encodes secreted signalling proteins.	5:15,059,839 - 15,062,939	MetaCore,
TAS2R9	This gene specifically expressed in the taste receptor cells of the tongue and palate epithelia. The functional expression studies show they respond to bitter taste.	5:63,741,431 - 63,794,981	KEGG

[44] http://www.genecards.org/

TAS1R3	The TAS1R3 gene is a major determinant of differences between sweet-sensitive and -insensitive mouse strains in their responsiveness to sucrose, saccharine, and other sweeteners.	6:57,756,164 - 57,809,595	MataCore, KEGG
ITPR3	This gene encodes a receptor for inositol 1,4,5-trisphosphate, it contains a calcium channel at the C-terminus and the ligand-binding site at the N-terminus. A knockout study shows their key role in exocrine secretion underlying energy metabolism and growth.	7:34,125,342 - 34,126,061	MataCore, KEGG
ADCY4	This gene encodes a member of the family of adenylate cyclases, which are membrane-associated enzymes that catalyze the formation of the secondary messenger cyclic adenosine monophosphate (cAMP). It is expressed in olfactory cilia which may couple with olfactory receptors.	7:79,938,055 - 79,942,518	KEGG
SCNN1B	Nonvoltage-gated, amiloride-sensitive, sodium channel; controls fluid and electrolyte transport across epithelia in many organs. This gene encodes the beta subunit, and mutations in this gene have been associated with pseudohypoaldosteronism type 1(PHA1), and Liddle syndrome.	10:340,718 - 392,716	MetaCore, KEGG
TAS2R41	TAS2R41 (taste receptor, type 2, member 41) is a protein-coding gene. This receptor may play a role in the perception	18:6,766,018 - 6,823,666	MetaCore, KEGG

	of bitterness, and also a role in sensing the chemical composition of the gastrointestinal content.		
TAS2R60	TAS2R60 (taste receptor, type 2, member 60) is a protein-coding gene. This receptor may play a role in the perception of bitterness. May play a role in sensing the chemical composition of the gastrointestinal content.	18:6,766,018 - 6,823,666	MetaCore, KEGG
TAS2R40	TAS2R40 (taste receptor, type 2, member 40) is a protein-coding gene. This gustducin-coupled receptor implicated in the perception of bitter compounds in the oral cavity and the gastrointestinal tract.	18:6,766,018 - 6,823,666	MetaCore, KEGG
NPY	This gene encodes a neuropeptide that is widely expressed in the central nervous system and influences many physiological processes, including cortical excitability, stress response, **food intake**, circadian rhythms, and cardiovascular function.	18:53,339,574 - 53,398,769	MetaCore

3.3.3.2.1.3 Bitter taste perception

Eight bitter taste receptor genes were annotated in the pig genome by EnsEMBL, of which five genes were assigned to chromosomes and three were found on unassigned scaffolds. Out of five mapped bitter-taste receptor genes, four were found in/near two EBRs on SSC18 (*TAS2R40, TAS2R41, TAS2R60*) and one on SSC5 (*TAS2R9*). In contrast, the human genome contains 25 bitter taste receptor genes that originated from a series of primate-specific duplication events (Fischer *et al.* 2005). An additional annotation of bitter taste receptor genes in the pig genome was performed to identify potentially unidentified genes. Apart from eight annotated bitter taste receptor genes

annotated by EnsEMBL 9 additional intact porcine bitter taste receptor genes were found. The predicted bitter taste receptor genes are listed with pig gene names and corresponding chromosome coordinates in Table 3.8. In a case where several different pig genes had the most significant match to the same member of the TAS2R gene family from other mammals, the extensions "A, B, C" were added at the end of porcine gene names to distinguish between the porcine gene family members.

Table 3.8 Identified intact porcine bitter taste receptor genes.

Gene name*	Pig chromosome and scaffolds coordinates	In/near EBR	Annotated by EnsEMBL
TAS2R42	5:63,867,091-63,868,041	YES	NO
TAS2R20	5:63,904,140-63,905,054	YES	NO
TAS2R7A	5:63,940,163-63,941,095	YES	NO
TAS2R7B	5:63,950,624-63,951,541	YES	NO
TAS2R10	5:63,965,446-63,966,375	YES	NO
TAS2R7C	5:63,985,142-63,986,080	YES	NO
TAS2R9	5:63,976,739-63,977,674	YES	YES
TAS2R134	18:5,876,579-5,877,487	NO	NO
TAS2R41	18:7,018,806-7,019,729	YES	YES
TAS2R60	18:7,045,247-7,046,597	YES	YES
TAS2R40	18:7,266,600-7,267,764	YES	YES
TAS2R39	18:7,358,848-7,359,855	NO	YES
TAS2R38	18:8,357,518-8,358,525	NO	NO
TAS2R16	18:25,883,452-25,884,354	NO	NO
TAS2R1	GL893464.1:28,052-29,033	NA	YES
TAS2R3	GL892960.2:34,965-35,915	NA	YES
TAS2R4	GL892960.2:41,686-42,576	NA	YES

* A, B, C at the end of porcine gene names to distinguish between putative porcine gene family members.

The previous studies indicate that pigs are not so sensitive to bitter tastes and respond to higher concentrations of bitter compounds than humans (Nelson and Sanregret 1997, Hellekant and Danilova 1999) suggesting that pigs are able to use some additional food

sources that humans cannot. This makes it temping to hypothesize that this feature coupled with a fast growing rate made pigs an attractive species for domestication somewhere around 9,000 year ago (Groenen *et al.* 2012).

The review of the taste transduction network from the KEGG (Figure 3.11) shows additional genes affected by chromosome rearrangements and related to taste transduction. This demonstrates that the pig genome rearrangements tend to affect the *apical* cell membrane layer and nodes of *taste receptor* processes of the network.

3.3.3.2.2 GO Molecular function analysis

In addition to GO molecular processes GO *molecular functions* enriched in the porcine EBRs were looked into separately. The results are shown in Fig. 3.12. It was observed that there was an overrepresentation of genes related to *receptor activity* and *binding* categories in the pig EBRs. The top 5 processes were related to *adrenergic receptor activity* which is a member of G-coupled receptor protein superfamily that plays an important role in smooth muscle contraction and relaxation. These muscles contribute to vasoconstriction in many blood vessels, including those of the skin, gastrointestinal system, kidney (renal artery) (Schmitz *et al.* 1981). These data confirm other results suggesting that chromosomal rearrangements in the *Sus* lineage could have significantly contributed to various lineage-specific adaptations.

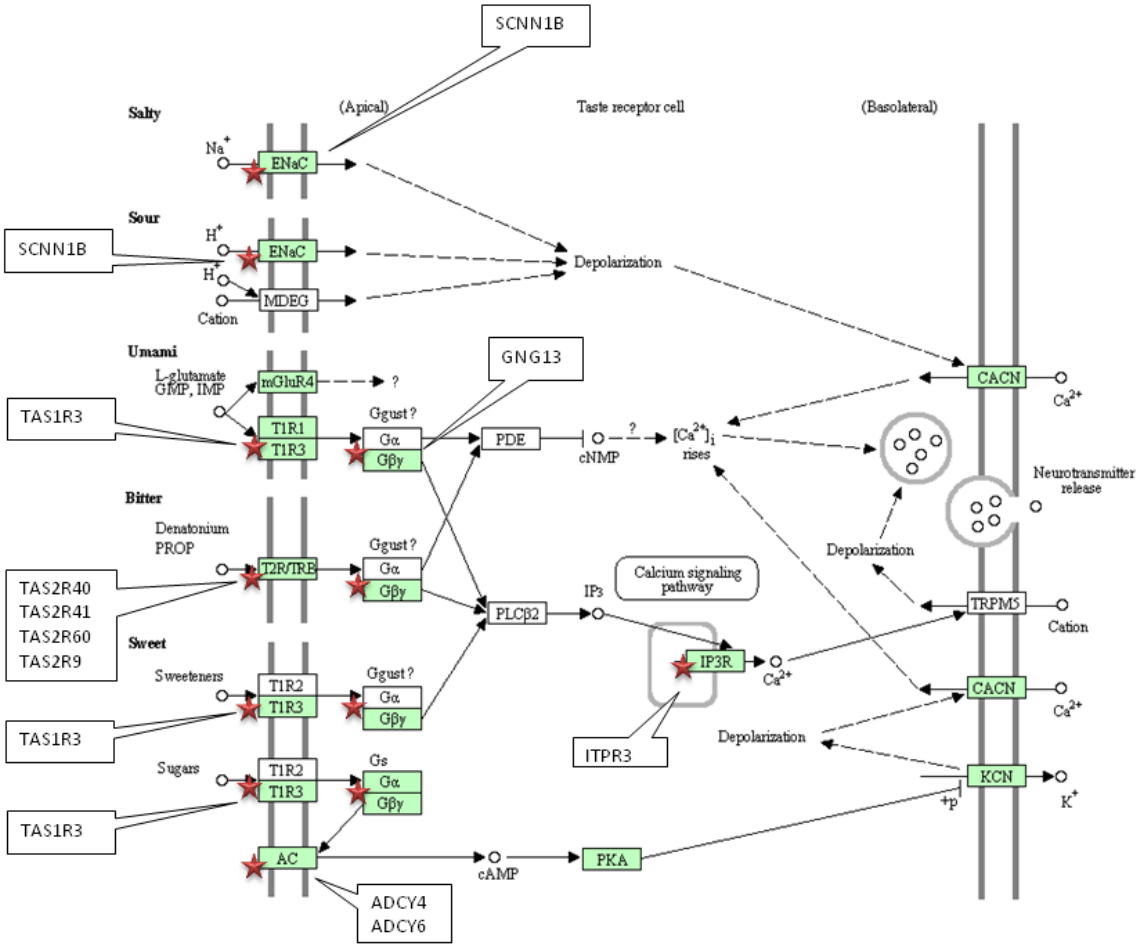


Figure 3.11 Pig KEGG taste transduction pathway. Red stars indicate nodes affected by porcine genome rearrangements. The genes from the affected nodes found near/in the EBRs are shown in boxes.

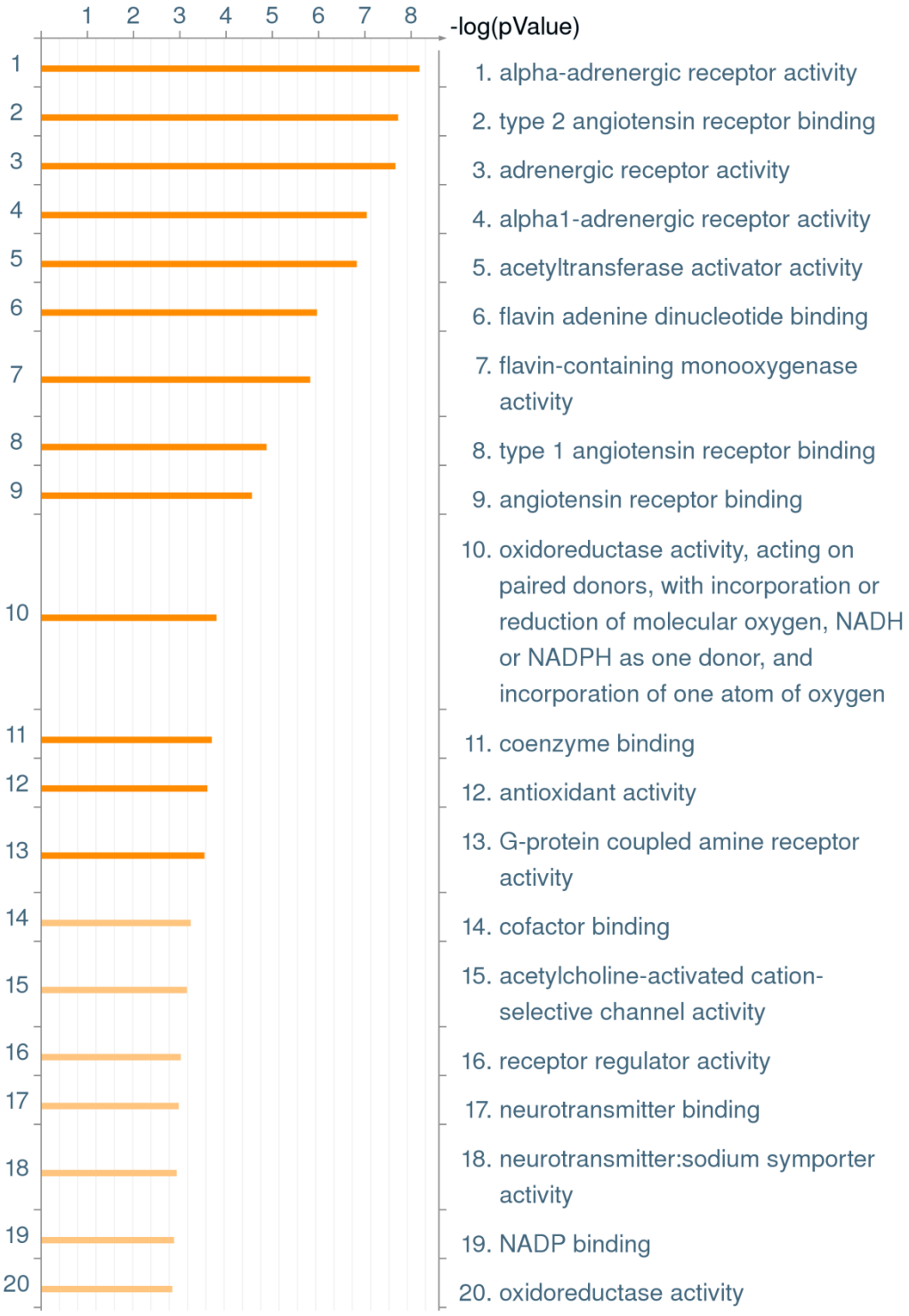


Figure 3.12 Gene Ontology (GO) molecular functions enrichment analysis in the pig EBRs with pig genes used as reference.

3.3.4 Differences in the Results of GO Enrichment Analyses using Human and Pig Genomes as References

The results of the GO enrichment analyses in EBRs using human and pig genomes as references show some differences. For instance, *keratinization* and *epidermal cell differentiation* pathway genes were significantly enriched in the pig EBRs in the analysis that used the human genome as a reference, whereas in the analysis that used the pig genome as a reference these processes were not found significantly enriched in pig EBRs. It was possible to observe these differences because of an incomplete gene annotation of the pig genome (19,094 annotated pig genes vs. 37,299 genes in the human genome). The data on TAS2R genes indeed demonstrate that the pig genome annotation is highly incomplete. This incompleteness could affect the GO analysis and make the pig genome-based analysis less statistically powerful. As such, MetaCore identifies 1,513 genes near the pig EBRs in the pig genome-as-reference set, whereas in a carefully annotated human genome–as-reference set 2,839 gene ids were recognized associated with pig EBRs. This difference in gene numbers could alter the GO enrichments resulting in different GO groups found significantly enriched. However, the occurrence of the "sensory perception of taste" biological process enrichment in both analyses provides independent confirmation for the validity of the result.

3.4 CONCLUSION

In summary, pig has been a matter of interest for many centuries due to its economical, evolutionary and medical importance. With the availability of a large number of mammalian genomes assembled to the chromosome level it is now possible to provide a basis for the identification of major chromosome evolutionary changes that contributed to biology of existing species or clades (including pigs). Using a comparative genomics approach, I demonstrated that the ancestral and lineage-specific chromosomal rearrangements in the pig genome have contributed to the formation of the pig-specific biology. For the first time EBRs were detected in the porcine and artiodactyl genomes with a high accuracy using complete pig genome assembly and sequence alignments to other genomes. These EBRs were used to reveal some adaptive changes in the pig genome that are found to be linked to the pig-specific biology.

In the study the focus was on the role of genes, gene networks, and TEs directly associated with EBRs, on sensory perception in pigs. The GO analysis has revealed how the pig lineage could have attained an omnivorous status via the metabolic adjustment for taste. The further study of EBRs in pigs and artiodactyls could influence genomic selection approaches in agriculture in order to improve pig feeding strategies. The study of the pig genome in general will empower genetic-based improvements within pork industry, which will allow fulfilling the worldwide food demands.

Summary of Novel Contributions

I identified 192 pig- specific EBRs and 20 artiodactyl breakpoints in the pig genome.The rate of chromosomal rearrangements in cattle and pig lineage were ~1.7 – 2.4 large scale rearrangements per million years of evolution. The LTR-ERV1 and Satellite repeats were found to be significantly enriched in the pig-specific EBRs. The Artiodactyl breakpoints were found to be enriched for SINE-tRNA-Glu transposable elements. We examined the EBRs regions for gene enrichments and identified that the pig EBRs was found to be enriched with the genes related to the sensory perception of taste. The genes DBH, GNG13, ADCY6, WNT10B, TAS2R9, TAS1R3, ITPR3, ADCY4, SCNN1B, TAS2R41, TAS2R60, TAS2R40, NPY from taste transduction pathways were found around pig EBRs. Seven genes, namely *GNB, IVR, LOR, SHARPIN, SPRR2G, SPRR3, and TGM1* were very close to the position of chromosomal rearrangements events in the pig genome. The GO analysis revealed how the pig lineage could have attained an omnivorous status by the adjustment of the taste transduction pathway.

In this chapter I described the importance of comparative genomics in evolutionary studies. The gene enrichments studies for the gene ontology categories showed how chromosomal rearrangements produce variations in the gene networks used in the natural selection for adaptation. Apart from that, the transposons and satellite repeats studies suggest how certain repetitive sequences have contributed to chromosomal evolution in the pig lineage. While working with pig as a reference data, I noticed chromosome evolution depends entirely depends upon breakpoints. It is crucial to study of chromosome evolution and provides answers to the evolutionary questions. In the next chapter, I devised a new method to detect EBRs in multispecies, and classify them.

4. AN ALGORITHMIC APPROACH TO IDENTIFY AND CLASSIFY EBRS IN SEQUENCED AMNIOTE GENOMES

4.1 INTRODUCTION

Chromosomal rearrangements play an important role in genome evolution and adaptation by providing a substantial source of genomic variation for natural selection, in addition to point mutations occurring in nucleic acids. Genome rearrangements alter relative positions of multiple genes from the same (inversions) or multiple (translocations, fusions) chromosomes and contribute to speciation due to the reproductive isolation of geographically separated populations (Francisco J Ayala and Mario Coluzzi 2005). The hotspots of genome evolution associated with chromosome rearragements are EBRs, regions of chromosomes where the DNA strands break and re-join due to non-allelic homologous recombination (Venturin *et al.* 2004) and end-joining processes (Critchlow and Jackson 1998, Hefferin and Tomkinson 2005).

Multiple studies have demonstrated that EBRs possess DNA features that make them distinct from other regions of the genome. EBRs are gene rich (Everts-van der Wind et al. 2004, Wind et al. 2005, Larkin et al. 2009), are associated with the repositioning of centromeres and telomeres, and contain a higher than expected frequency of segmental duplications and some families of TEs (W.J. Murphy et al. 2005, Bulazel et al. 2007, Martien AM Groenen et al. 2012). EBRs also are frequently associated with fragile chromosome sites (Ruiz-Herrera *et al*, 2006) and positions of recurrent DNA aberrations observed in certain cancers (Murphy *et al*, 2005; Darai-Ramqvist *et al*, 2008). In non-vertebrates EBRs are involved in adaptation processes causing changes in gene regulation (Iriarte and Hasson 2000). A well-known example is the reciprocal translocation between chromosomes VII and XVI in yeasts that changes regulation of the gene *SSU1,* making the mutated strand able to survive in media containing a high concentration of sulphate (Pérez-Ortín *et al.* 2002). There are accumulating evidences that EBRs play a critical role in lineage-specific adaptations in mammals as well. In cattle, an immune-related gene *HSTN* has moved to a casein genes regulatory sequence during chromosomal rearrangement. Which lead to an expression of *HSTN* gene in milk and subsequently

provide an additional immune protection for calves (Elsik et al. 2009, Danielle G Lemay et al. 2009). In addition, an expansion of β-defencin antimicrobial peptide genes in cattle has occurred in an EBR in cattle chromosome 27, likely facilitating appearance of the rumen in evolution (Elsik *et al.* 2009, Larkin 2012). A comprehensive analysis of the gene content within primates has revealed that primate EBRs are enriched for the genes involved in *inflammatory response*, *neutrophil activation*, *chemotaxis*, and *muscle contraction* (Larkin *et al*, 2009). In pigs I demonstrated that the *taste transduction* genes are overrepresented in the EBRs that are pig-specific, possibly making omnivouros pigs an attractive trarget for domestication about 9,000 ya (Martien AM Groenen et al. 2012). These examples demonstrate how the functional analysis of EBRs in the genomes of individual species or groups of related species provides an additional level of understanding of lineage-specific phenotypes, the origin of which could not be understood completely if the genetic analysis does not consider gene synteny.

With the advent of next generation sequencing technologies, genome sequencing became a trivial endeavour that provides a basis for ambitious projects aiming at sequencing large number of human individuals (1000 genomes) (Siva 2008) or species (10,000 genomes (Haussler et al.), i5K genomes (Levine 2011). While these projects have already resulted in numerous publications of individual human or various species genomes, providing important insides into the population or species-specific biology (Abecasis *et al.* 2012, Cho *et al.* 2013, Ge *et al.* 2013, X. Zhan *et al.* 2013), most genome studies do not include analyses of EBRs due to the (1) fragmented nature of the majority of the available genomes and (2) lack of automated tools to identify EBRs and assign them accurately to phylogenetic nodes. To deal with the problem of genome fragmentation, several computational tools have been developed to use existing reference genomes to predict chromosome or nearly chromosomal organization of the genomes that lack chromosome assemblies. Some of them, e.g., Reference Assistant Chromosome Assembly (RACA) (Kim *et al.*, 2013) use a combination of the comparative information and evidences from the long-range mate pair read libraries to identify the most likely organization of chromosomes, while others Hi-C methods produce genome-wide interaction maps and provide means to reconstruct chromosome architectures of fragmented assemblies (Belton *et al.* 2012). These tools provide a highly accurate prediction of chromosome structure for the *de novo* sequenced species, but underestimate the number of lineage-specific genome rearrangements. When RACA was used to reconstruct predicted

chromosome fragments of the Tibetan antelope chromosomes, it was able to achieve about 95% accuracy in the chromosome fragment reconstruction, but detected only 2 antelope-specific EBRs while the cattle genome assembled to chromosomes was reported to contain 64 cattle-specific EBRs in the same analysis (Kim *et al.*, 2013). Therefore, there is a need for a computational tool that would be suitable to perform identification of EBRs from *de novo* sequenced genomes in an automated way and assign these EBRs to phylogenetic lineages based on the EBR presence in chromosomes of species from the same clade. This tool should be suitable for detection of EBRs from a large number of genomes assembled to chromosomes or scaffolds to make use of genomes generated by G10K and other large-scale genome projects. It should also take into account the quality of individual genome assemblies when assigning probabilities of EBRs occurring in a specific lineage or clade.

Here I present the evolutionary breakpoint analyser (EBA), an algorithm that detects and assigns BRs to individual lineages or clades using several sets of pairwise HSBs, defined at different resolutions of rearrangement detection. The previous approaches of chromosomal breakpoint identifications were based on a manual extraction and classification of EBRs using certain user-defined threshold values. Such manual classification was feasible for small sets of genomes assembled to chromosomes, but when it comes to the analysis of a large amount of genomes with varied levels of assembly, it is not feasible to perform this analysis manually. The advantage of the new tools is that it parses the phylogenetic relationships of species from the NCBI taxonomy database, or alternatively uses a user-defined phylogenetic relationship. The algorithm detects the reference genome coordinates of BRs from all input HSB sets and estimates probabilities of failing to detect BRs for each target genome at each resolution (β) due low resolution of HSBs detection, assembly issues or other reasons. Using a Poisson process approximation, the algorithm estimates the probabilities that BRs from different target genomes or groups of phylogenetically related genomes have been reused in evolution (*R*). At the next step, the algorithm tests different hypotheses about the phylogenetic origin of each BR using the respective β and *R* values. Then, the tool selects the most probable hypothesis and classifies each BR for each resolution. At the final step, the algorithm compares the positions of BRs from different resolutions, identifies the narrowest and widest coordinate interval for each BR and performs classification of the final EBR set using the data from all resolutions.

EBA was applied to a set of 25 birds and 7 mammalian genomes out of which 11 were assembled to complete chromosomes and 21 were scaffold-based assemblies with N50 > 2 Mbp. The tool detected 2,066 EBRs in bird lineages at 100Kb resolution, out of which 1,796 (86.93%) were assigned to phylogenetic nodes (see Chapter 5 section 5.3.1). In mammals, 90 EBRs were detected and 86 (95.55%) assigned to individual lineages or clades.

4.2 MATERIALS AND METHODS

4.2.1 Genome datasets

4.2.1.1 Bird genome homologous synteny blocks

Five published and 15 unpublished bird genomes that were sequenced as part of the International bird genome sequencing project (Zhang *et al.* 2014) were aligned against the chicken genome sequence (ICGSC Gallus_gallus 4.0). Apart from that five reptile genomes and 3 mammalian genomes were also aligned to the chicken genome and used as outgroups in the EBR classification experiments. Details of bird genome synteny block definition are at Chapter 5 material and method section 5.2.

4.2.1.2 Mammalian homologous synteny blocks

In order to test and verify if this algorithm is capable of a proper classification of EBRs in sequenced animal genomes, the new methodology was applied to the previously published cattle genome dataset that had EBRs identified (Bovine Genome *et al.* 2009). To identify EBRs in the cattle genome, alignment was carried using human (hg19), rhesus macaque (rheMac2), dog (canFam4), mouse (mmu9), and pig (susScr3) chromosome assemblies to cattle chromosomes (UMD3.1) using the Satsuma Synteny program (M. G. Grabherr *et al.* 2010). HSBs were defined at three resolutions (≥ 100Kbp, ≥ 300Kbp, and ≥ 500Kbp) using SyntenyTracker (Donthu *et al.* 2009). Thereafter, the algorithm was applied to detect and classify EBRs in the cattle genome using the following topology: (((human, rhesus), mouse), (dog, (cattle, pig))).

Two EBR classifications were performed: (i) using EBR intervals exactly as they were defined from the HSBs sets, (ii) allowing EBR intervals to be extended by 20Kbp. The extended set could potentially allow for the detection of additional cattle-specific EBRs

in the regions of the reference genome where exact identification of HSB boundaries was complicated due to duplications or local misalignments. Translation was performed using the manually-defined and published cattle EBR coordinates from the Btau4.0 assembly (Bovine Genome *et al.* 2009) to the UMD3.1 assembly used in this analysis. This was done using the UCSC Genome Browser LiftOver tool. Out of 100 cattle-specific EBR intervals identified in the Btau4.0 assembly, 98 were successfully translated into the UMD3.1 coordinates. The EBRs found on cattle chromosome X (BTAX) were excluded from the comparison because the Btau4.0 assembly had a very incomplete BTAX assembly. The translated UMD3.1 coordinates of the remaining 90 cattle-specific autosomal EBRs were compared to cattle EBRs detected by this algorithm.

4.2.2 EBA algorithm

A novel algorithm was devised to detect, characterize and classify EBRs which are genomic intervals demarcating the boundaries between two adjacent HSBs or Syntenic Fragments (SFs) on the same reference chromosome or same scaffold, respectively. The multi-step automated EBRs detection and classification algorithm was implemented in a set of Perl scripts to detect, define, and classify the EBRs.

The custom *Perl* script identified all intervals between two adjacent SFs in the reference genome chromosomes. This was done separately for each SF set at every resolution included in the analysis. If a target genome was not assembled to the chromosomal level, only breakpoint regions (BRs) found within the scaffolds of the target species were used and classified as EBRs at the final step. All remaining BRs from all target genomes from the same SF set were coordinate-wise cross-compared for reference genome coordinate overlaps. A target genome BR that overlapped with more than one non-overlapping BRs in any other target genome(s) was treated as a *gap* and genomes containing gaps at any reference chromosome position were excluded from classification of EBRs at that position. All intervals in a reference genome chromosome between adjacent scaffolds from a single target genome that overlapped a BR in any other target genome were treated as gaps as well. Breakpoint regions were assigned to phylogenetic lineages using an updated version of the phylogenetic tree containing only the branches leading to species used in the BR analysis and the out-group species. Phylogenetic classification of BRs was performed, and later assigned EBR status using an *ad hoc* likelihood ratio approach. Since BRs identification was statistically independent for each target genome, it was possible to estimate the likelihood of any given

breakpoint classification hypothesis (H_i) with respect to the phylogenetic classification of a BR as:

$$L(H_i) = \prod_{j=1}^{n} P_{ij}(D_j|H_i),$$

Where Pij (Dj|Hi) is the conditional probabilities of occurrence of the observed data in species j (Dj), given that Hi was correct. The values of these conditional probabilities P_{ij} $(D_j|H_i)$ could be one of the following four values corresponding to four mutually exclusive events as given in the braces below:

$$P_{ij}(D|H_i) = \begin{cases} \beta_j, \\ 1-\beta_j, \\ R_{jk}, or \\ 1 \end{cases}$$

The first probability, β_j, or the probability of failing to detect a BR, was assigned when the occurrence of a BR in species *j* was expected under hypothesis H_i, but no BR was detected. The second probability (1 - β_j) corresponded to the opposite event (i.e., when the occurrence of a BR in species *j* was expected under hypothesis H_i, and a BR was indeed detected). The third probability, R_{jk}, or the probability of random overlaps between a BR in species *j* and interval of interest *k*, was assigned when no BR was expected in species *j* under hypothesis H_i, but a BR was detected. Finally, when no BR was expected or detected, a value of 1 was assigned (i.e., ignoring the very small probability of failing to detect a BR that would overlap with the interval of interest by chance).

For example, HSBs relative to a reference genome were identified in five species (1-5), three of which (1-3) are of the same order based on phylogenetic information: (1, 2, 3), 4, 5. Overlapping BRs were observed in species 1 and 2, but not in the remaining three species. In such a case scenario the algorithm will probabilistically assess three hypotheses:

i. H1: The observed BR is order-specific (i.e., due to a rearrangement that occurred in the common ancestor of species 1-3);
ii. H2: The observed BR is species-specific and was caused by independent rearrangements in species 1 and 2 (i.e., the BR was reused in evolution);
iii. H3: The observed BR is reference-specific (i.e., present in species 1-5 compared to the reference genome).

After estimating appropriate β and R values (see below) the algorithm will assess the likelihoods of the three hypotheses above as follows:

- $Pr(Data|H_1) = L(H_1) \sim (1-\beta_1)(1-\beta_2)\beta_3$
- $Pr(Data|H_2) = L(H_2) \sim (1-\beta_1)(1-\beta_2)R_{1,2}$
- $Pr(Data|H_3) = L(H_3) \sim (1-\beta_1)(1-\beta_2)\beta_3\,\beta_4\,\beta_5$

Then, the algorithm selects the most probable hypotheses and classifies each BR for each resolution. Given the hypothetical example presented above, the algorithm would calculate likelihood ratios and quantify the probabilistic support for each hypothesis. For example:

- $LR(H_1, H_2) = \beta_3/R_{1,2}$

It is expected that $R_{ij} << \beta_i << (1-\beta_i)$, and H_1 will therefore be favoured. The support will be weaker for higher R_{ij} (e.g., for gaps or widely defined breakpoint regions) and for HSB data resulting from lower quality alignments or inaccurate HSB definition (i.e., when $\beta_i \sim (1-\beta_i)$). At the final step, the algorithm compares the positions of breakpoint regions from different resolutions, identifies the narrowest and widest coordinate intervals for each breakpoint region and performs classification of the final BR set using data from all resolutions.

4.2.2.1 Probability of missing EBR (β_j)

The β_j was calculated for all of the BRs detected in species at each possible resolution by cross-validating those with higher and lower resolutions. In other words, the estimate of β_j for any intermediate resolution was calculated as the proportion of BRs that were not detected at that resolution, but were detected at both higher and lower resolutions. For the highest resolution, β_j was calculated as the proportion of BRs that were not detected at that resolution, however detected at the other two lower resolutions. For the lowest resolution, the value of β_j was extrapolated on basis of β_j values estimated at higher resolutions by means of general regression neural network algorithms (Specht 1991).

4.2.2.2 Probability of random EBR overlaps (R_{jk})

The probability of random overlap between a BR and a genome region of interest (R_{jk}) was approximated with aid of non-homogeneous Poisson process (Ross 1996). For defining the parameters, all BRs were first grouped size-wise, then the frequency rate of each class in target genomes estimated, and eventually the value of (R_{jk}) determined as:

$$R_{jk} \approx \lambda_{L(j)}(L_j + M_k),$$

where $\lambda_{L(j)}$ was the rate of occurrence of BRs from size class L in species j, L_j was the average size of BRs from class L in species j, and M_k was the size of the genome region of interest k.

4.2.2.3 "Unique", "uncertain", and "reuse" EBR classifications

After determining the likelihoods for each hypothesis of a BR classification, likelihood ratios were estimated between the first and second most likely hypotheses. These ratios were used for assigning BRs to phylogenetic branches in order to qualify them as EBRs. The EBRs with a ratio between the top hypotheses that is equal to or close to one were classified as *uncertain* due to the inability of the algorithm to select a most likely classification hypothesis. The only possible scenario to classify a BR as *uncertain* is when two or more phylogenetic nodes are indistinguishable based on the set of species used for the BR classification (see results for an example). BRs with a ratio >1 but less than a user defined threshold (T) between the top two (or more) classifications containing a set of species from distinct phylogenetic nodes was classified as *reuse,* suggesting that the EBR has likely occurred at the same position in two or more phylogenetic nodes independently. The final ratio of probabilities for these EBRs will be recalculated as a ratio between T < *average_of_all_probabilities* >1 to the highest probability outside this set. In other words for reuse breakpoints a new score was calculated by using two different approaches:

Assume there are six classifications for a single breakpoint, namely A, B, C, D, E, and F.

a) If the reuse EBRs are A and B then the new ratio is calculated between the average of reuse EBR ratio values and the maximum ratio values of non-reuse EBRs in the same cluster (C,D,E,F).

b) If the reuse EBRs are A and D then the new ratio is calculated by calculating an average value ratio of all non overlapping EBRs in the classification cluster.

The remaining EBRs (i.e., with likelihood ratios >*T*) were classified as *unique*, suggesting that they could be assigned to a specific lineage or clade. In order to reduce the false positive classifications, were removed from further analyses those EBRs whose likelihood ratios values were lying in the lowest 5% values within the respective classification group.

4.2.2.4 Generating and analysing a merged BR set

After all individual resolution datasets were analysed and individual resolution result files were generated the algorithm generated an artificial BR set by merging BR data from all resolutions with a base resolution dataset selected by the user. This was done to avoid possible misclassification of some BRs at the base resolution due to failing to detect BR in some of the target species at that resolution. However, if the lower or higher resolutions are taken into account, missed BRs might be observed and added back to the analysis. The algorithm does not change the total number of EBRs in the base resolution detected in the reference genome, but it will ensure that for each potential EBR region BRs from all target species are analysed. The *merged* set was then used to perform the calculation of new *R* values (but β values would correspond to the base resolution because the total number of EBRs does not change) and re-classification of BRs taking into account the data from all other resolutions. The final classification of the BRs from the *merged* set is expected to be the most accurate one because this set contains information about BRs that were not detected at the base resolution due to the alignment, assembly or HSB definition problems.

4.3 RESULTS

4.3.1 Algorithm implementation

The EBA algorithm was implemented as (i) a core program and (ii) a set of modules that are called by the main script at different stages to perform individual operations. The EBA package reads data folders containing files with information on HSB blocks defined for multiple target genomes. The user provides a file with the reference chromosome sizes; optional species phylogeny and β score files (see Methods). The HSB dataset submitted for the analysis can optionally be validated by the *CheckData*

module by default (user can applies an optional *-v* flag to ignore it). The module verifies presence of expected folders named after resolutions of HSB files (e.g., "10" for the 10 Kbp resolution, "100" for the 100 Kbp resolution, etc.). The module also checks input file formats. During the next cycles the package analyses data in each folder separately.

The structure of a single *EBA* cycle is shown in Figure 4.1. First putative BRs are identified in the reference genome coordinates as intervals in between two HSB boundaries that are adjacent in a reference chromosome. The algorithm ignores the putative BRs that could be present at the ends of reference chromosomes. The program gives a special status of *pseudobreakpoints* to the intervals of the reference chromosomes found in between adjacent synteny blocks that belong to different scaffolds in a target species that lacks chromosome assembly given that both breakpoint boundaries belong to either the start/end of two target species adjacent scaffolds. The BR detection step is performed by the *BreaksFinder* module. At the next step all BRs defined in the reference chromosome coordinates are checked for overlaps in reference chromosomes. Overlapping intervals are detected and recorded by the *BreaksAmongstSpecies* module. Next, the *BreaksMatrix* module checks all BRs that belong to the recorded intervals for overlaps with more than one BR in the same or different target genome. The pseudobreakpoint regions that do not overlap with at least one BR are eliminated from the further analysis. All BRs that overlap with >1 BR from the same or different target species are classified as *gaps* because the exact position of the regions in target species genomes cannot be identified. *Gap* status is also assigned to all pseudobreakpoint regions remaining in the set. The final matrix of BRs is recorded into a temporary file by the *CreateFinalMatrix* module.

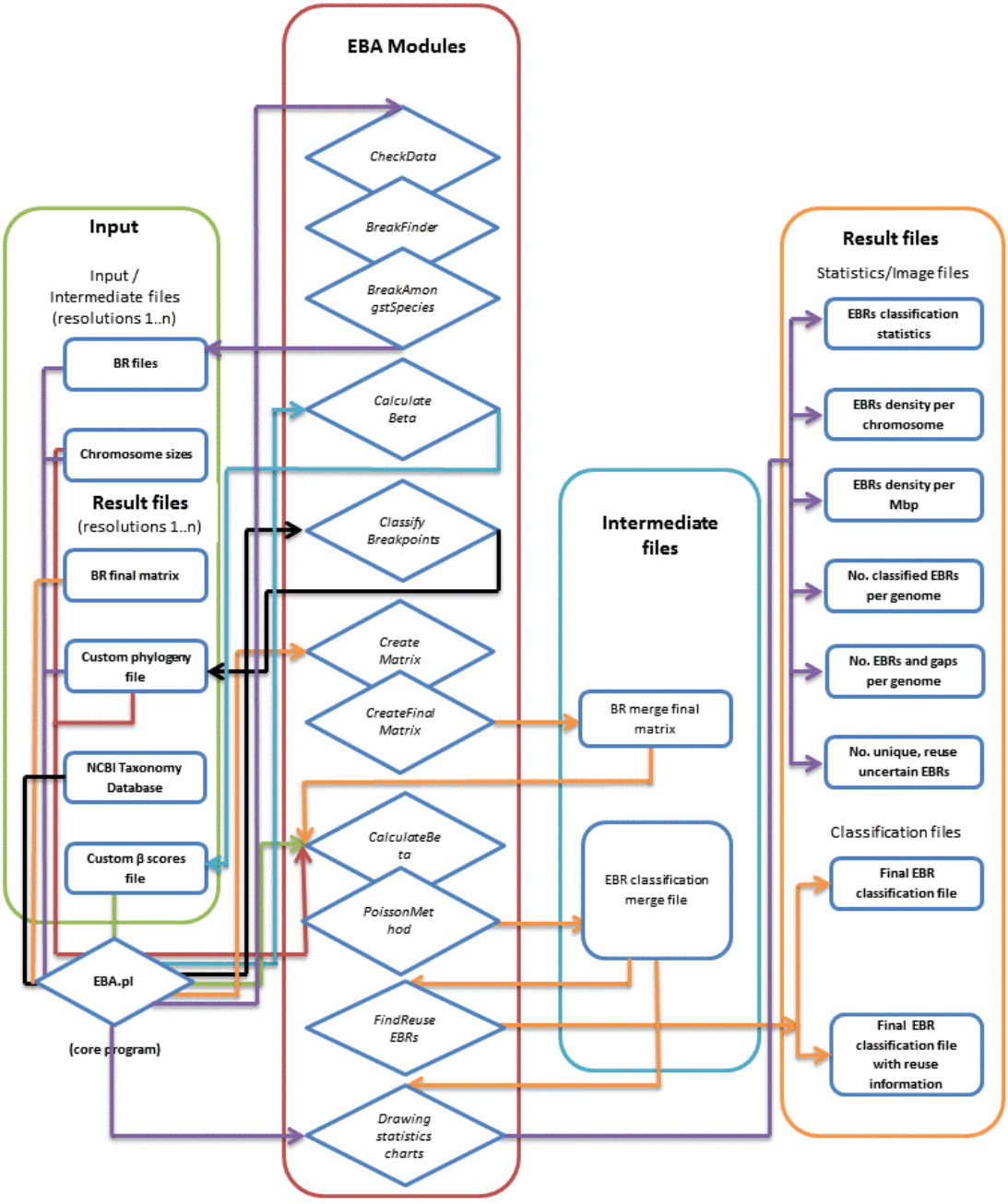


Figure 4.1 The workflow of EBA tool framework. The optional modules were not inclided in figure. The overlapping boxes are to show they used the output of conneting modules.

The *ClassifyBreakpoints* module parses the phylogenetic relationships among all species in the user dataset using the NCBI Taxonomy Database (Federhen 2012) stored in the *taxdump* folder. The module identifies groups of related species and the corresponding phylogenetic nodes from *names.dmp* and *nodes.dmp* files respectively. The results are stored in the *classification.eba* file. Alternatively, if the user wants to provide custom

phylogenetic relationships among species they need to produce a custom classification file and use the *-c* flag with *classification.eba* file to make the package using that file during the further steps. Providing a custom phylogeny file is useful when the phylogeny of species is not well established and the user would try different phylogenies to identify the one that fits data best. The classification file is used by the core program to identify suitable classification hypotheses to test for each BR recorded in the *final.eba6* file. However, to estimate the probability of a BR to belong to a certain phylogenetic node a special case needs to be considered when the BR is not observed in some of the species expected. To account for that the *CalculateBeta* module estimates probabilities of not observing a BR in each target genome at each resolution as described in the methods section. This probability is calculated and used for estimating the chances of missing BRs for the species in the dataset that are expected to contain the EBR. This is done to reduce chances of a breakpoint to be assigned to a wrong phylogenetic node with high probability when HSB files have a high fraction of missed BRs due to the alignment or assembly issues. Next the module *PoissonMethod* estimates the probability of BRs from individual genomes to overlap due to a random breakage process rather than a recent evolutionary relation of the species (see material and methods section 4.2 for a detailed description of the approach). The module tests how each classification hypothesis fits each BR in the dataset by comparing all appropriate probabilities. The hypotheses with the top two scores are compared and the ratio between them is calculated. Based on the ratio the user could judge how reliable is the classification produced by the package. The results of individual EBR classifications are stored as *result_<resolution>.final* files in the folders named after individual resolutions.

In a few cases the ratio between the top hypotheses could be equal to one suggesting that it is not possible to distinguish these hypotheses using the current dataset. The corresponding BRs will be classified as uncertain (see material and methods section 4.2). This happens when a species critical for the assignment of a BR to a phylogenetic node would have a gap at the position overlapping with the BR in another species from the same clade. In such a case the species containing gap would be excluded from the classification reducing the number of genomes from the clade used in the BR classification in a specific chromosome interval. e.g., if two carnivore genomes (cat and dog) are used to assign BRs to the carnivore clade and the dog dataset has a gap while the cat dataset has a BR at the overlapping position, then the dog data would be

excluded from BR classification resulting in the ambiguity between the lineage-specific (cat) and clade-specific (carnivore) classification of the BR.

The datasets that have gaps at a given BR position are excluded from the classification of the BR because of the lack of data. If this results in exclusion of a significant number of species from the BR classification the reliability of EBR classification will suffer. Therefore, in addition to the ratio between the top hypotheses the output file *result_<resolution>.final* contains the information about the fraction of species informative for each EBR classification.

In the cases when several hypotheses for the same BR have ratios > 1 but less than a user defined threshold *T* the *FindReuseEBRs* module recalculates the ratio for the BR and assigns it to the reuse set in addition to the individual lineages. These results are written to the *resultreuse_<resolution>.final* file.

After all data folders are analysed and individual resolution result files are generated the module *MergeResolutions* parses the base resolution (selected by the user (-p flag) and adds BRs from additional target species from other resolutions present in the reference chromosome positions where the base resolution has the BRs present (Figure 4.1). This *merge* set is then used to perform a new calculation of *R* values and re-classification of BRs taking into account the data from all other resolutions. The final classification of the EBRs from the *merge* set is expected to be the most accurate one because this set contains information about EBRs that were not detected in the base resolution due to alignment, assembly or the HSB definition issues.

4.3.2 Testing the algorithmic approach of EBR detection using a published EBR set

We applied our algorithm to a set of seven mammalian genomes to compare how precisely a set of previously published cattle EBRs (Elsik *et al.* 2009) will be detected. Out of the 90 EBRs our algorithm classified 76 (84.44%) as cattle-specific in the non-extended EBR set at the same resolution of HSBs detection (500Kbp). When I allowed the EBR intervals to be extended by 20Kbp, 86 (95.55%) of the 90 EBRs were reported as cattle-specific by our approach. In the extended set the remaining four EBRs were reported as gaps and were excluded from the EBR classification step (Figure 4.2). As expected, in the extended set a decrease in the number of lineage-specific EBRs compared to the non-extended set (up to 25% for the rhesus-specific EBRs) was

observed and also a higher fraction of EBRs was classified as reuse (8% in the extended set vs. 7% in the non-extended set) (Table 4.1).

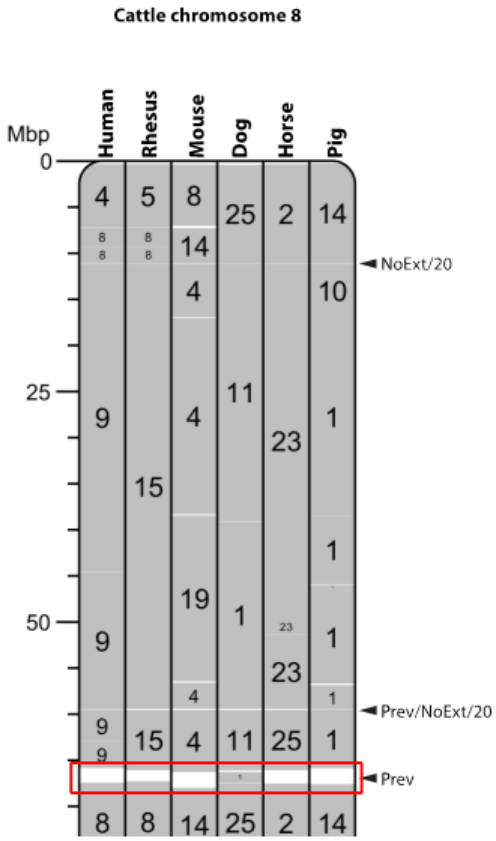


Figure 4.2 Comparison of the algorithmic approach to manually defined cattle EBR set (Bovine Genome *et al.* 2009). Cattle chromosome 8 showing the EBRs previously detected and published (Prev), newly detected EBRs not extending the boundaries (NoExt) and extending by 20Kbp (20). The red rectangle demarcates an example of an EBR classified as a "gap" by our algorithm.

While the extension of EBR intervals may help recovering additional reference-specific EBRs (11% in our set), on the other hand it leads to the overestimation of the number of reuse EBRs and to the underestimation of the number of lineage-specific EBRs. Such problematic EBRs would need to be carefully verified using FISH, PCR or other techniques. Therefore, in the computational analysis of the avian chromosomal rearrangements, it was decided to be conservative and not to extend EBR boundaries (see Chapter 5).

Table 4.1 Comparison of the not extended EBRs definition to 20Kbp the extended EBRs definition (autosomes only).

	Non-filtered		Filtered	
	Not Ext	**20Kbp**	**Ext**	**20Kbp**
Ferungulate	22	20	8	8
Dog	85	76	82	73
Rhesus	41	32	40	30
Human	49	38	49	37
Artiodactyla	24	26	19	20
Mouse	156	151	156	151
Catarrhini	33	37	30	36
Cattle	125	132	119	126
Pig	100	97	100	96
Total	635	609	603	577
Species-specific only	556	526	546	513
Reuse	42	49	42	47

In our analysis of the UMD 3.1 cattle genome assembly 35 additional cattle-specific EBRs were identified that were not reported in the Btau4.0 assembly (Bovine Genome *et al.* 2009). While these EBRs were not useful to verify the EBR-detection algorithm tracing their origin was of interest. When these EBRs were translated into the sequence

coordinates of Btau4.0 using the LiftOver tool[45], it was found that 29 of them did not match a synteny break in the Btau4.0 compared to other species. Six additional EBRs were not reported previously as EBRs in the pig genome, and therefore were classified as artiodactyl- rather than cattle-specific in the cattle genome paper. The 29 EBRs represent BRs that result from differences between Btau4.0 and UMD3.1 or differences in the methodology of genome comparisons: the cattle genome comparison (Bovine Genome *et al.* 2009) was performed using the alignment of a limited number of cattle BAC-end sequences against other species while in our analysis complete whole-genome sequence alignments were used. The six EBRs not reported previously in the pig genome are results of differences between the pig genome assembly (susScr3) used in our analysis and the pig physical map used in the cattle genome paper (Bovine Genome *et al.* 2009).

4.3.3 EBR detection in 25 bird genomes

The EBRs were assigned to different bird phylogenetic lineages using a custom classification file based on the TENT tree containing only the branches leading to species used in the EBR analysis (Jarvis and al. 2014). This algorithm was run to define and classify EBR for four resolutions (500Kbp, 300Kbp, 100Kbp and 50Kbp) of HSB detection by an alignment of 20 avian and five outgroup genomes to chicken genome. For more detail see Chapter 5.

4.3.3.1 Reuse filtration

In the bird dataset reuse EBRs were classified as the EBRs with the ratio between the first and the second classification <20 but >1. These thresholds have been chosen because these EBRs sets in this analysis were found enriched for classifications that belong to distinct phylogenetic nodes. If the ratio between first and second classification is more than the threshold (20) then the EBR classifications will be enriched for groups that tend to belong to the same clade. For more in-depth introduction of reuse breakpoint please see section 4.2.2.3. Finally, filtering was performed on those EBRs that had the ratio between the first and the second classification <45 and <50% of the studied species which were used to classify the EBR (Table 4.1).

[45] http://genome.ucsc.edu/cgi-bin/hgLiftOver

4.4 DISCUSSION

Comparative analyses of mammalian genomes have facilitated the discovery of the important features of chromosome evolution. Moreover, the comparative genome analyses get adversely affected by fragmented NGS data or partially assembled genomes. Therefore, concentrating on assemblies' limitations, EBA tools which can handle scaffold data with ease was developed. This method uses HSBs together with the phylogenetic relationship to deduce fragile region of the chromosome with high accuracy. The EBA framework is generic enough to accommodate other available information such as width size, phylogeny, chromosome size to analyse EBRs.

The genomes assembled to chromosomes have enabled the discovery of important genomic features of chromosome evolution. The critical limiting factor of genome evolution study lies in the quality of genome assemblies and analytical tools. In many cases in the current era of genome sequencing, the sequenced genomes are fragmented and not fully assembled to chromosomes, which either hinder to gain greater insight into the biology of genome evolution or generate unrealistic results. Therefore, a reliable computational method, such as EBA tool, was needed for reliable chromosomal breakpoint detection from NGS assemblies. The EBA algorithm detects EBRs and classifies them using their phylogenetic relationships. The EBA method uses homologous synteny blocks data and the Poisson approach to detect chromosomal breakpoint with high accuracy. The EBA tool also handles and accommodates other available genomic information, such as scaffolds, and phylogenetic data. The utility and effectiveness of EBA tool was demonstrated by defining chromosomal breakpoints and their classification using avian birds genomes (see Chapter 5). Moreover, with the availability of EBA tool results for chromosomal breakpoints, it will then be possible to address questions about the rates of chromosomal rearrangements and other genomic features of chromosome evolution, which may exhibit many unique adaptations.

In this study, I tested the EBA algorithm on already published cattle breakpoint classification data (Elsik *et al.* 2009). The EBA classified, and correctly assigned 84.44% of cattle EBRs to their phylogenetic nodes. In addition, once I tested the same by extending (20Kb) the EBRs size, which increased the accuracy to 95.55% for cattle-specific breakpoints. Our results show that EBA can detect and classify breakpoints with high accuracy. Therefore, it's clear from the findings that I can reach high accuracy

for both precision of detection of chromosomal breakpoints and their classification using the EBA tool. In addition to that, the EBA tool efficiently handles the challenge of fragmented assemblies due to the limited length of sequence reads. As I have shown, EBA permits the detection of chromosome breakpoints from SFs on a genome-wide basis in a *de novo* sequenced species. Our EBA tool therefore, efficiently handles the large scale genomic data and identifies classifies the EBRs amongst multiple genomes. Which, therefore, allows detailed evolutionary studies of genomes and better understanding of the unique adaptations that have occurred in different lineages (Lewin *et al.* 2009).

4.5 CONCLUSION AND REQUIREMENT

With EBA[46], this offers the geneticist a tool specially developed for non-specialists, which is user-oriented, fast, ready-to-use and standalone. In other words, the EBA is a offline tool and, thus, does not dependent on an Internet connection and a browser. The EBA tool provides a collection of modules for the EBR analysis, with an emphasis on the classification of EBRs regions. The EBA has been tested to work with Linux and Windows 7. The EBA capability has been tested on an Ubuntu Linux-based operating system with 16 GB RAM and an Intel i5 processor. It requires Perl (≥5.14.2), and Perl::GD module. For program's usability and requirements for complete novices are provided at EBA webpage[47]. The jobs that detects EBRs for 10 genomes with three different resolutions can be processed in 4 hours. Upon job completion, users can retrieve their results from Output folder. EBA tool provides an option to g*et all* intermediate files, output and figures for further analysis by the user. I provide the EBA tool in the hope that it will be useful for evolutionary study, but it is provided as is without any warranty of any kind, expressed or implied. Finally, as research continues in our lab, I will continue to make additions and updates to EBA tool.

4.6 FUTURE PLANS

The EBA provides a simple and user friendly approach to identify, and classify any number of chromosome breakpoints (with different resolutions) using their

[46] http://www.bioinformaticsonline.com/EBA
[47] http://bioinformaticsonline.com/mod/EBA/Manual.pdf

phylogenetic relationship. On-going EBA methodological developments focus more about phylogenetic distances, rate of chromosomal rearrangements, putative genome assembly error regions, ancestral breakpoints calculations and comparisons. Future development will also address the possibility to compare ancestral breakpoints and predict future possible breakpoint regions.

Summary of Novel Contributions

I developed a novel algorithm to detect and classify EBRs in genome assemblies at both the chromosomal and scaffold levels. The tool named "EBA" works with any number of genomes and resolutions in order to predict statistically significant breakpoints. It predicts and assigns EBRs breakpoints scores which were estimated using poisson process. The EBA algorithm detects breakpoint regions in genomes assembled to chromosome or scaffolds and classifies them using their phylogenetic relationships. While validating with real data, we noticed the EBA tool detected EBRs with 84.44% accuracy in a non-extended set, whereas 95.55% accuracy was observed once we extended the EBRs size by 20Kbp.

In this chapter, I developed new algorithms, and examined their accuracy on a real cattle chromosome breakpoint dataset. It has been shown to detect and consistently classify lineage- and group-specific evolutionary breakpoint regions efficiently. In the next chapter, I will test the EBA tool to identify chromosomal breakpoints in real avian genomes, chromosomes or scaffolds and then classify them.

5. COMPARATIVE ANALYSIS OF AVIAN GENOMES EVOLUTION IN BIRDS, ARCHOSAURIANS, AND REPTILES

5.1 INTRODUCTION

The non-random rearrangements of chromosomes are considered to be one of the most prominent features of animal genome evolution (Pevzner and Tesler 2003b, Larkin *et al.* 2009). The reshuffling of genome fragments in evolution still maintains large blocks of conserved synteny (chromosomal fragments) for billions of years of cumulative evolution. They are demarked by dynamic and changing EBRs from a single or both sides. Multiple evidence suggests that HSBs and EBRs appear to be evolving in different ways, and have different gene functional category enrichments (Larkin *et al.* 2009). However, to date, these conclusions have been drawn through study of the genome assemblies of sequenced mammalian genomes and may not necessarily hold the same pattern in other genomes.

Various evidence suggests that segmental duplications or repetitive DNA sequences promote chromosomal rearrangements in mammals (Bovine Genome et al. 2009, Larkin et al. 2009, Farre et al. 2011, M. A. Groenen et al. 2012). Additionally, it has been reported for mammals that lineage-specific active TEs promote species-specific rearrangements (Bovine Genome et al. 2009, M. A. Groenen et al. 2012), which point to connections between the mechanisms of chromosomal rearrangements and TE activity.

As mentioned previously, in mammals HSBs and EBRs are enriched for strikingly different functional gene content. This has been first reported by the Larkin and co-workers (2009). They have pointed out that the genes related to organismal development are preferentially located in HSBs (Larkin *et al.* 2009), whereas, the lineage-specific EBRs often affect the order and chromosomal positions of genes related to lineage-specific biology and adaptive features (Bovine Genome et al. 2009, Larkin et al. 2009, M. A. Groenen et al. 2012). These discoveries shed light on the mechanism of mammalian chromosome evolution and their potential influence on lineage-specific phenotypes.

However, there are a range of genomics features that make mammalian genome structurally different from the other amniotes genomes. Several major features that make them unique are:

a) A mammalian genome, on average, contains a large proportion of TEs and other repetitive DNA sequences (~50%) with many duplicated genes (E. S. Lander et al. 2001, Gibbs et al. 2004, Lindblad-Toh et al. 2005).
b) Mammalian genomes are organised in relatively large chromosomes with a few exceptions of micro-chromosomes (Becker *et al.* 2011, Trifonov *et al.* 2013).
c) The karyotypes of mammals are evolutionary variable with multiple inter-chromosomal rearrangements found in species (Pontius *et al.* 2007).
d) The mammalian chromosome numbers range widely from 2n=6 in Indian Muntjac to 2n=102 in Viscacha rat (Ruiz-Herrera *et al.* 2012).

The extent to which mammalian genomes are different from other amniote clades or representative of other amniotes remains an open question. With this in mind, it is quite fascinating to study the chromosomal evolution in another phylogenetic class, e.g., birds.

Birds, the only living descendants of dinosaurs (Padian and Chiappe 1998), are widely distributed all around our surroundings within a diverse group, bright showy displays of colours, distinct melodious natural songs and calls, which add an enjoyment to our daily lives. The birds' diverse plumages and behaviours are not only observed all round the world, but also play a critical role in the many food chains and webs that exist in many ecosystems. Birds transport a variety of things in our environment such as seeds, fish eggs, pollen, and certain diseases. They also control pests and work as a bio-indicators for environmental pollution. In addition to that, domesticated birds are kept for the eggs they produce, their meat, and feathers. Birds are an affordable and tasty source of protein enjoyed by people around the world. The poultry meat industry is growing by leaps and bound to fulfil the growing demand of food by the world population[48]. Despite having such a huge biological and environmental impact, their evolutionary relationships are poorly understood. The birds are a highly varied group with ~10,000 recognised species adapting to a wide variety of habitats across a broad geographic

[48] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/430104/poultry-statsnotice-23apr15.pdf

distribution (del Hoyo 1992-2013) and having a huge variety in body size and weight. The global bird speciation rates across space and time has been explored with 9,993 known living bird species using their DNA-sequence data (6,663 species and 3,330 species with no genetic data)[49] (W Jetz *et al.* 2012).

The birds show greater phenotypic variability, whereas their genomes mostly exhibit lower genome variability and are more compact (about one third the size) than of mammalian species (Gregory 2014). The repetitive DNA elements proportion in bird chromosomes is less than in mammals and constitutes only around ~15% of a bird genome (International Chicken Genome Sequencing 2004, Shedlock 2006, Zhang *et al.* 2014). A bird genome contains deletions of many duplicated gene family members, which are found in mammals and other amniotes (Zhang *et al.* 2014). The short intronic and intergenic regions in bird genomes could lead to a relatively short bird genome size (Hughes and Hughes 1995, Alekseyev and Pevzner 2009). In addition, the avian karyotypes show less variability that those of mammals (Ellegren 2010, Ruiz-Herrera *et al.* 2012), with most of them containing ~2n=80 chromosomes (D. K. Griffin *et al.* 2007). A typical avian karyotype contains 5 to 10 large chromosomes (macro-chromosomes) and a large number of small (<20 Mbp) chromosomes (micro-chromosomes).

Reconstructions of an ancestral avian karyotype based mostly on the zoo-FISH studies of macro-chromosomes suggest that most of them maintain conserved synteny in descendant genomes. These were often not disrupted by interchromosomal rearrangements over the period of avian evolution (Ellegren 2010, Skinner and Griffin 2012). Until very recently, however, a comprehensive study of avian genome at sequence level was impossible due to insufficient availability of sequenced genomes. The few available genetic maps and chromosomal assemblies of the chicken, turkey, and zebra finch genomes did provide an important insight into avian chromosomal evolution (D. W. Burt et al. 1999, Dalloul et al. 2010, Völker et al. 2010, Warren et al. 2010, Skinner and Griffin 2012). However, the nature and patterns of bird genome evolution, and their differences from mammalian genome evolution were unclear.

A comparative study using chicken, turkey, and zebra finch genomes have shown enrichment for repetitive sequences within chicken EBRs. However, the role and impact

[49] http://birdtree.org/

of TEs in the formation of species-specific EBRs was not resolved (Skinner and Griffin 2012). Volker and co-workers (Völker *et al.* 2010) have demonstrated that bird EBRs co-localize with copy number variants (CNVs). The similar trend was also reported for mammalian EBRs (W. J. Murphy et al. 2005, Larkin et al. 2009). However, in contrast to what was found in mammals and insects, association of bird EBRs with hotspots of recombination in chicken chromosomes might indicate the mechanisms of chromosome evolution between birds and other groups may differ (Völker *et al.* 2010). Due to availability of only three avian genomes at the time, it has not been possible to address the question of whether spatial organization of ancestral gene networks is maintained in bird and other reptile lineages. With the availability of multiple avian genome sequences, we are now equipped and able to test hypotheses in birds that lineage-specific EBRs alter the gene order in networks that had adaptive value and cross check the previously known evolutionary pattern in mammalian genomes.

In order to test the hypothesis, 21 available bird genomes were used, which were either assembled to chromosomes or to large scaffolds (N50 >2 Mbp). For the comparative genome analysis, we used the previously known methodologies along with our newly developed techniques, which were previously tested in mammalian genomes. By using all required methodologies, in the present study, we identified EBRs, HSBs, and stable intervals of ancestral avian, archosaurian, archosaurian/testudines, sauropsid, and amniote chromosomes (msHSBs). The rates of chromosomal rearrangements in 21 bird genomes, TEs densities in EBRs and other genome interval, and presence of genes in evolutionary stable ancestral chromosome regions were also investigated to better understand the processes that occurred during billions of years of independent bird genome evolution. Later, the gene networks, which were preferentially reshuffled during the course of bird chromosome evolution, were detected. Together the results are the first comprehensive sequence-based analysis of chromosome evolution in birds and other reptiles. These results demonstrate how the chromosomal evolution has acted upon and ruled the formation of ancestral and lineage-specific phenotypes.

Throught this chapter, the EBRs detection, scaffold handling, classification, validation, testing, and breakpoint reconstruction in target species was done by me, whereas the EBRs enrichment analyses were performed by Dr. Marta Ferre Belmonte.

5.2 MATERIALS AND METHODS

5.2.1 Syntenic fragments (SFs) detection

The chicken chromosome sequences (ICGSC Gallus_gallus 4.0) were aligned against eighteen bird genome assemblies with N50>2 Mbp (Zhang *et al.* 2014) (common cuckoo, peregrine falcon, American crow, little egret, crested ibis, domestic pigeon, hoatzin, golden-collared manakin, medium ground finch, downy woodpecker, Adelie penguin, Emperor penguin, Anna's hummingbird, chimney swift, killdeer, Pekin duck, budgerigar and ostrich) using Satsuma Synteny, a genome-wide synteny detection, program (M. G. Grabherr *et al.* 2010). In addition to that, the chicken chromosome alignments were also done with two previously published bird assemblies: turkey (TGC Turkey_2.01) and zebra finch (WUGSC 3.2.4) and five outgroup genomes: Anole lizard (AnoCar2.0), boa constrictor snake (snake 5C; (Keith R. Bradnam et al. 2013), painted turtle (Chrysemys picta bellii-3.0.1), Chinese alligator (ASM45574v1), and opossum (monDom5).

Later the pairwise alignments were further checked and cleaned from overlapping fragments and duplicated matches. The filtered set were used to define syntenic fragments (SFs) using the SyntenyTracker program (Donthu *et al.* 2009). These SFs were identified using sets of parameters that allowed the detection of genome rearrangements at ≥ 500Kbp, ≥ 300Kbp, ≥ 100Kbp in the chicken chromosome sequences. The SF sets identified were further classified as complete HSBs and SFs. The SFs found in the genomes assembled to chromosomes represent complete HSBs, whereas SFs detected in fragmented assemblies are referred to as partial synteny blocks (SFs). The HSBs and SFs were made publicly available through the Evolution Highway [50] comparative chromosome browser (Figure 5.1).

[50] http://evolutionhighway.ncsa.uiuc.edu

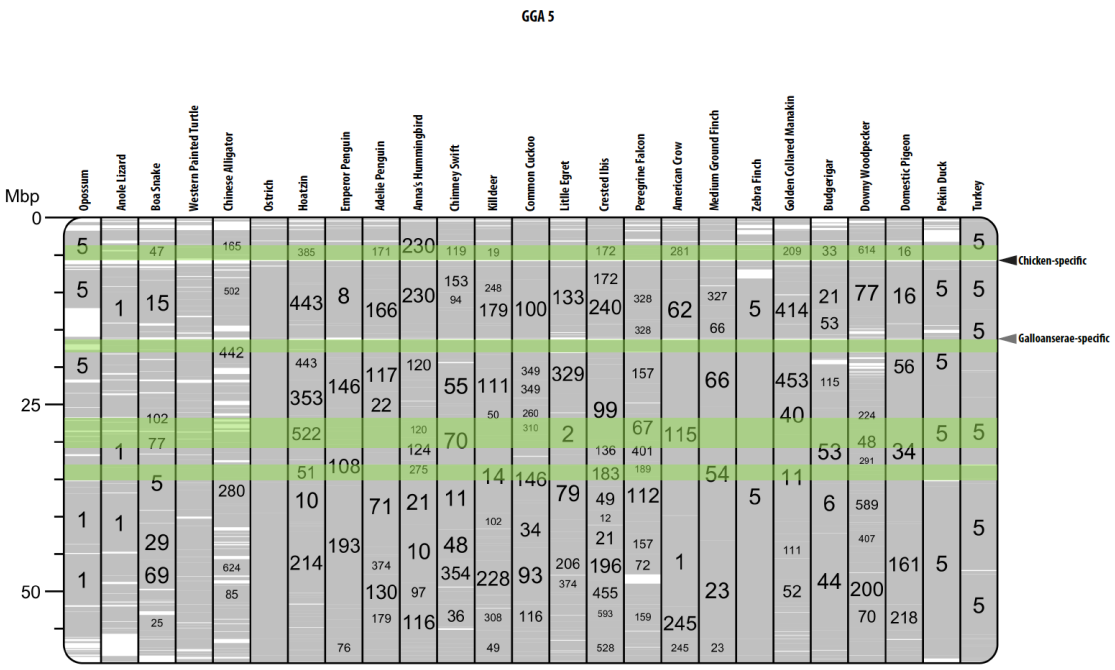


Figure 5.1 Evolutionary breakpoint regions (EBRs), syntenic fragments (SFs) and homologous synteny blocks (HSBs) identified in the chicken chromosome 5. Grey blocks define SFs in target genomes compared to the chicken chromosome, with target species scaffold or chromosome numbers indicated inside the blocks. Only the rows with genomes assembled to chromosomes (turkey, duck, zebra finch, Anole lizard and opossum) contain complete HSBs while blocks in the remaining rows represent either HSBs or SFs. EBRs are defined as white intervals in between either two adjacent SFs originating from the same scaffold in a target genome or two adjacent HSBs. Reference-specific EBRs are represented by the white intervals that overlap in all species. The arrowheads point to a reference-specific and Galloanserae-specific EBRs. Pale green boxes demarcate avian msHSBs that are longer than 1.5Mbp in the chicken chromosome.

5.2.2 Identification and classification of evolutionary breakpoint regions

Evolutionary breakpoint regions (EBRs), intervals delimited by two adjacent HSB boundaries on the same reference chromosome, were identified and classified using our algorithm. The multi-step automated EBRs detection and classification algorithm was implemented in Perl to detect, define, and classify the EBRs (see Chapter 4 section 4.3.1 for detail).

5.2.3 Identification of multispecies homologous synteny blocks (msHSBs)

In order to identify multispecies homologous synteny blocks (msHSBs), the regions of reference chromosomes that had no EBRs or *uncertain* BRs (see Chapter 4 section 4.2.3.3) detected in a set of target species, the 100Kbp SF and EBR sets were used. These higher resolution sets were used to ensure that regions of the genomes that had no rearrangements even at a relatively high level of resolution compared to the 300Kbp and 500Kbp sets were identified. The five different sets of msHSBs (avian, archosaurian, archosaurian/testudines, sauropsida, amniote msHSBs) have been defined using the above mentioned msHSBs identification approach. The msHSBs were defined based on occurrences in a selected number of the study species. The msHSBs defined in this analysis are as follows:

i. Archosaurian (birds and crocodiles)
ii. Archosaurian/testudines (birds, crocodiles, turtles, and dinosaurs)
iii. Avian (all bird species)
iv. Sauropsida (all reptile species)
v. Amniote (all species studied)

Later, the distribution of msHSB sizes in each set was tested for goodness-of-fit to measure the largest difference between the observed and theoretical distribution of msHSB. These exponential distribution analyses were done using the Kolmogorov-Smirnov test following Churchill *et al.* (1990) and Pevzner and Tesler (2003). The probabilities of each msHSB to be detected under the Poisson process were calculated.

5.2.4 Functional analysis of genes in EBRs and msHSBs

In order to perform functional analysis of genes present in the regions of interest, the following steps have been taken.

5.2.4.1 Gene selection

The gene sequence coordinates with a single known ortholog in the chicken and human genomes were downloaded from EnsEMBL using Biomart (v.74)[51]. At the time of data extraction the focus was on the chicken genes with a single known ortholog in the

[51] http://www.biomart.org/

human genome. This was done because the follow-up analyses used functional annotation of genes generated for mammalian genomes. Some genes were filtered out from the list if they were mis-assembled or had erroneous ortholog definition. These gene errors were identified with the SyntenyTracker program by building chicken-human HSBs using only the genes coordinate information. This led to the detection of "singleton" and "out-of-place" genes located in unexpected positions within HSBs or representing a single-gene HSB (see Chapter 3 section 3.2.1).

5.2.4.2 Overlapping gene selection

The remaining filtered set of genes was assigned to EBRs and msHSBs based on overlaps of gene coordinates in chicken chromosomes. In order to identify the functional categories of genes over-represented in msHSBs, only blocks larger than 1.5Mbp were considered to avoid genes that could be located in proximity to EBRs. Similarly, in order to evaluate gene functional enrichment in and near EBRs, genes that were located within EBRs or within 300Kbp from EBR boundaries were considered.

5.2.4.2 GO analysis

The DAVID (Huang *et al.* 2008) servers has been used to detect gene ontology (GO) categories for the genes that were overrepresented in these datasets. The GO terms enriched in these gene lists were examined using the DAVID functional annotation chart tool. The terms with >1.3 fold-enrichment in EBRs or msHSBs relative to all other regions on chicken chromosomes were considered significantly enriched (Huang *et al.* 2009). In order to minimise the number of potentially false positive discoveries the number of such categories were limited to a maximum of two, which in this dataset was delimited by a false discovery rate (FDR) of 6%.

5.2.5 Comparing densities of transposable elements (TEs) in EBRs and other parts of the bird genomes

The densities of TEs in EBRs and other parts of the bird genome were calculated using the following steps:

5.2.5.1 Coordinate translation

The lineage-specific EBRs identified in chicken genome coordinates were translated into the coordinates of target bird genomes using the correspondence between SFs boundary

coordinates in the chicken and target genomes. These translations were performed with a custom Perl script.

5.2.5.2 Density of TEs

In the resulting 20 target bird EBR sets and the chicken-specific EBRs, the densities of TEs (RepeatMasker[52] output, RepBase v.18) from the major families, most abundantly occurring repeats family, were calculated and compared to those in other intervals of each target genome. Following our previous publications a *t-test* with unequal variances was used to identify TE families that were enriched in the 10 Kbp genome intervals overlapping EBR positions (Bovine Genome et al. 2009, Larkin et al. 2009, M. A. Groenen et al. 2012). Local false discovery rate (FDR) critical values (Efron *et al.* 2001) were calculated to control for false positive discovery rate using the fdrtool (Strimmer 2008).

5.2.6 Density of bird-specific highly conserved non-coding elements (CNE) and genes in msHSBs

Bird-specific highly conserved elements (Zhang *et al.* 2014) were filtered to remove the elements present in coding parts of chicken genes and all mRNA sequences mapped to the chicken genome (UCSC genome browser dataset). This approach leaves only putative conserved non-coding elements (CNE). Consequently, the UCSC genome browser LiftOver tool was used to translate the CNE coordinates to the galGal4 genome assembly to make the data compatible with the HSB sets. The set of conserved elements that was not found overlapping with coding sequences after two filtering steps represents the bird-specific conserved non-coding elements (CNE) in the chicken genome. Later, the densities of CNEs and chicken genes (UCSC all known gene set) were calculated in avian, archosaurian/testudines, and sauropsida msHSBs and compared to the rest of the reference genome using the same published pipeline used to compare densities of TEs in EBRs and other genome intervals (see above).

[52] http://www.repeatmasker.org/

5.3 RESULTS

5.3.1 Syntenic fragments and evolutionary breakpoint regions

The SFs were detected at three resolutions: 100Kbp, 300Kbp, and 500Kbp by an alignment of 20 avian and five outgroup genomes to the chicken genome. At the highest resolution (i.e. 100kbp) a total of 12,761 avian SFs were detected. Out of which 914 HSBs were identified in genomes assembled to chromosomes, whereas the remaining 11,847 SFs share boundaries either with EBRs or scaffolds ends (Figure 5.1). The average and maximum sizes of an avian HSB were 3.13Mbp (± 338Kbp) and 65.84Mbp, respectively. The average size of an avian SF was 1.46Mbp (± 360Kbp) and maximum was 38.99Mbp. The number of pairwise HSBs ranged from 261 between the chicken and duck to 330 between chicken and zebra finch chromosomes. On average 89.90% of the chicken genome was covered by the avian pairwise SFs. The pairwise coverage of the chicken genome in SFs ranged from 85.74% in the chicken-to-Downy woodpecker to 91.61% in the chicken-to-Emperor penguin comparisons. Once the five outgroup genomes were added, the pairwise SFs number increased to 16,457, with an overall average size of 1.61Mbp.

The SFs from all three resolutions were used to identify EBRs. After comparing the number of EBRs in all resolutions of all studied species, the 100Kbp resolution was set for the final estimation of chromosomal rearrangement rates and the msHSB definition. The highest 100Kbp resolution was selected to avoid possible false EBR estimation and errors in msHSB detection. It also matches the ~300Kbp resolution in mammalian genomes, the resolution commonly used to define mammalian EBRs. In contrast, the 500Kbp set contained the fewest number of BRs that could be assembly errors and was selected for gene enrichment analysis in EBRs. A total of 2,066 avian EBRs were detected at 100Kbp resolution, out of which 1,796 (86.93%), with average size 18.54Kbp, were assigned to phylogenetic nodes. These EBRs cover almost 32.99Mbp (3.7%) of the chicken chromosome sequences. The 16 chicken lineage- and 42 Galliformes-species EBRs were detected. The reuse EBRs analysis revealed 211 reuse EBRs in avian genomes, which is 11.75% of the total number of avian EBRs (Table 5.1). Once the outgroup genomes were added, it increased the number of unambiguously classified EBRs to 2,589; with 486 reuse EBRs (18.77%; Table 5.1). In order to compensate for the fragmentation of some genome assemblies, the recovery

rate of reference-specific EBRs in each target genome was used to calculate the 'expected' number of EBRs in each lineage (See Table 5.1). The expected number of EBRs in each lineage was estimated using the recovery rate of the reference-specific EBRs. Thus, we used the "expected number of EBRs" in the calculation of the rates of genome rearrangements to compensate for the fragmented nature of some genomes. Moreover, for the genomes assembled at scaffold level, we analysed the possible effect of scaffold length on the EBR detection and found that these two variables do not correlate (r=-0.4960, p-value=0.06), suggesting that our EBRs estimation is not biased towards genomes with longer scaffolds.

Table 5.1 Number of detected and expected EBRs in each avian lineage at 100Kbp resolution.

	Detected no. EBRs	Expected no. EBRs
Species-specific		
Anas platyrhyncos	113	130
Aptenodytes forsteri	38	43
Calypte anna	92	99
Chaetura pelagica	45	56
Charadrius vociferous	25	27
Gallus gallus	16	16
Columba livia	102	114
Corvus brachyrhynchos	37	40
Cuculus canorus	106	110
Egretta garzetta	40	47
Falco peregrinus	86	99
Geospiza fortis	35	40
Manacus vitellinus	35	44
Meleagris gallopavo	255	255
Melopsittacus undulates	181	199
Nipponia nipon	39	42
Opisthocomus hoazin	39	50
Picoides pubescens	147	184

Pygoscelis adeliae	47	55
Struthio camelus	124	137
Taeniopygia guttata	47	53
Clade-specific		
Galliformes	42	42
Galloanserae	15	23
Trochiliformes + Apodiformes	2	2
Ciconiiformes	3	3
Passeroidea	14	14
Passeroidea + Corvoidea	19	21
Passeriformes	16	17
Sphenisciformes	4	4
Passeriformes + Psittaciformes + Falconiformes + Piciformes + Ciconiiformes + Sphenisciformes + Charadriiformes + Opisthocomiformes	2	4
Non-galloanserae	11	13
Non-galloanserae + non-columbiformes	1	1
Neognathae	9	14
Avian	9	14
Total avian EBR	1796*	
Reuse		211 (11.7%)

* Total number of EBRs does not include the reuse EBRs, because these were counted as lineage or order specific in corresponding lineages.

5.3.2 Rates of chromosomal rearrangements

In order to estimate rates of chromosomal rearrangements, a published bird TENT tree was used (Jarvis and al. 2014). The observed number of EBRs in bird genomes varies from 16 (chicken) to 181 (budgerigar) with an average of 48 EBRs (Table 5.1). In order to estimate global rates of chromosome rearrangements in birds and other lineages the expected number of EBRs in each node were normalized by the node branch length in million years (MY) (Table 5.1). The rearrangement rates were defined as: (i)"low", < 1.22 EBRs/MY, (ii) "intermediate", 1.22-2.13EBRs/MY, and (iii) "high", > 2.13 EBRs/MY based on the average number of EBRs and 95% confidence intervals (Figure

5.2). The turkey EBRs and rearrangements rates were excluded from the further analysis because of the large differences in number of observed EBRs at 100Kbp and 300Kbp resolutions. This large number of observed differences points towards the probable large number of local miss-assemblies present in the turkey genome at <300Kbp resolution. Such differences were not observed in other genome which suggests that the other assemblies were reliable enough to perform chromosomal rearrangement studies at the resolutions selected. The estimated bird chromosomal rearrangement rates were identical with the previously reported rearrangement rates by us (Zhang *et al.* 2014) with minor differences because of inclusion of some additional outgroup genomes (Chinese alligator, painted turtle, and boa snake). Due to the inclusion of the alligator genome it was found that the divergence of the ancestral bird lineage from crocodiles was accompanied with one of the lowest rates of rearrangements in bird genome evolution (0.098 EBRs/MY), whereas the branch leading to Neognathae (~5.73 EBRs/MY) contains the highest rate of rearrangements.

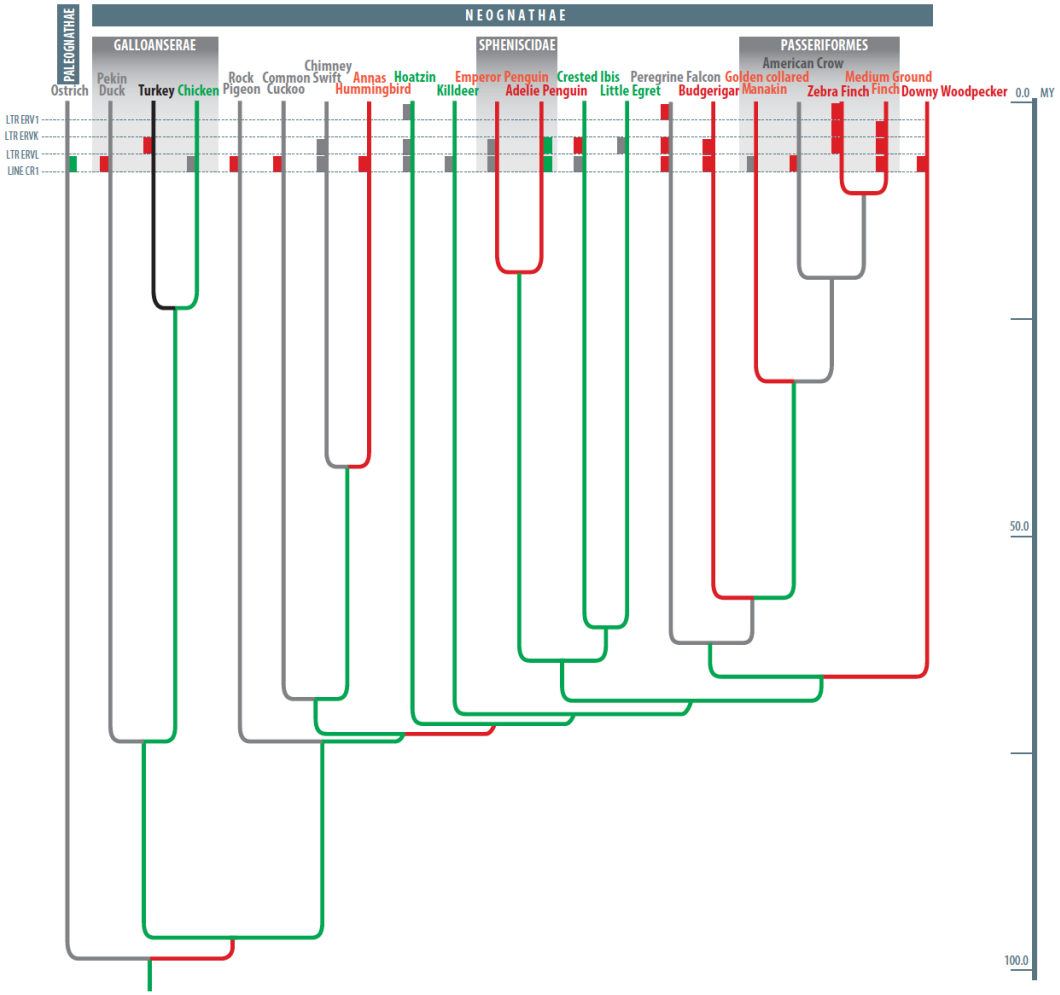


Figure 5.2 Chromosomal rearrangement rates in avian lineages. The phylogenetic tree is based on the total evidence nucleotide (TENT) tree (Jarvis and al. 2014). Rearrangement rates (RR) for the 100Kbp resolution dataset are plotted on each branch. Green lines represent low rates (<1.22 EBRs/MY); grey, medium rates (1.22-2.13 EBRs/MY), and red high rates (>2.13 EBR/MY). Turkey rate was omitted from the calculation of these intervals (black line). Red bars represent a significant enrichment of TEs (LINE-CR1, LTR-ERVL, LTR-ERVK or LTR-ERV1) in species-specific EBRs; green bars show negative association of TEs with species-specific EBRs and grey bars indicate elevated numbers of the TE families in species-specific EBRs.

The cytogenetic studies, performed by Griffin *et al.* in 2007, on bird chromosomes suggested that birds have a stable karyotype. Therefore the chromosomal rearrangements affecting the chromosome numbers have occurred sporadically (D. K. Griffin *et al.* 2007). The estimation of inter-chromosomal rearrangement rates in each species relative to the chicken reference genome has been made using the SFs dataset.

The study shows a high rate of chromosomal rearrangements in Emperor penguin (2.23 EBRs/MY) and Adelie penguin (2.82 EBRs/MY). In addition, Emperor penguin and Adelie penguin genomes contain many lineage-specific interchromosomal EBRs (18 and 20, respectively). In comparisons, Passeriformes exhibit significantly fewer interchromosomal EBRs than other bird lineages (t-test=-2.9224, p-value=0.0096). However, the global rearrangement rate in Passeriformes is significantly higher than in other bird lineages due to a large number of interchromosomal rearrangements (t-test=2.48, p-value =0.029). In contrast to the previous cytogenetic study (DK Griffin *et al.* 2007), ostrich (2n=80) seems to have a large number of interchromosomal rearrangements (26) and an intermediate rearrangement rate (1.38 EBRs/MY).

5.3.3 Density of transposable elements in avian EBRs

The lineage specific EBRs were tested for enrichment of the abundant group of TEs (>100bp on average in the EBR- or non-EBR-containing non-overlapping 10Kbp genome intervals). Due to a comparatively small fraction of TEs in bird genomes (4-19%) compared to mammalian genomes (~50%) only four families of TEs: LINE-CR1, LTR-ERVL, LTR-ERVK and LTR-ERV1 passed this threshold in at least one of the bird genomes. A significant enrichment or elevated number of at least one of these TEs groups was observed in the majority of avian lineage-specific EBRs (Figure 5.2). However, no significant enrichment for these TEs families was found in ostrich and Adelie penguin-specific EBRs. Moreover, the analysis of ostrich EBRs and TEs shows a significant negative association of the EBRs with LINE-CR1 elements (p-value=1.1e-6). Similarly, the LINE-CR1 and LTR-ERVL elements also show a negative association with Adelie penguin EBRs (p-value=0.005 and p-value=0.0002, respectively).

The potential for a correlation between the rates of chromosome rearrangements and the total number of TEs in individual bird genomes was investigated. When all species and all TEs families were considered there was no correlation detected (r=0.23, p-value = 0.314). Similarly, no significant correlation was found when four highly represented families of TEs were considered r=0.24, p-value=0.301). However, the correlation coefficient increases to 0.66 (p-value =0.001) if only LTR-ERVL and LTR-ERV1 were analysed in the all bird genomes. When Passeriformes were analysed separately from other species a strong correlation was observed (Figure 5.3) between the total number of Passeriform TEs and chromosomal rearrangements rates in the same species (r=0.96, p-value= 0.033).

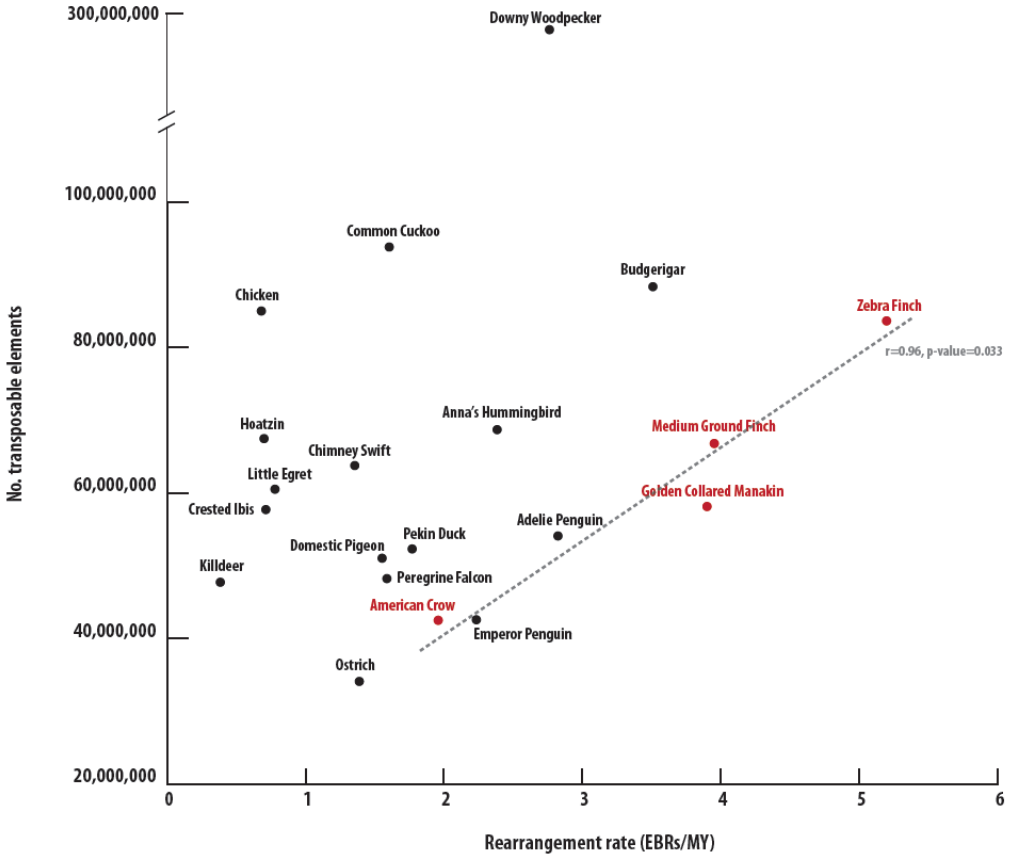


Figure 5.3 Number of transposable elements (TEs) and rearrangement rates (EBRs/MY) in bird species. Red dots show Passeriformes, while black dots show the other species. The dotted line depicts the correlation of rearrangement rates and TEs in Passeriformes. Turkey was not included in this analysis as explained in the text.

5.3.4 Multispecies HSBs

The 100Kbp resolution was used to define the msHSBs using the pairwise SFs dataset. An msHSB was defined as a chromosomal region, which was not interrupted by EBRs across all or a subset of genomes. By applying this approach, 1,746 avian msHSBs were detected, covering 76.29% of the chicken genome. The longest avian msHSB with the size of 4.81Mbp was found on GGA1. The chicken genome coverage in the msHSBs was reduced to 53.33% in 1,514 amniotes msHSBs after the addition of five species (Table 5.2). The analysis of msHSB length distribution approximates to an exponential distribution in four of the msHSB sets (Additional data section 5). These distributions are consistent with a random distribution of evolutionary chromosomal breakages in the

genomes. However, some msHSBs were detected (6 amniote, 4 reptile, 3 archosaurian/testudines, 3 archosaurian and 5 avian msHSBs) that were longer than the maximum length expected from a random distribution of EBRs in the corresponding genome sets. Out of these it was noticed that one amniote msHSB on GGA1, two sauropsid msHSBs on GGA6 and GGA3, one archosaurian/testudines msHSB on GGA3, one archosaurian msHSB and one avian msHSB on GGA1 were significantly larger than expected if all EBRs were distributed randomly (p-value <0.05).

Table 5.2 Multispecies Homologous Synteny Blocks (msHSBs) present in different subsets of the species studied.

	Avian	Archosaurian	Archosaurian & Testudines	Sauropsid	Amniote
Total length (Mbp)	765.21	665.03	651.74	545.95	534.92
Coverage of chicken genome (%)	76.29	66.3	64.98	54.43	53.33
Max length (Mbp)	4.81	4.46	4.67	4.67	4.17
Expected max length (Mbp)*	3.52	3.25	3.06	2.73	2.79

* The expected maximum msHSB sizes were calculated assuming an exponential distribution as $L[\gamma + \ln(n+1)]$, where L is mean fragment length and $\gamma = 0.5772$ is Euler's constant (Churchill *et al.* 1990).

5.3.5 Bird-specific conserved non-coding elements (CNEs) in msHSBs

A total of 215,998 bird-specific CNEs with sizes longer than 10bp have been discovered in the chicken genome (Zhang *et al.* 2014). They were significantly enriched in all msHSBs sets (p-value < 3e-12). The ratio between the number of CNEs in msHSBs and other genome intervals ranged from 1.42 for reptile and amniote msHSBs to 1.72 for avian msHSBs. Similarly, the density of chicken genes were also checked in msHSBs, which followed the opposite trend with msHSBs having significantly fewer genes than other chromosome intervals with the ratios ranging from 0.58 for avian msHSBs to 0.74 for reptile and amniote ones (p-value < 3e-12).

5.3.6 Functional analysis of genes within msHSBs

Comparison of gene ontologies (GO) associated with genes located in the evolutionary stable (msHSBs) and dynamic (EBRs) regions allowed identification of preferred gene functional categories located in EBRs and msHSBs regions of avian and reptile genomes. In order to perform these analyses 11,153 genes with a single known ortholog in the chicken and human genomes were extracted from BioMart[53]. The gene sets were filtered for misassembled regions and unreliable orthology, which finally produced 10,830 genes in this reference dataset.

Table 5.3 msHSBs >1.5Mbp in each subset of species with the total number of genes in each msHSB set.

	No. msHSBs >1.5Mbp	Coverage of chicken genome (%)	No. genes	Percentage of total genes used (10,830)
Birds	85	18.12	1315	12.14
Archosaurian	67	14.07	1024	9.45
Archosaurian + Testudines	62	13.17	959	8.85
Sauropsid	45	9.16	706	6.52
Amniote	44	8.03	676	6.24

[53] http://biomart.org/

In order to perform gene enrichment analysis only those msHSBs which were longer than 1.5Mbp in the chicken genome were used. They covered from 8.03% to 18.12% of the chicken genome in amniote and avian msHSBs, respectively and contained 6-12% of genes from the reference list (Table 5.3). The GO categories which were significantly enriched and passed the FDR threshold in each of the five msHSB sets were identified (Figure 5.4). Once the avian, archosaurian, and archosaurian/testudines msHSBs were checked for GO term enrichments, we found a significant enrichment for the *regulation of gene expression* and *biosynthetic processes* in avian, archosaurian, and archosaurian/testudines msHSBs. Interestingly, these processes were also found enriched in sauropsida or amniote msHSBs but did not pass the FDR threshold. The GO terms enriched in avian, archosaurian and archosaurian/testudines msHSBs show a very strong correlation but not all of these categories reached the FDR threshold in archosaurian or archosaurian/testudines msHSBs sets ($r=0.95$, p-value<0.0001 for avian and archosaurian comparison and $r=0.86$, p-value<0.0001 for avian and archosaurian/testudines). However, this correlation pattern fails when the bird, sauropsida and amniote msHSBs were compared ($r=-0.40$, p-value$=0.22$; $r=-0.65$, p-value$=0.17$, respectively).

Figure 5.4 Gene Ontology (GO) terms enriched in four sets of msHSBs. Green boxes show a fold enrichment >1.3 while red boxes depict a fold enrichment >2. White crosses inside boxes show categories that passed the FDR significance threshold of 6%.

Later, highly enriched (>2 fold) GO categories in the msHSBs were analysed (Figure 5.4). The *development of primary sexual characteristics* category was found highly enriched in all msHSBs sets but passed the FDR threshold in avian, archosaurian and archosaurian/testudines msHSBs only. These 17 genes were found in 14 avian msHSBs. These msHSBs were distributed across 12 chicken chromosomes, one of them, the bone morphogenetic protein receptor 1B *BMPR1B* gene involved in chondrogenesis and growth of wings was found present only in an avian msHSB. The avian and archosaurian msHSBs show a significant enrichment of *retina development in camera-eye type* category. This category contains nine genes in six avian msHSBs and found distributed across six chicken chromosomes. The avian, archosaurian and archosaurian/testudines msHSBs sets were found significantly enriched for the *appendage* and *limb development* categories. But the FDR threshold was passed by the avian set only. The nineteen genes of these categories were distributed across 12 avian msHSBs in eight chicken chromosomes. Out of nineteen genes, five genes, namely *SHOX*, *DLX5*, *DLX6*, *HOXA11*, and *BMPR1B* were in the msHSBs found only in bird genomes

5.3.7 Functional analysis of genes within or around EBRs

In order to perform the gene enrichment analysis in EBRs, only enriched GO terms (fold-enrichment >1.3; FDR<6%) with genes found in >1 EBR region were considered as significant. Using this approach we most likely detected the gene networks affected by multiple chromosomal rearrangements rather than functional enrichments in individual genome intervals.

Table 5.4 Gene Ontology terms enriched in EBRs

EBR classification	GO term	No. genes	Fold-enrichment	FDR (%)	No. EBRs
Downy woodpecker	Histidine metabolism	6	8.12	0.51	5
Adelie penguin	Regionalization	7	6.48	0.83	7
	Anterior/Posterior pattern formation	6	7.78	1.23	6
	Pattern specification process	7	4.89	3.59	7
Killdeer	Transmembrane transport	9	4.03	1.47	4
Little egret	Neurological system process	6	7.67	0.97	5
	Feeding behaviour	3	34.23	3.94	3
Manakin	Cytokine-cytokine receptor interaction	6	4.91	4.61	3
Peregrine falcon	Metal ion transmembrane transporter activity	11	3.26	2.12	8
	Synapse	8	4.03	3.28	7
	Nucleoside-triphosphatase regulator activity	10	3.09	5.49	9
	Cation channel activity	10	3.38	3.03	7
Budgerigar	Forebrain development	12	2.74	5.47	11

A total of 13 significantly enriched GO categories have been detected in species-specific EBRs of seven bird species (Table 5.4). The GO term *regionalisation* and *pattern specification* were found enriched in the Adelie penguin-specific EBRs, which include seven genes (*NR2F2*, *LHX1*, *KIF3A,* and *GBX2* among them). All these seven genes were found in/near seven EBRs and were distributed amongst five reference chromosomes. Similarly, the Budgerigar-specific EBRs study indicates that certain genes, namely *NOTCH1*, *DRAXIN*, *GATA2,* and *NUMB* are involved in *forebrain development* processes which tend to reshuffle during chromosomal rearrangement. Some genes *NPY1R*, *APLP2,* and *BSX*, related to *feeding behaviour* and *RGS9BP*, *TECTA,* and *P2RX4*, related to *neurological system process* have been found co-localised with the little egret-specific EBRs. The GO term enrichment analysis for falcon-specific EBRs shows enrichments of three GO terms: *metal ion transmembrane transporter activity, synapse* and *nucleoside-triphosphatase regulator activity*. Further descriptive gene analysis of each process indicates 11 genes for *metal ion transmembrane transporter activity* distributed among six reference chromosomes, eight genes for *synapse* in six chromosomes and 10 genes for *nucleoside-triphosphatase regulator activity* in seven chromosomes. The six genes (*ALDH6*, *HAL* and *CNDP1* among others) related to *histidine metabolism* process was found co-localise with downy woodpecker-specific EBRs.

5.4 DISCUSSION

The 26 sequenced avian genomes were made available recently due developments and advancements in sequencing technologies (Mardis 2008) and subsequent initialization of various large-scale projects (Genome 2009). This work used a set of avian (21) and reptile (4) genomes to perform a comprehensive study of chromosome rearrangements in birds.

A similar alternate pattern of faster and slower rates of chromosomal rearrangements (Figure 5.2) in avian species as was earlier reported for mammals (W. J. Murphy et al. 2005, Larkin et al. 2009) was observed. For instance, the split between Paleognathae and Neognathae ~100 MYA was accompanied with an enhanced rate of chromosomal rearrangements in the Neognathae ancestral lineage. The observed low chromosome rearrangement rate in Neognathae is similar to the rates observed in the eutherian mammals until the Cretaceous-Tertiary (K-T) boundary (W. J. Murphy et al. 2005). In

the majority of Neognathae clades, chromosome rearrangement rates were lower in comparison to mammalian orders. In contrast, the fraction of reuse EBRs in case of birds (11.7%) is higher when compared to mammals (8.0%; (Ma *et al.* 2006, Larkin *et al.* 2009) and is similar to the earlier estimates on a lower number of bird species (Skinner and Griffin 2012).

If the smaller genome sizes in birds (~1.05 Gbp compared to ~3.0 Gbp in mammals) is considered and observed slow ancestral chromosomal rearrangements rates, then it can perhaps be hypothesized that almost all bird clades had a stable genome organization which was needed to be maintained in order not to affect important gene networks and phenotypes. This hypothesis could be checked through the following studies: (i) by relating the global rates of rearrangements in birds to diversification rates and phenotypes, (ii) by comparing the rates of rearrangements to densities of TEs, and (iii) by observing signatures of gene network enrichments in evolutionary stable and dynamic chromosome intervals.

It was noticed that the lowest rates of lineage-specific chromosomal rearrangements in those avian species which retain most ancestral phenotypes like chicken (Romanov *et al.* 2014) and hoatzin (Mayr and De Pietri 2014) while highly diverged species/clades such as penguins and budgerigar contain more rearranged chromosomes. The highest level of chromosome rearrangements was found in the lineage leading to Passeriformes, falcons, parrots and woodpeckers. From the seven species belonging to this clade in this analysis, it was found out six that have shown fast rates of chromosomal rearrangement with the fastest rate found in zebra finch and medium ground finch genomes. It makes it tempting to compare finches with murid rodents because of the highest levels of genome rearrangements found in both groups in birds and mammals, respectively (Bourque *et al.* 2004, Zhao and Bourque 2009). In comparison to mammals, it was noticed that there was a noteworthy difference in terms of percentage of intra- and inter-chromosomal rearrangement. For instance, ~89-100% of rearrangements in finches are intra-chromosomal and in rodents ~16-36% is inter-chromosomal. Conversely, there was a high correlation between the diversification rates (especially when diversification rates are high) and the rates of chromosomal rearrangements in birds ($r=0.92$, p-value=0.025). These results suggest that the link between the stable bird karyotypes and ancestral phenotypes does exist. The reproductive isolation, adaptation (see below) and speciation could be ensured and appear eventually as a result

of intra-chromosomal rearrangements. In other words, the derived phenotypes and speciation may appear without significant disruption of karyotype structure (F. J. Ayala and M. Coluzzi 2005).

The lineage- and order-specific EBRs are enriched for TEs and other duplicated sequences that were active at the time of lineage/order formation in mammals as reported by many earlier works (Schibler et al. 2006, Larkin et al. 2009, M. A. Groenen et al. 2012). Repeats could have promoted the process of chromosomal rearrangements through the non-allelic homologous recombination process (NAHR) (Bailey *et al.* 2004). Therefore if bird EBRs are also enriched for TEs, a lower fraction of TE in avian species (~4-20%) compared to mammals (~50%) could be accountable for the evolutionary stability of bird karyotypes. In this analysis 19 out of 21 bird lineage-specific EBR sets were either significantly enriched (p-value<0.05; FDR<0.05) or had an elevated fraction of at least one of the highly abundant families of TEs (Figure 5.2). It was observed that there was a significant negative association between the two sets of lineage-specific EBRs (budgerigar and ostrich) with the density of TE elements, a pattern consistent with ancestral TE families in mammalian EBRs (M. A. Groenen et al. 2012). This suggests that some unknown families of TEs may contribute toward the genome rearrangements in ostrich and budgerigar. The TEs contributed to the genome rearrangements in birds and mammals which comply with past studies which stated that in birds, LTRs and LINEs were significantly enriched in EBRs but not in the HSBs (Skinner and Griffin 2012). From in-depth analysis of the clade with the higher rate of chromosomal rearrangements (Passeriformes), a highly positive correlation (r=0.96, p-value =0.033) of the total number of TEs in the passeriform species with the corresponding rearrangement rates was found, which suggests a connection between these two characteristics in Passeriformes. A similar trend was found in the two penguins and ostrich genomes, but not in other species where the number of observed EBRs is significantly lower than would be expected from the total number of TEs, suggesting a likely negative selection of chromosomal rearrangements at the germ cell level. This is strongly supported by the woodpecker genome data. For instance, in the woodpecker has a large expansion of LINE-CR1 elements comprising ~19% of the genome. The woodpecker-specific EBRs were highly enriched with LINE-CR1 elements. The chromosome rearrangements rate in woodpecker is high (2.78 EBR/MY) but comparatively lower than what would be expected from the number of TEs (Figure

5.3), suggesting a strong negative selection against chromosome breakage in various parts of woodpecker chromosomes. Besides, it was discovered that those parts of avian chromosomes that are devoid of EBRs (msHSBs) are highly enriched for bird-specific CNEs and not having many genes in comparison to other regions. This confers a rationale behind the negative selection for evolutionary breakage in these gene deserts enriched for regulatory sequences. Eventually, it could be suggested that TE families have taken part in the formation of lineage-specific EBRs in birds. A smaller fraction of TEs in the bird genomes in comparison to mammals coupled with the selection against chromosomal rearrangements in some lineages or genome intervals might be responsible for more stable karyotype in birds compared to other lineages.

Identification of several long regions in amniote chromosomes were shown to be non-randomly maintained in evolution (Larkin *et al.* 2009) and were enriched for the genes responsible for development of organ systems in the human genome (Larkin *et al.* 2009). The current study puts emphasis on the functional gene categories overrepresented in various ancestral reptile and bird msHSBs. These msHSBs covered 9-18% of chicken chromosomes, contained about 6-12% of chicken genes with well-established orthologs in the human genome. In avian, archosaurian, and archosaurian/testudines msHSBs, a significant correlation for GO categories relevant to *regulation of gene expression* and *biosynthetic processes* was found, however, the correlation was not present in the reptile and amniote msHSB sets. This suggests that the ancestral archosaurian/testudines lineage went through re-organisation of genes which are either controlling gene expression or taking part in biosynthetic processes; and that some ancestral syntenies are maintained in the descendant lineages. The slow chromosomal rearrangement rate occurred at the split of archosaurian and testudines affirms this hypothesis. From all the GO terms enriched in all msHSBs sets, only the *development of primary sexual characteristics* has passed the FDR significance threshold in avian, archosaurian, and archosaurian/testudines msHSBs. A *BMPR1B* gene resided uniquely inside an avian msHSB (GGA4: 57,866,704-59,398,610). Besides having an effect on ovulation (Onagbesan *et al.* 2003), it has a foremost function in the process of digit condensation in mammalian limbs and bird wings. In one recent study, the arrest of digit I formation in bird wings was attributed to the low expression of *BMPR1B* and *SOX9* genes, which could explain the appearance of the three digit wings in birds (Welten *et al.* 2005). A presence of bird-specific CNE situated 100bp upstream of *BMPR1B* was detected in this work, whereas

the closest CNE present in all vertebrate species was placed 34.5Kbp upstream, indicating that the presence of *BMPR1B* in an avian msHSB may be related to change of its regulation and expression.

Retina development in camera-type eye GO category was found to be significantly enriched in both the avian and archosaurian msHSBs but not in any other msHSBs. Crocodiles and birds share a similar organization of the retinal centrifugal visual system that varies from other reptiles and possibly incepts in their archosaurian ancestor (Ferguson *et al.* 1978). Remarkably, *SOX2* gene, found in an archosaurian msHSB, is accountable for reprogramming of non-neural retinal pigment epithelium cells to differentiate towards retinal neurons in chicken embryos (Ma *et al.* 2009). Thus, this work links the morphological similarity of the retina in birds and crocodiles to their corresponding identical genome regions containing the genes involved in retina development.

Whereas GO category related to *limb development* was found to be enriched in avian, archosaurian, and archosaurian/testudines msHSBs, only in the avian msHSBs this particular category was highly enriched and passed the FDR threshold. The five genes namely *DLX5, DLX6, BMPR1B, SHOX, and HOXA11* were identified that differentiate the bird msHSBs set from its evolutionarily closest archosaurian group. Therefore, these genes are prime candidates to contribute to bird-specific limb phenotypes. The *BMPR1B* gene is responsible for the formation of the three-digit bird limb as described above. *DLX5*, a representative of the distal-less (*DLX*) family of homeobox-containing genes is found to be expressed in the apical ectodermal ridge guiding the outgrowth and patterning of limb mesoderm (Ferrari *et al.* 1995). This gene is also involved in early feather bud development and is actively expressed in the bud epidermis. Activity of *DLX5* in chicken embryos could be responsible for feather fusions and loss (Rouzankina *et al.* 2004), supporting the fact that *DLX5* is one of the crucial genes accountable for the origin of feathered animals. A bird-specific CNE 1.9Kbp upstream of the *DLX5* was found and could be accountable for a distinct expression of the gene in birds and non-feathered species. An idiopathic short stature and skeletal malformation which was regularly seen in human patients with Turner, Leri–Weill and Langer syndromes (Tiecke *et al.* 2006) happened as a consequence of mutation in *SHOX*, another homeobox-containing gene. *SHOX gene* is primarily related with both developing cartilage and muscle elements of limbs in chicken, however such function is not present in the case of human muscle formation (Tiecke *et al.* 2006).

Over-expressed *SHOX* in the chicken embryos tends to enhance the length of skeletal elements significantly, demonstrating that *SHOX* modulates length of bones (Tiecke *et al.* 2006). Eventually, the *HOXA11* gene expressed in zeugopod territory (Zeller *et al.* 2009) during the proximodistal limb bud development aids in the formation of the ulna and radius bones.

Additionally, this work focused on genomic regions and gene networks that define species-specific characteristics (Larkin et al. 2009, Danielle G. Lemay et al. 2009, M. A. Groenen et al. 2012) by performing the GO analysis in lineage-specific EBRs. There were 13 GO categories found to be significantly enriched in the lineage-specific EBRs of seven bird species (Table 5.4). Five out of 13 GO terms may be associated with adaptive changes in the bird lineages. Adelie penguin's EBRs were enriched for genes connected with *pattern specification* and *regionalization* including the *NR2F2* and *KIF-3* genes. Both genes are likely to be expressed during the spinal motor neuron development (Lutz *et al.* 1994) and left-right determination (Hirokawa *et al.* 2009) as demonstrated earlier. Spatial reorganizations in the genome could change the regulation and expression of these genes (Marques-Bonet *et al.* 2004), which in turn, made the body structure adaptation in such a way that Adelie penguins may able to swim deeper and spend less energy than other penguins (Culik *et al.* 1994). Another case is the GO terms overrepresented in the little egret-specific EBRs. They cause the re-shuffling of *feeding behaviour* related genes, including the spatial reorganization of a genomic region having the gene *NPY1R*, wherein mutations could connect it to carbohydrate intake (Elbers *et al.* 2009) in humans. Hence, it is tempting to state that *NPY1R* reorganization is associated with the specific diet of egrets. The budgerigar-specific EBRs are enriched with the genes whose functions are connected with *forebrain development*. These parts of the brain are responsible for producing vocalizations in vocal-learner bird species and are called 'vocal brain nuclei'. Parrots surprisingly have unique neuronal connections in comparisons to other vocal-learners (songbirds and hummingbirds) (Jarvis 2004). Three genes (*NUMB*, *NOTCH1* and *DRAXIN*) out of those related to the forebrain development in budgerigar EBRs were found to be responsible for neuron differentiation (Wakamatsu *et al.* 1999, Islam *et al.* 2009). According to analysis of the current data, these genes will be primarily the topmost candidates for the appearance of neurological features in the parrot's forebrain. Peregrine falcon EBRs were found to be enriched with genes responsible for *cation channel activity* and *synapse,* in similar line, with

the nervous system- and sodium ion transport-related genes evolving rapidly in two falcon species founded on the basis of latest whole genome-based comparison (Xiangjiang Zhan *et al.* 2013).

5.5. CONCLUSIONS

In summary, this work demonstrated that multiple genome synteny comparison is a powerful tool to understand the chromosomal rearrangements and their impact on evolution. In addition, it also enabled detection of ancestral and lineage-specific genome-rearrangements as well as evolutionary stable chromosomal intervals in birds and other reptiles. The study also demonstrated that chromosomal breakage in reptiles and birds is not random but is connected to multiple genome features including the number of TEs, regulatory sequences and a relative gene order. It was found that the rates of genome rearrangements over evolutionary time in birds are not constant but vary significantly, in agreement with earlier findings in mammals (W. J. Murphy et al. 2005). They correlate positively with diversification rates (W. Jetz *et al.* 2012), but on average are lower than in mammals. The lower density of TEs in birds is most likely an important factor in part responsible for the evolutionary stable avian karyotype. Apart from this some other factors like selection against EBRs in the genome intervals containing genes and regulatory sequences related to some pathways established in the common ancestor of birds, crocodiles, and turtles or formation of micro-chromosomes (D. W. Burt et al. 1999, Burt 2002) should also be considered. Moreover, with the availability of a larger number of genomes assembled to the chromosomal level, this approach coupled with ancestral genome reconstruction (Ma *et al.* 2006) will provide a basis for the identification of major chromosome changes that contributed to the formation of existing species or clades.

Summary of Novel Contributions

The comparative analysis of avian genomes indicates that there are lower rates of chromosome evolution as well as the presence of a lower fraction of transposable elements in bird genomes compared to mammals. The study revealed enrichment for GO terms related to regulation of gene expression and biosynthetic processes in bird, crocodile and turtle HSBs. These findings point towards the order of these genes being established in the archosaurian/testudines ancestor about 300 million years ago (MYA)

and then maintained in the descendant species. The archosaurian HSBs were found enriched for genes that are responsible for the similar retina structures in birds and crocodiles, while the avian HSBs contain genes involved in the bird skeleton and limb development. The analysis of gene content in and around avian EBRs revealed enrichments for genes likely to be related to lineage-specific phenotypes, such as GO terms related to regionalisation in the Adelie penguin and forebrain development in the Budgerigar. The lower fraction of TE in avian species (~4-20%) compared to mammals (~50%) could be accountable for the evolutionary stability of avian karyotypes.

In this chapter I showed the importance of the EBA tool, and its application in evolutionary research. Apart from that, I also reported some of the noble findings and application of EBRs on avian genome evolution (see above paragraph). In the next chapter, I discuss my research findings and limitations.

6. GENERAL DISCUSSION AND CONCLUSION

6.1 INTRODUCTION

Recent developments in sequencing technology have led to a breakthrough in studies of chromosomal evolution and the rapidly developing field of comparative genomics. The research presented in this thesis demonstrates the role and application of computational techniques into modern comparative genomics and genome evolution studies. Keeping the technological development in mind, it is obvious that large scale genome wide analyses have fundamentally changed the current perspective in which evolutionary problems are considered. In other words, these large scale genome wide analyses have permitted the explanation of long-standing evolutionary biology problems such as how reshuffling of genomes and evolutionary forces works for the same. While on the other hand, they have also raised many challenging questions and avenues of research. The study of amniote (mammalian, avian and non-avian reptile) evolution has come to a new era, where genomic sequences are used to explore the evolutionary perspective**.** The complete amniote genome sequences generated by whole genome sequencing (WGS) provide genomic sequences that empower us to clarify key aspects of early amniote evolution. The comparative genome analysis of fully or partially assembled genomes has demonstrated that many key genomic elements play a vital role in adaptive changes that occur over the course of evolution. This research helps exploration and understanding of the molecular mechanisms behind chromosome evolution and their adaptive consequences. Moreover, there is a wide scepticism not only regarding genome sequence data, but also regarding the outcomes of actual computational analyses, which are believed to be often erroneous due to various genome sequencing approaches, genome assembly algorithms, and inaccurate phylogenies. Once the amount of data and dimension increases, the problem to be solved becomes more complex, and therefore more sophisticated analytical tools are needed to address this complexity (Chapter 4). This is especially true in the case of comparative analyses (see Chapter 3 and 5), as the most prevalent drawbacks of comparative genomics are the misleading results.

Comparative genomics, as applied in our amniote chromosomal rearrangements study, is a powerful method for high-resolution, cross-species genomic inference. This approach is not only used to detect the boundaries of conserved synteny, but also to

determine the chromosomal rearrangements and breakpoints within genomes. Such powerful a high-resolution comparative genomics approach was applied to analyse mammalian species (described in Chapter 3) and avian chromosome evolution (described in Chapter 5). In this thesis, the comparative genomics approach has been applied to amniote genomes, representing a period of 300 million years of divergent evolution (Saitou 2013). In the course of this work, the pig genome was examined first and analysed the chromosomal rearrangements and breakpoints to understand their impact on pig evolution. Ultimately, the chromosomal rearrangement events were explored at 100Kb, 300Kb, and 500Kb resolution and cross compared among these resolutions to get the most accurate result. The breakpoints inferred across the multiple genomes characterise the EBRs or the genomic regions where breaks happened in evolution. These EBRs provide the ultimate resource for attempting to understand the adaptation and speciation mechanism at the genomic level. In addition, it explains how an amniote chromosome evolves and contributes to lineage-specific phenotypes. The comparative genome analysis strategy was applied to mammalian genomes and proceeded step-wise to avian genomes. This approach was successful because of the genomic resources already available from the Genome 10K consortium[54], as well as the pipeline and tools that have been developed to detect multispecies EBR using computational resources available at the IBERS, Aberystwyth University, UK.

This discussion of my work is intended to detail what these computational analysis results suggest more broadly, with reference to early mammalian and avian evolution. However, before discussing the limitations of this work, future directions and conclusions of the results presented in this thesis, I will discuss some important considerations that recur throughout this thesis.

6.2 COMPARATIVE GENOMIC APPROACHES TO AMNIOTES GENOME

Whole genome sequencing has created a watershed of research opportunities in biology that have helped to elucidate genome evolution and understand the mechanism of adaptation. For these purposes, comparative genome analysis is the primary method which has been used for investigation. Moreover, the species chosen are crucial, as some

[54] https://genome10k.soe.ucsc.edu/

inferences are dependent upon their positions on the tree of life. Thus amniotes have become a focus of great attention in comparative genomics, because the taxon comprises all extant land-dwelling vertebrates. It is therefore clear that amniote genomes will play an essential role in elucidating the genetic background of phenotypic evolution.

Chapter 3 describes the distribution of EBRs within the pig genome and demonstrates how chromosomal rearrangements produce variations in the gene networks likely used by the natural selection for adaptation to environment. The comparative study of seven mammalian genomes has provided a glimpse into the dynamic nature of gene networks by discovering the EBRs linked to the pig-specific biology. Chapter 3 demonstrates, that the functional genes categories in and around pig EBRs are found significantly enriched for the gene ontology (GO) category *sensory perception of taste* and mostly affect the periphery of metabolic networks pathway. In addition, these findings illustrate the adaptation throughout the course of pig genome evolution.

Chapter 4 demonstrates the importance of computational techniques, along with application of scripting languages, which are applied to develop a novel chromosomal breakpoint identification tool named "evolutionary breakpoint analyser" (EBA). To the best of our knowledge, this tool is the only existing tool that precisely determines the EBRs demarking rearrangements in chromosomes. EBRs are enriched for segmental duplications, TEs and are often associated with lineage-specific expansions of gene families. To investigate a potential adaptive role of EBRs in different animal lineages, a bioinformatics tool is required that would identify EBRs reliably and assign them to the correct phylogenetic nodes. This task becomes more complicated when the genomes are not assembled to complete chromosomes and are represented by relatively short DNA scaffolds. To allow the detection of EBRs from a large number of sequenced genomes available through high throughput genome projects; an algorithm was developed to perform an automated identification and classification of EBRs from a large number of animal genomes, taking into account their phylogenetic relationships. In short, EBA was shown to detect and consistently classify lineage- and group- specific EBRs efficiently.

In Chapter 5, an extensive comparative study has been carried out using the EBA algorithm on a large number of genomes (i.e., on 21 avian, and five non-avian species) to address fundamental questions of genome organisation and chromosome evolution.

As it was shown previously for mammalian genomes (W. J. Murphy et al. 2005, Larkin et al. 2009, Larkin 2012) EBRs are enriched for segmental duplications, TEs and genes related to lineage-specific phenotypes. In order to investigate if similar patterns hold in avian genomes, the avian and non-avian genomes were exploited. The EBA algorithm and tool (see Chapter 4 section 4.2.2) was used to detect the EBRs and classify them using phylogenetic relationships of birds and other reptiles. The application of the EBA tool to avian genomes revealed many chromosomal rearrangements, which shed light on chromosome evolution in reptiles. In addition, these results provide novel evolutionary insights on the nature of karyotype stability in birds and the contribution of chromosomal rearrangements to the maintenance of ancestral phenotypes and formation of novel phenotypes in birds and reptiles. The wealth of genomes merits additional investigations of these data, which will hopefully provide more insights on the role and importance of chromosomal evolution.

6.3 CHROMOSOMAL REARRANGEMENTS AND THEIR IMPACT ON EVOLUTION

Systematic genome analysis has been used over the last decade to identify various features associated with EBRs. The analysis of the genomic landscape in and around EBRs has yielded important insights into the possible mechanism of breakpoint use, reuse and genome evolution. The results in this thesis in mammalian (Chapter 3) and bird (Chapter 5) genomes suggest that a key role is played by chromosomal rearrangements in adaptation to the environment.

The pig-based analysis results (Chapter 3) corroborate the previous observation (W. J. Murphy et al. 2005, Larkin 2011) that chromosomal rearrangements play an important role in genome evolution and adaptation. In amniote genomes, as expected, the largest fraction of EBRs was found to be lineage-specific. Moreover, avian evolution shows an alternation of faster and slower rates of chromosomal rearrangements, as reported in Neognathae clades where EBR frequency is however lower than in the mammalian orders. In contrast, the reuse EBRs within birds were found to be more frequent than in mammals. In addition, the amniote multi-species HSBs (which represent the regions of chromosomes where synteny and order of genes that have been maintained for over million of years) shows enrichment for developmentally important genes.

Similarly, as reported in mammals the lineage- and order-specific EBRs are enriched for TEs and other duplicated sequences that were active at the time of lineage and order formation. I found that almost all studied avian lineage-specific EBRs were either significantly enriched for or had elevated fraction of at least one of the highly abundant families of TEs. This implies that in avian lineages TEs contributed to the genome rearrangements, as has been reported in mammals (Larkin 2012). The analysis of gene ontologies for lineage-specific EBRs indicated that the genomic regions and gene networks are related to species-specific characteristics. The GO category enrichment analysis in EBRs identified five GO terms that were directly linked to adaptive changes in bird lineages. These results suggest that at least some evolutionary chromosome rearrangements may have adaptive value by creating novel configurations of structural and regulatory loci involved in responses to environmental challenges.

6.4 RECOMMENDATION

6.4.1 Limitation

The molecular biology research has reached the genomics era, where genome sequences are commonly used for comparative genome analysis. Even if there are many biological software and tools for storing, comparing and visualising the wealth of genomic data, these resources suffer several major computational as well as biological limitations. One major flaw could be the sensitivity of the genome alignment algorithms and procedures. Apart from that, another difficulty is that the findings from a given reference genome cannot be directly used within another target genome context, but that rather painstaking genomic and computational validation is needed. Comparative genomics has a huge potential in evolutionary genomics research, but there are a number of limitations as well:

1. Most of analysis requires a large number of high-quality DNA sequences, which can be difficult to handle in small computational labs.
2. Various molecular and computational protocols are needed, which vary depending on the nature of research and priorities of the experiment.
3. Whole genome comparative genomics are computationally very expensive in terms of memory and time needed.

4. The computational analysis of the results needs a high computational and technical knowledge.
5. The availability of reliable and correct assembly of full-length chromosomes using NGS data(Alkan *et al.* 2010).

The availability of low cost genome sequencing in a reasonable timeframe makes comparative genomics a main focus for research over the next decade. Moreover, this advancement has taken into consideration evolutionary studies amongst phylogenetically related species and has inferred evolutionary mechanisms. Therefore, a great deal of effort is needed to develop computational algorithms that are able to cope with multispecies WGS. The computational alignment and analysis of assemblies to scaffolds, genomes, intra- and inter- chromosomal rearrangements and the identification of functional elements are some research areas that need extensive computational and algorithmic support to allow analysis by comparative genomics approaches. In addition to that, the visual interpretation of such biological information requires an improved interface to elucidate patterns.

6.4.2 Future work

This thesis explored the comparative genomics approaches with WGSs to shed light on genome evolution and adaptive biological processes. Phylogenetic information guided the inspection of the ways in which chromosomes change over evolutionary time. In each case, the computational method depends on well-posed questions based on current established biological knowledge.

One fruitful extension of this comparative genome work would be the examination of high resolution 3D genome architecture maps for syntenic and non-syntenic blocks within the genomes of other closely- and distantly- related amniotes. This work on chromosomal rearrangements could also be extended to the comprehensive study of mammals and more distantly related birds. These studies may provide more information and evidence that may accumulate for the specific mechanisms, which have caused evolutionary rearrangements and shaped amniote genomes. Furthermore, as more sequence-level studies in eukaryotes accumulate, it will help to assess whether there is any correspondence of rearrangement breakpoints across the genomes of multiple organisms.

There is a wealth of information encoded in eukaryotic and prokaryotic genomes. The basis for intelligence, immunity and development is all encoded within genome sequences. In forthcoming years, it will be interesting to articulate new hypotheses from 3D genomic data. New biological models will be needed to discover novel aspects of epigenetic regulation, and their very discovery will result from genome-wide studies. Development of new algorithms, statistical and computational methods and tool will be needed for exploration of biological data. The following are interesting projects that still lay ahead.

1. The demonstration and analysis of genomic interactions using HiC is possible even within single cells. This approach is likely to initiate the generation of a whole new wave of analytic tools. This will enable genome organisation and regulation to be investigated in much more depth than is currently possible. Accordingly, the EBA tool will be improved to use HiC data and detect EBRs.
2. The ENCODE project has revolutionised the biological understanding of non-coding DNA. This project changed the perception of "junk" DNA by demonstrating that non-coding DNA not only works as a genome operating system but also contains lots of genetic regulatory switches. In the future, it will be illuminating to look at these regions with respect to EBRs and HSBs. In addition, non-coding DNA, which is the regulatory fragment of biological function, needs to examined for its impact on HSB and EBRs. The understanding of genome evolution is will be possible following exploration and analysis of all various kinds of noisy and neutral biological processes.
3. The 3D organisation of amniote genomes and the functional relationship of gene expression during evolution remain largely unexplored. Studying these topics will help to determine whether the chromosome threads found on the surface of the nucleus are only affected by evolutionary forces or not, and if so, then which forces are responsible. In addition, I can also study some genes which have been activated and deactivated over periods of evolutionary time (Chapter 3). Therefore, it is important to develop a tool to trace genes of interests over evolutionary time and determine their evolutionary impacts.
4. While genetics research scientists are actively involved in discovering chromosomal rearrangements and synteny involved in complex biological mechanism, EBA detection and classification methods will need to be

continually reassessed and possibly redesigned for optimal prediction of complex evolutionary breakpoints. Although our EBA tool currently accounts for a large proportion of the chromosomal breakpoints and contributes to the understanding of chromosome evolution, new bioinformatics tools and methods that evaluate chromosome breakpoints for regulatory, functional enrichments and splicing will broaden our understanding of evolutionary mechanisms. Moreover, it is largely unclear where it is possible to detect EBRs without a reference genome or not. In future I will dedicate my time to resolve them and develop an algorithm to make that determination.

In the near future, there will be more insight into the effects and nature of chromosome rearrangements using 3D models of the genome. In addition, their vital role in various evolutionary mechanisms, and the regulation of gene expression, both local and genome-wide, will be better explained. With this thesis, it has been possible to increase the understanding of chromosomal rearrangements and adaptation throughout amniote evolution. There remain several challenging goals that need to be accomplished (examples are mentioned above), and further efforts are needed to understand these complex natural phenomena. The exploration and understanding of the position of EBRs within 3D chromosome models is certainly one topic for future research. A better understanding of chromosomal and evolutionary dynamics of closely- and distantly-related species is yet another goal. Finally, it can be suggested that a further focus on 3D amniote genome evolution is necessary to understand specific differences between HSBs and EBRs.

6.5 CONCLUSION

To the best of knowledge, this study is the first large-scale genome analysis to investigate the role of chromosomal rearrangements and their impact on amniote genome evolution. The previous comparative evolutionary studies on several species were either applied to very small genomes, or limited to a certain group of species. This thesis exploited a wide range of biological information from the sequenced amniote genomes (see Chapter 3 and 5). The computational analyses presented in this thesis have discovered unique biological findings that are non-discoverable by traditional molecular techniques, regardless of the time or effort spent. The computational approach

presented is general and has the competitive advantage that one can increase its power by increasing the number of species studied; as sequencing costs decrease and sequencing capacity increases, obtaining additional genomes becomes only a question of time. The computational comparison of multiple distantly- or closely-related species might present a new paradigm for understanding genome evolution. In particular, our multi-species comparative genome analysis methods are currently being applied to amniote genomes. This study reveals the power of comparative genome analysis, which can be applied to closely- or distantly-related species in order to infer a wide range of evolutionary mechanisms occur over the course of evolution.

The GO analysis using the MetaCore database shows that porcine EBRs and adjacent intervals are enriched for the genes involved in *sensory perception of taste* suggesting that taste phenotypes may be affected by the events associated with genomic rearrangements in pigs. On the other hand, there were 13 GO categories found to be significantly enriched in the lineage-specific EBRs of seven bird species. The 5 out of 13 GO terms may be associated with adaptive changes in the bird lineages. Adelie penguin's EBRs were enriched for genes connected with *pattern specification* and *regionalization* including the *NR2F2* and *KIF-3* genes. In addition, the GO terms overrepresented in the little egret-specific EBRs, cause the re-shuffling of *feeding behaviour* related genes, including the spatial reorganization of a genomic region having the gene *NPY1R*. The *NPY1R* reorganization is believed to be associated with the specific diet of egrets. The budgerigar-specific EBRs are enriched with the genes whose functions are connected with *forebrain development*. Three genes *NUMB*, *NOTCH1* and *DRAXIN,* out of those related to the forebrain development in budgerigar EBRs were found to be responsible for neuron differentiation. Peregrine falcon EBRs were found to be enriched with genes responsible for *cation channel activity* and *synapse,* in similar line, with the nervous system- and sodium ion transport-related genes evolving rapidly in two falcon species. All these EBRs enrichment study in avian genomes were done by Dr. Marta Ferre Belmonte. The distribution of TEs and other repetitive sequence families in and around pig-specific EBRs were enriched for LTR-ERV1 transposons and satellite repeats suggesting that these two families of repetitive sequences have contributed to the chromosomal evolution in the pig lineage. In contrast, due to a comparatively small fraction of TEs in bird genomes only four families of TEs: LINE-CR1, LTR-ERVL, LTR-ERVK and LTR-ERV1 passed this threshold in at least one of the bird genomes. Comparatively

small fraction of TEs in the bird genomes compared to mammals might explain the stable karyotype in birds compared to other lineages. My approach, henceforth, demonstrates how chromosomal rearrangements produce variations in the gene networks used by the natural selection for adaptation.

In other words, these results show that comparative analysis with closely related species can be invaluable in understanding the adaptive mechanism at a genomic level. It also reveals the way different EBR regions affect chromosomes during evolution and provides clues as to their evolutionary importance. These comparative genome studies show consistency with previous studies in mammals that suggest that chromosomal breakage in amniotes is not random but is connected to multiple genome features. Moreover, the enrichments study that used assembled and fragmented genomes, found functional categories of genes that are enriched in lineage- or order-specific breakpoint intervals. In many cases, these genes were directly related to ancestral- or lineage-specific adaptive biology. In birds, the rates of genome rearrangements are found to be lower than in mammals. A lower density of TEs in birds and the formation of micro-chromosomes are a likely factor responsible for the evolutionary stability of the avian karyotype. Continued advances and the availability of more genomes provide a basis for the identification of major chromosome changes that contributed to the formation of existing species or clades. This progress also contributes greatly toward an improved understanding of the role of chromosome rearrangements in adaptation and speciation.

APPENDIX A

List of published full length papers

Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491, 393-398.

Comparative genomics reveals insights into avian genome evolution and adaptation. Science, 346(6215), 1311-1320.

REFERENCES

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. and McVean, G. A. (2012) 'An integrated map of genetic variation from 1,092 human genomes', *Nature,* 491(7422), 56-65.

Abrusán, G., Krambeck, H.-J., Junier, T., Giordano, J. and Warburton, P. E. (2008) 'Biased distributions and decay of long interspersed nuclear elements in the chicken genome', *Genetics,* 178(1), 573-581.

Ackers, G. K. and Smith, F. R. (1985) 'Effects of site-specific amino acid modification on protein interactions and biological function', *Annual review of biochemistry,* 54(1), 597-629.

Al-Shahrour, F., Díaz-Uriarte, R. and Dopazo, J. (2004) 'FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes', *Bioinformatics,* 20(4), 578-580.

Alekseyev, M. and Pevzner, P. (2011) 'Limited lifespan of fragile regions in mammalian evolution', *Comparative Genomics*, 198-215.

Alekseyev, M. A. and Pevzner, P. A. (2009) 'Breakpoint graphs and ancestral genome reconstructions', *Genome Research,* 19(5), 943-57.

Alekseyev, M. A. and Pevzner, P. A. (2010) 'Comparative genomics reveals birth and death of fragile regions in mammalian evolution', *Genome Biology,* 11(11), R117.

Aleyasin, A. and Barendse, W. (1999) 'Comparative mapping of genes from human chromosome 12 by genetic linkage mapping in cattle', *Journal of Heredity,* 90(5), 537-542.

Alkan, C., Sajjadian, S. and Eichler, E. E. (2010) 'Limitations of next-generation genome sequence assembly', *Nature methods,* 8(1), 61-65.

Anderson, S. I., Lopez-Corrales, N. L., Gorick, B. and Archibald, A. L. (2000) 'A large-fragment porcine genomic library resource in a BAC vector', *Mammalian Genome,* 11(9), 811-814.

Archibald, A. L., Bolund, L., Churcher, C., Fredholm, M., Groenen, M. A. M., Harlizius, B., Lee, K. T., Milan, D., Rogers, J. and Rothschild, M. F. (2010) 'Pig genome sequence-analysis and publication strategy', *BMC Genomics,* 11(1), 438.

Armant, M. A. and Fenton, M. J. (2002) 'Toll-like receptors: a family of pattern-recognition receptors in mammals', *Genome biology,* 3(8), reviews3011.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S. and Eppig, J. T. (2000) 'Gene Ontology: tool for the unification of biology', *Nature genetics,* 25(1), 25-29.

Assis, R., Kondrashov, A. S., Koonin, E. V. and Kondrashov, F. A. (2008) 'Nested genes and increasing organizational complexity of metazoan genomes', *Trends in Genetics,* 24(10), 475-478.

Ayala, D., Ullastres, A. and González, J. (2014) 'Adaptation through chromosomal inversions in Anopheles', *Frontiers in Genetics,* 5.

Ayala, F. J. and Coluzzi, M. (2005) 'Chromosome speciation: humans, Drosophila, and mosquitoes', *Proc Natl Acad Sci U S A,* 102 Suppl 1, 6535-42.

Ayala, F. J. and Coluzzi, M. (2005) 'Chromosome speciation: humans, Drosophila, and mosquitoes', *Proceedings of the National Academy of Sciences of the United States of America,* 102(Suppl 1), 6535-6542.

Baertsch, R., Diekhans, M., Kent, W. J., Haussler, D. and Brosius, J. (2008) 'Retrocopy contributions to the evolution of the human genome', *BMC Genomics,* 9(1), 466.

Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. and Eichler, E. E. (2004) 'Hotspots of mammalian chromosomal evolution', *Genome Biology,* 5(4), R23.

Bantignies, F. and Cavalli, G. (2011) 'Polycomb group proteins: repression in 3D', *Trends in Genetics,* 27(11), 454-464.

Barlow, D. P., Stöger, R., Herrmann, B., Saito, K. and Schweifer, N. (1991) 'The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus', *Nature,* 349(6304), 84-87.

Barreiro, L. B. and Quintana-Murci, L. (2009) 'From evolutionary genetics to human immunology: how selection shapes host defence genes', *Nature Reviews Genetics,* 11(1), 17-30.

Bartolomé, C. and Charlesworth, B. (2006) 'Rates and patterns of chromosomal evolution in Drosophila pseudoobscura and D. miranda', *Genetics,* 173(2), 779-791.

Barton, R. C. and Scherer, S. (1994) 'Induced chromosome rearrangements and morphologic variation in Candida albicans', *Journal of bacteriology,* 176(3), 756-763.

Bártová, E. and Kozubek, S. (2006) 'Nuclear architecture in the light of gene expression and cell differentiation studies', *Biology of the Cell,* 98(6), 323-336.

Bastide, M. and McCombie, W. R. (2007) 'Assembling genomic DNA sequences with PHRAP', *Current Protocols in Bioinformatics.*

Bateson, W. and Punnett, R. C. (1911) 'On the inter-relations of genetic factors', *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character,* 84(568), 3-8.

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P. and Lander, E. S. (2002) 'ARACHNE: a whole-genome shotgun assembler', *Genome Research,* 12(1), 177-189.

Becker, S. E., Thomas, R., Trifonov, V. A., Wayne, R. K., Graphodatsky, A. S. and Breen, M. (2011) 'Anchoring the dog to its relatives reveals new evolutionary breakpoints across 11 species of the Canidae and provides new clues for the role of B chromosomes', *Chromosome Research,* 19(6), 685-708.

Becker, T. S. and Lenhard, B. (2007) 'The random versus fragile breakage models of chromosome evolution: a matter of resolution', *Molecular Genetics and Genomics,* 278(5), 487-491.

Beißbarth, T. and Speed, T. P. (2004) 'GOstat: find statistically overrepresented Gene Ontologies within a group of genes', *Bioinformatics,* 20(9), 1464-1465.

Bell, M. A., Aguirre, W. E. and Buck, N. J. (2004) 'Twelve years of contemporary armor evolution in a threespine stickleback population', *Evolution,* 58(4), 814-824.

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y. and Dekker, J. (2012) 'Hi–C: A comprehensive technique to capture the conformation of genomes', *Methods,* 58(3), 268-276.

Bennetzen, J. L., Ma, J. and Devos, K. M. (2005) 'Mechanisms of recent genome size variation in flowering plants', *Annals of botany,* 95(1), 127-132.

Bentley, D. R. (2006) 'Whole-genome re-sequencing', *Current opinion in genetics & development,* 16(6), 545-552.

Betrán, E. and Long, M. (2002) 'Expansion of genome coding regions by acquisition of new genes', *Genetica,* 115(1), 65-80.

Betrán, E., Thornton, K. and Long, M. (2002) 'Retroposed new genes out of the X in Drosophila', *Genome Research,* 12(12), 1854-1859.

Betts, M. J. and Russell, R. B. (2003) 'Amino acid properties and consequences of substitutions', *Bioinformatics for geneticists,* 317, 289.

Beutler, B. and Rehli, M. (2002) 'Evolution of the TIR, tolls and TLRs: functional inferences from computational biology' in *Toll-Like Receptor Family Members and Their Ligands*, Springer, 1-21.

Beyer, A., Bandyopadhyay, S. and Ideker, T. (2007) 'Integrating physical and genetic maps: from genomes to interaction networks', *Nature Reviews Genetics,* 8(9), 699-710.

Biémont, C. and Vieira, C. (2006) 'Genetics: junk DNA as an evolutionary force', *Nature,* 443(7111), 521-524.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z. and Galon, J. (2009) 'ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks', *Bioinformatics,* 25(8), 1091-1093.

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C. and Apweiler, R. (2009) 'QuickGO: a web-based tool for Gene Ontology searching', *Bioinformatics,* 25(22), 3045-3046.

Birney, E. (2012) 'The making of ENCODE: lessons for big-data projects', *Nature,* 489(7414), 49-51.

Boeke, J. D., Garfinkel, D. J., Styles, C. A. and Fink, G. R. (1985) 'Ty elements transpose through an RNA intermediate', *Cell,* 40(3), 491-500.

Boisvert, S., Laviolette, F. and Corbeil, J. (2010) 'Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies', *Journal of Computational Biology,* 17(11), 1519-1533.

Bonhommeau, S., Dubroca, L., Le Pape, O., Barde, J., Kaplan, D. M., Chassot, E. and Nieblas, A.-E. (2013) 'Eating up the world's food web and the human trophic level', *Proceedings of the National Academy of Sciences,* 110(51), 20617-20620.

Bose, P., Hermetz, K. E., Conneely, K. N. and Rudd, M. K. (2014) 'Tandem repeats and G-rich sequences are enriched at human CNV breakpoints', *PLoS ONE,* 9(7), e101607.

Botstein, D., White, R. L., Skolnick, M. and Davis, R. W. (1980) 'Construction of a genetic linkage map in man using restriction fragment length polymorphisms', *American journal of human genetics,* 32(3), 314.

Bourque, G., Pevzner, P. A. and Tesler, G. (2004) 'Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes', *Genome Research,* 14(4), 507-16.

Bovine Genome, S., Analysis, C., Elsik, C. G., Tellam, R. L., Worley, K. C., Gibbs, R. A., Muzny, D. M., Weinstock, G. M., Adelson, D. L., Eichler, E. E., Elnitski, L., Guigo, R., Hamernik, D. L., Kappes, S. M., Lewin, H. A., Lynn, D. J., Nicholas, F. W., Reymond, A., Rijnkels, M., Skow, L. C., Zdobnov, E. M., Schook, L., Womack, J., Alioto, T., Antonarakis, S. E., Astashyn, A., Chapple, C. E., Chen, H. C., Chrast, J., Camara, F., Ermolaeva, O., Henrichsen, C. N., Hlavina, W., Kapustin, Y., Kiryutin, B., Kitts, P., Kokocinski, F., Landrum, M., Maglott, D., Pruitt, K., Sapojnikov, V., Searle, S. M., Solovyev, V., Souvorov, A., Ucla, C., Wyss, C., Anzola, J. M., Gerlach, D., Elhaik, E., Graur, D., Reese, J. T., Edgar, R. C., McEwan, J. C., Payne, G. M., Raison, J. M., Junier, T., Kriventseva, E. V., Eyras, E., Plass, M., Donthu, R., Larkin, D. M., Reecy, J., Yang, M. Q., Chen, L., Cheng, Z., Chitko-McKown, C. G., Liu, G. E., Matukumalli, L. K., Song, J., Zhu, B., Bradley, D. G., Brinkman, F. S., Lau, L. P., Whiteside, M. D., Walker, A., Wheeler, T. T., Casey, T., German, J. B., Lemay, D. G., Maqbool, N. J., Molenaar, A. J., Seo, S., Stothard, P., Baldwin, C. L., Baxter, R., Brinkmeyer-Langford, C. L., Brown, W. C., Childers, C. P., Connelley, T., Ellis, S. A., Fritz, K., Glass, E. J., Herzig, C. T., Iivanainen, A., Lahmers, K. K., Bennett, A. K., Dickens, C. M., Gilbert, J. G., Hagen, D. E., Salih, H., et al. (2009) 'The genome sequence of taurine cattle: a window to ruminant biology and evolution', *Science,* 324(5926), 522-8.

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M. and Sherlock, G. (2004) 'GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes', *Bioinformatics,* 20(18), 3710-3715.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G. and Chikhi, R. (2013) 'Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species', *Gigascience,* 2(1), 1-31.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N. A., Ganapathy, G., Gibbs, R. A., Gnerre, S., Godzaridis, E., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J. B., Ho, I. Y., Howard, J., Hunt, M., Jackman, S. D., Jaffe, D. B., Jarvis, E. D., Jiang, H., Kazakov, S., Kersey, P. J., Kitzman, J. O., Knight, J. R., Koren, S., Lam, T.-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., Maccallum, I., Macmanes, M. D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T. D., Paten, B., Paulo, O. S., Phillippy, A. M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F. J., Richards, S., Rokhsar, D. S., Ruby, J. G., Scalabrin, S., Schatz, M. C., Schwartz, D. C., Sergushichev, A., Sharpe, T., Shaw, T. I., Shendure, J., Shi, Y., Simpson, J. T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B. M., Wang, J., Worley, K. C., Yin, S., Yiu, S.-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S. and Korf, I. F. (2013) 'Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species', *Gigascience,* 2(1), 10.

Britten, R. J. and Davidson, E. H. (1969) 'Gene regulation for higher cells: a theory', *Science,* 165(891), 349-357.

Brock, D. C. and Moore, G. E. (2006) *Understanding Moore's law: four decades of innovation,* Chemical Heritage Foundation.

Brosius, J. (1999) 'RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements', *Gene,* 238(1), 115-134.

Brown, J. D. and O'Neill, R. J. (2010) 'Chromosomes, conflict, and epigenetics: chromosomal speciation revisited', *Annual review of genomics and human genetics,* 11, 291-316.

Brown, T. A. (2002) *Genomes,* Wiley-Liss, Oxford.

Brownlie, R. and Allan, B. (2011) 'Avian toll-like receptors', *Cell and tissue research,* 343(1), 121-130.

Bryant, D., Wong, W. K. and Mockler, T. (2009) 'QSRA–a quality-value guided de novo short read assembler', *BMC bioinformatics,* 10(1), 69.

Bulazel, K. V., Ferreri, G. C., Eldridge, M. and O'Neill, R. J. (2007) 'Species-specific shifts in centromere sequence composition are coincident with breakpoint reuse in karyotypically divergent lineages', *Genome Biol,* 8(8), R170.

Burt, D. W. (2002) 'Origin and evolution of avian microchromosomes', *Cytogenet Genome Res,* 96(1-4), 97-112.

Burt, D. W., Bruley, C., Dunn, I. C., Jones, C. T., Ramage, A., Law, A. S., Morrice, D. R., Paton, I. R., Smith, J. and Windsor, D. (1999) 'The dynamics of chromosome evolution in birds and mammals', *Nature,* 402(6760), 411-413.

Burt, D. W., Bruley, C., Dunn, I. C., Jones, C. T., Ramage, A., Law, A. S., Morrice, D. R., Paton, I. R., Smith, J., Windsor, D., Sazanov, A., Fries, R. and Waddington, D. (1999) 'The dynamics of chromosome evolution in birds and mammals', *Nature,* 402(6760), 411-3.

Bush, G. L., Case, S., Wilson, A. and Patton, J. (1977) 'Rapid speciation and chromosomal evolution in mammals', *Proceedings of the National Academy of Sciences,* 74(9), 3942-3946.

Butlin, R. K. (2005) 'Recombination and speciation', *Molecular Ecology,* 14(9), 2621-2635.

Cahill, M. J., Köser, C. U., Ross, N. E. and Archer, J. A. C. (2010) 'Read length and repeat resolution: exploring prokaryote genomes using next-generation sequencing technologies', *PLoS ONE,* 5(7), e11518.

Callinan, P. and Batzer, M. (2006) 'Retrotransposable elements and human disease'.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B. and Lewis, S. (2009) 'AmiGO: online access to ontology and annotation data', *Bioinformatics,* 25(2), 288-289.

Caspi, A. and Pachter, L. (2006) 'Identification of transposable elements using multiple alignments of related genomes', *Genome research,* 16(2), 260-270.

Čepica, S., Masopust, M., Knoll, A., Bartenschlager, H., Yerle, M., Rohrer, G. and Geldermann, H. (2006) 'Linkage and RH mapping of 10 genes to a QTL region for fatness and muscling traits on pig chromosome X', *Animal genetics,* 37(6), 603-604.

Chandley, A. (1989) 'Asymmetry in chromosome pairing: a major factor in de novo mutation and the production of genetic disease in man', *Journal of medical genetics,* 26(9), 546-552.

Chang, S.-L., Lai, H.-Y., Tung, S.-Y. and Leu, J.-Y. (2013) 'Dynamic large-scale chromosomal rearrangements fuel rapid adaptation in yeast populations', *PLoS genetics,* 9(1), e1003232.

Chen, K., Baxter, T., Muir, W. M., Groenen, M. A. and Schook, L. B. (2007) 'Genetic resources, genome mapping and evolutionary genomics of the pig (Sus scrofa)', *International journal of biological sciences,* 3(3), 153.

Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., Fulton, R. S., Graves, T. A., Hillier, L. W., Mardis, E. R. and McPherson, J. D. (2002) 'Initial sequencing and comparative analysis of the mouse genome', *Nature,* 420(6915), 520-562.

Chiu, J. C., Lee, E. K., Egan, M. G., Sarkar, I. N., Coruzzi, G. M. and DeSalle, R. (2006) 'OrthologID: automation of genome-scale ortholog identification within a parsimony framework', *Bioinformatics,* 22(6), 699-707.

Cho, Y. S., Hu, L., Hou, H., Lee, H., Xu, J., Kwon, S., Oh, S., Kim, H. M., Jho, S., Kim, S., Shin, Y. A., Kim, B. C., Kim, H., Kim, C. U., Luo, S. J., Johnson, W. E., Koepfli, K. P., Schmidt-Kuntzel, A., Turner, J. A., Marker, L., Harper, C., Miller, S. M., Jacobs, W., Bertola, L. D., Kim, T. H., Lee, S., Zhou, Q., Jung, H. J., Xu, X., Gadhvi, P., Xu, P., Xiong, Y., Luo, Y., Pan, S., Gou, C., Chu, X., Zhang, J., Liu, S., He, J., Chen, Y., Yang, L., Yang, Y., Wang, J., Kim, C. H., Kwak, H., Kim, J. S., Hwang, S., Ko, J., Kim, C. B., Bayarlkhagva, D., Paek, W. K., Kim, S. J., O'Brien, S. J. and Bhak, J. (2013) 'The tiger genome and comparative analysis with lion and snow leopard genomes', *Nat Commun,* 4, 2433.

Chou, H.-H., Takematsu, H., Diaz, S., Iber, J., Nickerson, E., Wright, K. L., Muchmore, E. A., Nelson, D. L., Warren, S. T. and Varki, A. (1998) 'A mutation in human CMP-sialic acid hydroxylase occurred after the Homo-Pan divergence', *Proceedings of the National Academy of Sciences,* 95(20), 11751-11756.

Chowdhary, B., Frönicke, L., Gustavsson, I. and Scherthan, H. (1996) 'Comparative analysis of the cattle and human genomes: detection of ZOO-FISH and gene mapping-based chromosomal homologies', *Mammalian Genome,* 7(4), 297-302.

Chowdhary, B. P., Raudsepp, T., Kata, S. R., Goh, G., Millon, L. V., Allan, V., Piumi, F., Guérin, G., Swinburne, J. and Binns, M. (2003) 'The first-generation whole-genome radiation hybrid map in the horse identifies conserved segments in human and mouse genomes', *Genome Research,* 13(4), 742-751.

Church, G. M. and Gilbert, W. (1984) 'Genomic sequencing', *Proceedings of the National Academy of Sciences,* 81(7), 1991-1995.

Churchill, G. A., Daniels, D. L. and Waterman, M. S. (1990) 'The distribution of restriction enzyme sites in Escherichia coli', *Nucleic Acids Research,* 18(3), 589-97.

Clark, R. and Felsenfeld, G. (1971) 'Structure of chromatin', *Nature,* 229(4), 101-106.

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. (2009) 'Continuous base identification for single-molecule nanopore DNA sequencing', *Nature nanotechnology,* 4(4), 265-270.

Coghlan, A. and Wolfe, K. H. (2002) 'Fourfold faster rate of genome rearrangement in nematodes than in Drosophila', *Genome Research,* 12(6), 857-867.

Collins, M. and Rubin, G. M. (1983) 'Structure of chromosomal rearrangements induced by the FB transposable element in Drosophila', *Nature,* 308(5957), 323-327.

Consortium, E. P. (2004) 'The ENCODE (ENCyclopedia of DNA elements) project', *Science,* 306(5696), 636-640.

Consortium, G. O. (2008) 'The gene ontology project in 2008', *Nucleic Acids Research,* 36(suppl 1), D440-D444.

Consortium, G. O. (2010) 'The Gene Ontology in 2010: extensions and refinements', *Nucleic Acids Research,* 38(suppl 1), D331-D335.

Consortium, G. O. (2013) 'Gene Ontology annotations and resources', *Nucleic Acids Research,* 41(D1), D530-D535.

Consortium, S. (1998) '{Genome sequence of the nematode $ C. elegans $: A platform for investigating biology}', *Science,* 282, 2012-2018.

Cooper, G. M. and Hausman, R. E. (2000) *The cell,* ASM press Washington.

Cotton, J. A. and Page, R. D. (2005) 'Rates and patterns of gene duplication and loss in the human genome', *Proceedings of the Royal Society B: Biological Sciences,* 272(1560), 277-283.

Craig, N. L. (1996) 'Transposon Tn7' in *Transposable Elements*, Springer, 27-48.

Cremer, T. and Cremer, M. (2010) 'Chromosome territories', *Cold Spring Harbor perspectives in biology,* 2(3), a003889.

Critchlow, S. E. and Jackson, S. P. (1998) 'DNA end-joining: from yeast to man', *Trends in biochemical sciences,* 23(10), 394-398.

Crombach, A. and Hogeweg, P. (2007) 'Chromosome rearrangements and the evolution of genome structuring and adaptability', *Molecular biology and evolution,* 24(5), 1130-1139.

Crosland, M. W. J. and Crozier, R. H. (1986) 'Myrmecia pilosula, an ant with only one pair of chromosomes', *Science,* 231(4743), 1278-1278.

Culik, B. M., Wilson, R. P. and Bannasch, R. (1994) 'Underwater Swimming at Low Energetic Cost by Pygoscelid Penguins', *Journal of Experimental Biology,* 197(1), 65-78.

Dalloul, R. A., Long, J. A., Zimin, A. V., Aslam, L., Beal, K., Blomberg Le, A., Bouffard, P., Burt, D. W., Crasta, O., Crooijmans, R. P., Cooper, K., Coulombe, R. A., De, S., Delany, M. E., Dodgson, J. B., Dong, J. J., Evans, C., Frederickson, K. M., Flicek, P., Florea, L., Folkerts, O., Groenen, M. A., Harkins, T. T., Herrero, J., Hoffmann, S., Megens, H. J., Jiang, A., de Jong, P., Kaiser, P., Kim, H., Kim, K. W., Kim, S., Langenberger, D., Lee, M. K., Lee, T., Mane, S., Marcais, G., Marz, M., McElroy, A. P., Modise, T., Nefedov, M., Notredame, C., Paton, I. R., Payne, W. S., Pertea, G., Prickett, D., Puiu, D., Qioa, D., Raineri, E., Ruffier, M., Salzberg, S. L., Schatz, M. C., Scheuring, C., Schmidt, C. J., Schroeder, S., Searle, S. M., Smith, E. J., Smith, J., Sonstegard, T. S., Stadler, P. F., Tafer, H., Tu, Z. J., Van Tassell, C. P., Vilella, A. J., Williams, K. P., Yorke, J. A., Zhang, L., Zhang, H. B., Zhang, X., Zhang, Y. and Reed, K. M. (2010) 'Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis', *PLoS Biol,* 8(9).

Darwin, C. (1859) 'On the Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life: Murray',

Datta, R. S., Meacham, C., Samad, B., Neyer, C. and Sjölander, K. (2009) 'Berkeley PHOG: PhyloFacts orthology group prediction web server', *Nucleic Acids Research*, gkp373.

de Grouchy, J. (1972) *Proceedings of the Fourth International Congress of Human Genetics, Paris, 6-11 September 1971,* Excerpta Medica.

de Koning, D. J., Janss, L. L., Rattink, A. P., van Oers, P. A., de Vries, B. J., Groenen, M. A., van der Poel, J. J., de Groot, P. N. and van Arendonk, J. A. (1999) 'Detection of quantitative trait loci for backfat thickness and intramuscular fat content in pigs (Sus scrofa)', *Genetics,* 152(4), 1679-1690.

De Lorenzi, L., Molteni, L. and Parma, P. (2010) 'FISH mapping in cattle (Bos taurus L.) is not yet out of fashion', *Journal of applied genetics,* 51(4), 497-499.

Deakin, J. E. and Ezaz, T. (2014) 'Tracing the evolution of amniote chromosomes', *Chromosoma,* 123(3), 201-216.

del Hoyo, J. E., Andrew; Chrsitie, David A. (1992-2013) *Handbook of the birds of the world,* Spain: Lynx Edicions.

Delneri, D., Colson, I., Grammenoudi, S., Roberts, I. N., Louis, E. J. and Oliver, S. G. (2003) 'Engineering evolution to study speciation in yeasts', *Nature,* 422(6927), 68-72.

Delseny, M. (2004) 'Re-evaluating the relevance of ancestral shared synteny as a tool for crop improvement', *Current opinion in plant biology,* 7(2), 126-131.

Demuth, J. P., De Bie, T., Stajich, J. E., Cristianini, N. and Hahn, M. W. (2006) 'The evolution of mammalian gene families', *PLoS One,* 1(1), e85.

Dennis Jr, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. and Lempicki, R. A. (2003) 'DAVID: database for annotation, visualization, and integrated discovery', *Genome Biol,* 4(5), P3.

Derrien, T., André, C., Galibert, F. and Hitte, C. (2007) 'AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps', *Bioinformatics,* 23(4), 498-499.

Dewannieux, M., Esnault, C. and Heidmann, T. (2003) 'LINE-mediated retrotransposition of marked Alu sequences', *Nature genetics,* 35(1), 41-48.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012) 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature,* 485(7398), 376-380.

Dobzhansky, T. (1936) 'Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids', *Genetics,* 21(2), 113.

Dohm, J. C., Lottaz, C., Borodina, T. and Himmelbauer, H. (2007) 'SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing', *Genome Research,* 17(11), 1697-1706.

Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C. and Conklin, B. R. (2003) 'MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data', *Genome Biol,* 4(1), R7.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T. P., Bowden, D. W., Smith, D. R. and Lander, E. S. (1987) 'A genetic linkage map of the human genome', *Cell,* 51(2), 319-337.

Donthu, R., Lewin, H. A. and Larkin, D. M. (2009) 'SyntenyTracker: a tool for defining homologous synteny blocks using radiation hybrid maps and whole-genome sequence', *BMC Res Notes,* 2, 148.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J. and Nusbaum, C. (2006) 'Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements', *Genome Research,* 16(10), 1299-1309.

Drillon, G., Carbone, A. and Fischer, G. (2013) 'Combinatorics of chromosomal rearrangements based on synteny blocks and synteny packs', *Journal of Logic and Computation,* 23(4), 815-838.

Drillon, G., Carbone, A. and Fischer, G. (2014) 'SynChro: a fast and easy tool to reconstruct and visualize Synteny blocks along eukaryotic chromosomes', *PLoS ONE,* 9(3), e92621.

Dujon, B., Alexandraki, D., Andre, B., Ansorge, W., Baladron, V., Ballesta, J., Banrevi, A., Bolle, P., Bolotin-Fukuhara, M. and Bossier, P. (1994) 'Complete DNA sequence of yeast chromosome XI'.

Dunham, M. J., Badrane, H., Ferea, T., Adams, J., Brown, P. O., Rosenzweig, F. and Botstein, D. (2002) 'Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae', *Proceedings of the National Academy of Sciences,* 99(25), 16144.

Duret, L., Marais, G. and Biémont, C. (2000) 'Transposons but not retrotransposons are located preferentially in regions of high recombination rate in Caenorhabditis elegans', *Genetics,* 156(4), 1661-1669.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001) 'Empirical Bayes analysis of a microarray experiment', *Journal of the American Statistical Association,* 96(456), 1151-1160.

Eichler, E. E. and Sankoff, D. (2003) 'Structural dynamics of eukaryotic chromosome evolution', *Science's STKE,* 301(5634), 793.

Eisen, J. A., Heidelberg, J. F., White, O. and Salzberg, S. L. (2000) 'Evidence for symmetric chromosomal inversions around the replication origin in bacteria', *Genome Biol,* 1(6), 01-9.

Elbers, C. C., de Kovel, C. G., van der Schouw, Y. T., Meijboom, J. R., Bauer, F., Grobbee, D. E., Trynka, G., van Vliet-Ostaptchouk, J. V., Wijmenga, C. and Onland-Moret, N. C. (2009) 'Variants in neuropeptide Y receptor 1 and 5 are associated with nutrient-specific food intake and are under recent selection in Europeans', *PLoS ONE,* 4(9), e7070.

Elgar, G. and Vavouri, T. (2008) 'Tuning in to the signals: noncoding sequence conservation in vertebrate genomes', *Trends in Genetics,* 24(7), 344-352.

Ellegren, H. (2005) 'The avian genome uncovered', *Trends Ecol Evol,* 20(4), 180-186.

Ellegren, H. (2010) 'Evolutionary stasis: the stable chromosomes of birds', *Trends Ecol Evol,* 25(5), 283-91.

Elsik, C. G., Tellam, R. L. and Worley, K. C. (2009) 'The genome sequence of taurine cattle: a window to ruminant biology and evolution', *Science,* 324(5926), 522-528.

Emerson, J., Kaessmann, H., Betrán, E. and Long, M. (2004) 'Extensive gene traffic on the mammalian X chromosome', *Science,* 303(5657), 537-540.

Epstein, H. (1969) 'Domestic animals of China', *Boerma. Tech. Commun. Commonw. Bur. Anim. Breed. Genet.,* (18).

Esnault, C., Maestre, J. and Heidmann, T. (2000) 'Human LINE retrotransposons generate processed pseudogenes', *Nature genetics,* 24(4), 363-367.

Everts-van der Wind, A., Kata, S. R., Band, M. R., Rebeiz, M., Larkin, D. M., Everts, R. E., Green, C. A., Liu, L., Natarajan, S. and Goldammer, T. (2004) 'A 1463 gene cattle–human comparative map with anchor points defined by human genome sequence coordinates', *Genome Research,* 14(7), 1424-1437.

Faria, R. and Navarro, A. (2010) 'Chromosomal speciation revisited: rearranging theory with pieces of evidence', *Trends Ecol Evol,* 25(11), 660-669.

Farre, M., Bosch, M., Lopez-Giraldez, F., Ponsa, M. and Ruiz-Herrera, A. (2011) 'Assessing the role of tandem repeats in shaping the genomic architecture of great apes', *PLoS ONE,* 6(11), e27239.

Farré, M., Bosch, M., López-Giráldez, F., Ponsà, M. and Ruiz-Herrera, A. (2011) 'Assessing the role of tandem repeats in shaping the genomic architecture of great apes', *PLoS One,* 6(11), e27239.

Farré, M., Micheletti, D. and Ruiz-Herrera, A. (2013) 'Recombination rates and genomic shuffling in human and chimpanzee—a new twist in the chromosomal speciation theory', *Molecular biology and evolution,* 30(4), 853-864.

Federhen, S. (2012) 'The NCBI taxonomy database', *Nucleic Acids Research,* 40(D1), D136-D143.

Fedoroff, N. V. (2012) 'Transposable elements, epigenetics, and genome evolution', *Science,* 338(6108), 758-767.

Feng, Q., Moran, J. V., Kazazian Jr, H. H. and Boeke, J. D. (1996) 'Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition', *Cell,* 87(5), 905-916.

Ferguson-Smith, M. A. and Trifonov, V. (2007) 'Mammalian karyotype evolution', *Nature Reviews Genetics,* 8(12), 950-962.

Ferguson, J. L., Mulvanny, P. J. and Brauth, S. E. (1978) 'Distribution of neurons projecting to the retina of Caiman crocodilus', *Brain Behav Evol,* 15(4), 294-306.

Ferrari, D., Sumoy, L., Gannon, J., Sun, H., Brown, A. M., Upholt, W. B. and Kosher, R. A. (1995) 'The expression pattern of the Distal-less homeobox-containing gene Dlx-5 in the developing chick limb bud suggests its involvement in apical ectodermal ridge activity, pattern formation, and cartilage differentiation', *Mech Dev,* 52(2-3), 257-64.

Feuillet, C. and Keller, B. (2002) 'Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution', *Annals of botany,* 89(1), 3-10.

Finlan, L. E. and Bickmore, W. A. (2008) 'Porin new light onto chromatin and nuclear organization', *Genome Biology,* 9(5), 222.

Fischer, A., Gilad, Y., Man, O. and Pääbo, S. (2005) 'Evolution of bitter taste receptors in humans and apes', *Molecular biology and evolution,* 22(3), 432-436.

Fischer, G., James, S., Roberts, I., Oliver, S. and Louis, E. (2000) 'Chromosomal evolution in Saccharomyces', *Nature,* 405(6785), 451-454.

Fischer, G., Neuvéglise, C., Durrens, P., Gaillardin, C. and Dujon, B. (2001) 'Evolution of gene order in the genomes of two related yeast species', *Genome Research,* 11(12), 2009-2019.

Fischer, G., Rocha, E. P. C., Brunet, F., Vergassola, M. and Dujon, B. (2006) 'Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages', *PLoS genetics,* 2(3), e32.

Flisikowska, T., Kind, A. and Schnieke, A. (2013) 'The new pig on the block: modelling cancer in pigs', *Transgenic Research*, 1-8.

Fontanillas, P., Hartl, D. L. and Reuter, M. (2007) 'Genome organization and gene expression shape the transposable element distribution in the Drosophila melanogaster euchromatin', *PLoS genetics,* 3(11), e210.

Francis, N. J., Kingston, R. E. and Woodcock, C. L. (2004) 'Chromatin compaction by a polycomb group protein complex', *Science,* 306(5701), 1574-1577.

Gasser, S. M. (2002) 'Visualizing chromatin dynamics in interphase nuclei', *Science,* 296(5572), 1412-1416.

Ge, R. L., Cai, Q., Shen, Y. Y., San, A., Ma, L., Zhang, Y., Yi, X., Chen, Y., Yang, L., Huang, Y., He, R., Hui, Y., Hao, M., Li, Y., Wang, B., Ou, X., Xu, J., Wu, K., Geng, C., Zhou, W., Zhou, T., Irwin, D. M., Yang, Y., Ying, L., Bao, H., Kim, J., Larkin, D. M., Ma, J., Lewin, H. A., Xing, J., Platt, R. N., 2nd, Ray, D. A., Auvil, L., Capitanu, B., Zhang, X., Zhang, G., Murphy, R. W., Wang, J. and Zhang, Y. P. (2013) 'Draft genome sequence of the Tibetan antelope', *Nat Commun,* 4, 1858.

Geiser, D. M., Arnold, M. L. and Timberlake, W. E. (1996) 'Wild chromosomal variants inAspergillus nidulans', *Current genetics,* 29(3), 293-300.

Genome, K. (2009) 'Genome 10K: A proposal to obtain whole-genome sequences of 10,000 vertebrate species', *J Hered,* 100, 659-674.

Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E., Okwuonu, G., Hines, S., Lewis, L., DeRamo, C., Delgado, O., Dugan-Rocha, S., Miner, G., Morgan, M., Hawes, A., Gill, R., Celera, Holt, R. A., Adams, M. D., Amanatides, P. G., Baden-Tillson, H., Barnstead, M., Chin, S., Evans, C. A., Ferriera, S., Fosler, C., Glodek, A., Gu, Z., Jennings, D., Kraft, C. L., Nguyen, T., Pfannkoch, C. M., Sitter, C., Sutton, G. G., Venter, J. C., Woodage, T., Smith, D., Lee, H. M., Gustafson, E., Cahill, P., Kana, A., Doucette-Stamm, L., Weinstock, K., Fechtel, K., Weiss, R. B., Dunn, D. M., Green, E. D., Blakesley, R. W., Bouffard, G. G., De Jong, P. J., Osoegawa, K., Zhu, B., Marra, M., Schein, J., Bosdet, I., Fjell, C., Jones, S., Krzywinski, M., Mathewson, C., Siddiqui, A., Wye, N., McPherson, J., Zhao, S., Fraser, C. M., Shetty, J., Shatsman, S., Geer, K., Chen, Y., Abramzon, S., Nierman, W. C., Havlak, P. H., Chen, R., Durbin, K. J., Egan, A., Ren, Y., Song, X. Z., Li, B., Liu, Y., Qin, X., Cawley, S., Worley, K. C., Cooney, A. J., D'Souza, L. M., Martin, K., Wu, J. Q., Gonzalez-Garay, M. L., Jackson, A. R., Kalafus, K. J., McLeod, M. P., Milosavljevic, A., Virk, D., Volkov, A., Wheeler, D. A., Zhang, Z., Bailey, J. A., Eichler, E. E., et al. (2004) 'Genome sequence of the Brown Norway rat yields insights into mammalian evolution', *Nature,* 428(6982), 493-521.

Gibcus, J. H. and Dekker, J. (2013) 'The hierarchy of the 3D genome', *Molecular cell,* 49(5), 773-782.

Gilbert, W. and Maxam, A. (1973) 'The nucleotide sequence of the lac operator', *Proceedings of the National Academy of Sciences,* 70(12), 3581.

Giordano, J., Ge, Y., Gelfand, Y., Abrusán, G., Benson, G. and Warburton, P. E. (2007) 'Evolutionary history of mammalian transposons determined by genome-wide defragmentation', *PLoS computational biology,* 3(7), e137.

Girirajan, S., Chen, L., Graves, T., Marques-Bonet, T., Ventura, M., Fronick, C., Fulton, L., Rocchi, M., Fulton, R. S. and Wilson, R. K. (2009) 'Sequencing human–gibbon breakpoints of synteny reveals mosaic new insertions at rearrangement sites', *Genome research,* 19(2), 178-190.

Goddard, M., Foweraker, J. and Wallwork, J. (2000) 'Xenotransplantation—2000', *Journal of clinical pathology,* 53(1), 44-48.

Göndör, A. and Ohlsson, R. (2009) 'Chromosome crosstalk in three dimensions', *Nature,* 461(7261), 212-217.

Gordon, L., Yang, S., Tran-Gyamfi, M., Baggott, D., Christensen, M., Hamilton, A., Crooijmans, R., Groenen, M., Lucas, S. and Ovcharenko, I. (2007) 'Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions', *Genome Research,* 17(11), 1603-1613.

Gosden, C. (1995) 'Arboriculture and agriculture in coastal Papua New Guinea', *Antiquity,* 69(265), 807-817.

Grabherr, M. G., Russell, P., Meyer, M., Mauceli, E., Alfoldi, J., Di Palma, F. and Lindblad-Toh, K. (2010) 'Genome-wide synteny through highly sensitive sequence alignment: Satsuma', *Bioinformatics,* 26(9), 1145-51.

Grabherr, M. G., Russell, P., Meyer, M., Mauceli, E., Alföldi, J., Di Palma, F. and Lindblad-Toh, K. (2010) 'Genome-wide synteny through highly sensitive sequence alignment: Satsuma', *Bioinformatics,* 26(9), 1145-1151.

Gray, Y. H. (2000) 'It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements', *Trends in Genetics,* 16(10), 461-468.

Green, P. (2002) 'Whole-genome disassembly', *Proceedings of the National Academy of Sciences,* 99(7), 4143-4144.

Gregory, T. (2014) 'Animal Genome Size Database', [online], available: http://www.genomesize.com [accessed

Griffin, D., Robertson, L., Tempest, H. and Skinner, B. (2007) 'The evolution of the avian genome as revealed by comparative molecular cytogenetics', *Cytogenet Genome Res,* 117(1-4), 64-77.

Griffin, D. K., Robertson, L. B., Tempest, H. G. and Skinner, B. M. (2007) 'The evolution of the avian genome as revealed by comparative molecular cytogenetics', *Cytogenet Genome Res,* 117(1-4), 64-77.

Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C. and Gelbart, W. M. (1999) 'Modern genetic analysis'.

Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C. and Gelbart, W. M. (2000) 'Translocations'.

Groenen, M. A., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., Rogel-Gaillard, C., Park, C., Milan, D. and Megens, H.-J. (2012) 'Analyses of pig genomes provide insight into porcine demography and evolution', *Nature,* 491(7424), 393-398.

Groenen, M. A., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H. J., Li, S., Larkin, D.

M., Kim, H., Frantz, L. A., Caccamo, M., Ahn, H., Aken, B. L., Anselmo, A., Anthon, C., Auvil, L., Badaoui, B., Beattie, C. W., Bendixen, C., Berman, D., Blecha, F., Blomberg, J., Bolund, L., Bosse, M., Botti, S., Bujie, Z., Bystrom, M., Capitanu, B., Carvalho-Silva, D., Chardon, P., Chen, C., Cheng, R., Choi, S. H., Chow, W., Clark, R. C., Clee, C., Crooijmans, R. P., Dawson, H. D., Dehais, P., De Sapio, F., Dibbits, B., Drou, N., Du, Z. Q., Eversole, K., Fadista, J., Fairley, S., Faraut, T., Faulkner, G. J., Fowler, K. E., Fredholm, M., Fritz, E., Gilbert, J. G., Giuffra, E., Gorodkin, J., Griffin, D. K., Harrow, J. L., Hayward, A., Howe, K., Hu, Z. L., Humphray, S. J., Hunt, T., Hornshoj, H., Jeon, J. T., Jern, P., Jones, M., Jurka, J., Kanamori, H., Kapetanovic, R., Kim, J., Kim, J. H., Kim, K. W., Kim, T. H., Larson, G., Lee, K., Lee, K. T., Leggett, R., Lewin, H. A., Li, Y., Liu, W., Loveland, J. E., Lu, Y., Lunney, J. K., Ma, J., Madsen, O., Mann, K., Matthews, L., McLaren, S., Morozumi, T., Murtaugh, M. P., Narayan, J., Nguyen, D. T., Ni, P., Oh, S. J., Onteru, S., Panitz, F., Park, E. W., et al. (2012) 'Analyses of pig genomes provide insight into porcine demography and evolution', *Nature,* 491(7424), 393-8.

Groth, D., Lehrach, H. and Hennig, S. (2004) 'GOblet: a platform for Gene Ontology annotation of anonymous sequence data', *Nucleic acids research,* 32(suppl 2), W313-W317.

Groves, C. (2007) 'Current views on taxonomy and zoogeography of the genus Sus', *Pigs and Humans,* 10(000), 15-29.

Groves, C. P. and Grubb, P. (1993) 'The Eurasian suids: Sus and Babyrousa', *Pigs, Peccaries and Hippos–Status Survey and Conservation Action Plan*, 107-111.

Grubben, G. J. (2004) *Vegetables,* Prota.

Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. and Cristianini, N. (2005) 'Estimating the tempo and mode of gene family evolution from comparative genomic data', *Genome research,* 15(8), 1153-1160.

Haldane, J. and Smith, C. (1947) 'A new estimate of the linkage between the genes for colour-blindness and haemophilia in man', *Annals of Human Genetics,* 14(1), 10-31.

Hall, B. (2008) *Strickberger's evolution,* Jones & Bartlett Learning.

Halverson, J. D., Smrek, J., Kremer, K. and Grosberg, A. Y. (2014) 'From a melt of rings to chromosome territories: the role of topological constraints in genome folding', *Reports on Progress in Physics,* 77(2), 022601.

Hancks, D. C., Goodier, J. L., Mandal, P. K., Cheung, L. E. and Kazazian, H. H. (2011) 'Retrotransposition of marked SVA elements by human L1s in cultured cells', *Human molecular genetics,* 20(17), 3386-3400.

Hannenhalli, S. and Pevzner, P. (1995) *Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals,* translated by ACM, 178-189.

Harewood, L. and Fraser, P. (2014) 'The Impact Of Chromosomal Rearrangements On Regulation Of Gene Expression', *Human molecular genetics*, ddu278.

Harrison, P. M. and Gerstein, M. (2002) 'Studying genomes through the aeons: protein families, pseudogenes and proteome evolution', *Journal of molecular biology,* 318(5), 1155-1174.

Haussler, D., O'Brien, S. J., Ryder, O. A., Barker, F. K., Clamp, M., Crawford, A. J., Hanner, R., Hanotte, O., Johnson, W. E. and McGuire, J. A. (2009) 'Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species', *J Hered,* 100(6), 659-674.

Haussler, D., O'Brien, S., Ryder, O., Barker, F., Clamp, M., Crawford, A., Hanner, R., Hanotte, O., Johnson, W. and McGuire, J. 'Genome 10K Community of Scientists (2009) Genome 10K: A proposal to obtain whole-genome sequence for 10,000 vertebrate species', *J Hered,* 100, 659-674.

Havlak, P., Chen, R., Durbin, K. J., Egan, A., Ren, Y., Song, X. Z., Weinstock, G. M. and Gibbs, R. A. (2004) 'The Atlas genome assembly system', *Genome Research,* 14(4), 721-732.

Hawken, R. J., Murtaugh, J., Flickinger, G. H., Yerle, M., Robic, A., Milan, D., Gellin, J., Beattie, C. W., Schook, L. B. and Alexander, L. J. (1999) 'A first-generation porcine whole-genome radiation hybrid map', *Mammalian Genome,* 10(8), 824-830.

Hedges, S. and Dudley, J. (2006) 'S. KUMAR (2006): TimeTree: a public knowledge-base of divergence times among organisms', *Bioinformatics,* 22, 2971-2972.

Hefferin, M. L. and Tomkinson, A. E. (2005) 'Mechanism of DNA double-strand break repair by non-homologous end joining', *DNA repair,* 4(6), 639-648.

Hellekant, G. and Danilova, V. (1999) 'Taste in domestic pig, Sus scrofa', *Journal of Animal Physiology and Animal Nutrition,* 82(1), 8-24.

Hernandez, D., François, P., Farinelli, L., Østerås, M. and Schrenzel, J. (2008) 'De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer', *Genome Research,* 18(5), 802-809.

Hert, D. G., Fredlake, C. P. and Barron, A. E. (2008) 'Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods', *Electrophoresis,* 29(23), 4618-4626.

Heyen, D., Weller, J., Ron, M., Band, M., Beever, J., Feldmesser, E., Da, Y., Wiggans, G., VanRaden, P. and Lewin, H. (1999) 'A genome scan for QTL influencing milk production and health traits in dairy cattle', *Physiological Genomics,* 1(3), 165-175.

Hill, D. P., Smith, B., McAndrews-Hill, M. S. and Blake, J. A. (2008) 'Gene Ontology annotations: what they mean and where they come from', *BMC bioinformatics,* 9(Suppl 5), S2.

Hirokawa, N., Tanaka, Y. and Okada, Y. (2009) 'Left-right determination: involvement of molecular motor KIF3, cilia, and nodal flow', *Cold Spring Harb Perspect Biol,* 1(1), a000802.

Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A. and Birney, E. (2012) 'Integrative annotation of chromatin elements from ENCODE data', *Nucleic Acids Research*, gks1284.

Hohjoh, H. and Singer, M. F. (1997) 'Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon', *The EMBO journal,* 16(19), 6034-6043.

Hou, J., Friedrich, A., de Montigny, J. and Schacherer, J. (2014) 'Chromosomal Rearrangements as a Major Mechanism in the Onset of Reproductive Isolation in< i> Saccharomyces cerevisiae</i>', *Current Biology,* 24(10), 1153-1159.

Hoyle, F. and Wickramasinghe, N. C. (2000) *Biological Evolution,* Springer.

Hu, Z.-L., Park, C. A., Wu, X.-L. and Reecy, J. M. (2013) 'Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era', *Nucleic acids research,* 41(D1), D871-D879.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2008) 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nature Protocols,* 4(1), 44-57.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009) 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Research,* 37(1), 1-13.

Huang, X. and Madan, A. (1999) 'CAP3: A DNA sequence assembly program', *Genome Research,* 9(9), 868-877.

Hubisz, M. J., Lin, M. F., Kellis, M. and Siepel, A. (2011) 'Error and error mitigation in low-coverage genome assemblies', *PLoS ONE,* 6(2), e17034.

Hughes, A. L. (2002) 'Adaptive evolution after gene duplication', *Trends in Genetics,* 18(9), 433-434.

Hughes, A. L. and Hughes, M. K. (1995) 'Small genomes for better flyers', *Nature,* 377(6548), 391.

Humphray, S. J., Scott, C. E., Clark, R., Marron, B., Bender, C., Camm, N., Davis, J., Jenks, A., Noon, A. and Patel, M. (2007) 'A high utility integrated map of the pig genome', *Genome Biology,* 8(7), R139.

Humphray, S. J., Scott, C. E., Clark, R., Marron, B., Bender, C., Camm, N., Davis, J., Jenks, A., Noon, A. and Patel, M. (2007) 'A high utility integrated map of the pig genome', *Genome Biol,* 8(7), R139.

Hyman, E. D. (1988) 'A new method of sequencing DNA', *Analytical Biochemistry,* 174(2), 423-436.

Iacia, A. A. S. and Pinto-Maglio, C. A. (2013) 'Mapping pachytene chromosomes of coffee using a modified protocol for fluorescence in situ hybridization', *AoB plants,* 5, plt040.

International Chicken Genome Sequencing, C. (2004) 'Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution', *Nature,* 432(7018), 695-716.

Iriarte, P. F. and Hasson, E. (2000) 'The role of the use of different host plants in the maintenance of the inversion polymorphism in the cactophilic Drosophila buzzatii', *Evolution,* 54(4), 1295-1302.

Iriarte, P. J. F., Norry, F. and Hasson, E. (2003) 'Chromosomal inversions effect body size and shape in different breeding resources in Drosophila buzzatii', *Heredity,* 91(1), 51-59.

Irie, A., Koyama, S., Kozutsumi, Y., Kawasaki, T. and Suzuki, A. (1998) 'The Molecular Basis for the Absence ofN-Glycolylneuraminic Acid in Humans', *Journal of Biological Chemistry,* 273(25), 15866-15871.

Islam, S. M., Shinmyo, Y., Okafuji, T., Su, Y., Naser, I. B., Ahmed, G., Zhang, S., Chen, S., Ohta, K., Kiyonari, H., Abe, T., Tanaka, S., Nishinakamura, R., Terashima, T., Kitamura, T. and Tanaka, H. (2009) 'Draxin, a repulsive guidance protein for spinal cord and forebrain commissures', *Science,* 323(5912), 388-93.

Jarvis, E. D. (2004) 'Learned Birdsong and the Neurobiology of Human Language', *Annals of the New York Academy of Sciences,* 1016(1), 749-777.

Jarvis, E. D. and al., e. (2014) 'Avian phylogenomics paper', *Science,* submitted.

Jean, G. and Nikolski, M. (2011) 'SyDiG: Uncovering synteny in distant genomes', *International Journal of Bioinformatics Research and Applications,* 7(1), 43-62.

Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., Hickenbotham, M. T., Magrini, V., Mardis, E. R., Dangl, J. L. and Jones, C. D. (2007) 'Extending assembly of short DNA sequences to handle error', *Bioinformatics,* 23(21), 2942-2944.

Jensen, R. A. (2001) 'Orthologs and paralogs–we need to get it right', *Genome Biol,* 2(8), 1002.1-1002.3.

Jetz, W., Thomas, G., Joy, J., Hartmann, K. and Mooers, A. (2012) 'The global diversity of birds in space and time', *Nature,* 491(7424), 444-448.

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K. and Mooers, A. O. (2012) 'The global diversity of birds in space and time', *Nature,* 491(7424), 444-448.

Jiang, N., Bao, Z., Zhang, X., Eddy, S. R. and Wessler, S. R. (2004) 'Pack-MULE transposable elements mediate gene evolution in plants', *Nature,* 431(7008), 569-573.

Jiang, Z., Michal, J. J., Melville, J. S. and Baltzer, H. L. (2005) 'Multi-alignment of orthologous genome regions in five species provides new insights into the

evolutionary make-up of mammalian genomes', *Chromosome Research,* 13(7), 707-715.

Jiang, Z. and Rothschild, M. F. (2007) 'Swine genome science comes of age', *International journal of biological sciences,* 3(3), 129.

Johannsen, W. (1911) 'The genotype conception of heredity', *The American Naturalist,* 45(531), 129-159.

Johansson, M., Ellegren, H. and Andersson, L. (1995) 'Comparative mapping reveals extensive linkage conservation--but with gene order rearrangements--between the pig and the human genomes', *Genomics,* 25(3), 682-690.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) 'Repbase Update, a database of eukaryotic repetitive elements', *Cytogenet Genome Res,* 110(1-4), 462-467.

Kazazian, H. H. (2004) 'Mobile elements: drivers of genome evolution', *Science,* 303(5664), 1626-1632.

Kazazian Jr, H. H. (1998) 'Mobile elements and disease', *Current opinion in genetics & development,* 8(3), 343-350.

Kececioglu, J. and Sankoff, D. (1995) 'Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement', *Algorithmica,* 13(1), 180-210.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003) 'Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes', *Proceedings of the National Academy of Sciences of the United States of America,* 100(20), 11484.

Kent, W. J. and Haussler, D. (2001) 'Assembly of the working draft of the human genome with GigAssembler', *Genome Research,* 11(9), 1541-1548.

Khan, S., Situ, G., Decker, K. and Schmidt, C. J. (2003) 'GoFigure: Automated Gene Ontology™ annotation', *Bioinformatics,* 19(18), 2484-2485.

Khandelwal, S. (1990) 'Chromosome evolution in the genus Ophioglossum L', *Botanical journal of the Linnean Society,* 102(3), 205-217.

Kim, J., Larkin, D. M., Cai, Q., Zhang, Y., Ge, R.-L., Auvil, L., Capitanu, B., Zhang, G., Lewin, H. A. and Ma, J. (2013) 'Reference-assisted chromosome assembly', *Proceedings of the National Academy of Sciences,* 110(5), 1785-1790.

King, M. (1995) *Species evolution: the role of chromosome change,* Cambridge University Press.

Kogelman, L. J., Kadarmideen, H. N., Mark, T., Karlskov-Mortensen, P., Bruun, C. S., Cirera, S., Jacobsen, M. J., Jørgensen, C. B. and Fredholm, M. (2013) 'An F2 pig resource population as a model for genetic studies of obesity and obesity-related diseases in humans: design and genetic parameters', *Frontiers in genetics,* 4.

Kohn, M. H., Murphy, W. J., Ostrander, E. A. and Wayne, R. K. (2006) 'Genomics and conservation genetics', *Trends Ecol Evol,* 21(11), 629-637.

Kordis, D. (2010) 'Transposable elements in reptilian and avian (sauropsida) genomes', *Cytogenet Genome Res,* 127(2-4), 94-111.

Kornberg, R. D. (1974) 'Chromatin structure: a repeating unit of histones and DNA', *Science,* 184(4139), 868-871.

Kumar, S. and Hedges, S. B. (1998) 'A molecular timescale for vertebrate evolution', *Nature,* 392(6679), 917-920.

Kwitek, A. E., Gullings-Handley, J., Yu, J., Carlos, D. C., Orlebeke, K., Nie, J., Eckert, J., Lemke, A., Andrae, J. W. and Bromberg, S. (2004) 'High-density rat radiation hybrid maps containing over 24,000 SSLPs, genes, and ESTs provide a direct link to the rat genome sequence', *Genome Research,* 14(4), 750-757.

Ladeveze, V., Aulard, S., Aulard, N., Periquet, G. and Lemeunier, F. (1998) 'Hobo transposons causing chromosomal breakpoints', *Proceedings of the Royal Society of London. Series B: Biological Sciences,* 265(1402), 1157-1159.

Lahbib-Mansais, Y., Mompart, F., Milan, D., Leroux, S., Faraut, T., Delcros, C. and Yerle, M. (2006) 'Evolutionary breakpoints through a high-resolution comparative map between porcine chromosomes 2 and 16 and human chromosomes', *Genomics,* 88(4), 504-512.

Lajoie, B. R., van Berkum, N. L., Sanyal, A. and Dekker, J. (2009) 'My5C: webtools for chromosome conformation capture studies', *Nature methods,* 6(10), 690.

Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G. and Cremer, T. (2007) 'Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions', *Nature Reviews Genetics,* 8(2), 104-115.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M. and FitzHugh, W. (2001) 'Initial sequencing and analysis of the human genome', *Nature,* 409(6822), 860-921.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001) 'Initial sequencing and analysis of the human genome', *Nature,* 409(6822), 860-921.

Lander, E. S. and Waterman, M. S. (1988) 'Genomic mapping by fingerprinting random clones: a mathematical analysis', *Genomics,* 2(3), 231-239.

Langkjær, R. B., Nielsen, M. L., Daugaard, P., Liu, W. and Piskur, J. (2000) 'Yeast chromosomes have been significantly reshaped during their evolutionary history1', *Journal of molecular biology,* 304(3), 271-288.

Larkin, D. (2011) 'Status of the cattle genome map', *Cytogenet Genome Res,* 134(1), 1-8.

Larkin, D. M. (2012) 'Cattle Comparative Genomics and Chromosomal Evolution', *Bovine Genomics*, 101.

Larkin, D. M., Everts-van der Wind, A., Rebeiz, M., Schweitzer, P. A., Bachman, S., Green, C., Wright, C. L., Campos, E. J., Benson, L. D. and Edwards, J. (2003) 'A cattle–human comparative map built with cattle BAC-ends and human genome sequence', *Genome Research,* 13(8), 1966-1972.

Larkin, D. M., Pape, G., Donthu, R., Auvil, L., Welge, M. and Lewin, H. A. (2009) 'Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories', *Genome Research,* 19(5), 770-777.

Larson, G., Albarella, U., Dobney, K. and Rowley-Conwy, P. (2007) 'Current views on Sus phylogeography and pig domestication as seen through modern mtDNA studies', *Pigs and humans: 10,000 years of interaction*, 30.

Latinis, D. K. (2000) 'The development of subsistence system models for Island Southeast Asia and Near Oceania: the nature and role of arboriculture and arboreal-based economies', *World Archaeology,* 32(1), 41-67.

Lechner, M., Hernandez-Rosales, M., Doerr, D., Wieseke, N., Thévenin, A., Stoye, J., Hartmann, R. K., Prohaska, S. J. and Stadler, P. F. (2014) 'Orthology detection combining clustering and synteny for very large datasets', *PLoS ONE,* 9(8), e105015.

Lemaitre, C., Zaghloul, L., Sagot, M. F., Gautier, C., Arneodo, A., Tannier, E. and Audit, B. (2009) 'Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation', *BMC Genomics,* 10(1), 335.

Lemay, D. G., Lynn, D. J., Martin, W. F., Neville, M. C., Casey, T. M., Rincon, G., Kriventseva, E. V., Barris, W. C., Hinrichs, A. S. and Molenaar, A. J. (2009) 'The bovine lactation genome: insights into the evolution of mammalian milk', *Genome Biology,* 10(4), R43.

Lemay, D. G., Lynn, D. J., Martin, W. F., Neville, M. C., Casey, T. M., Rincon, G., Kriventseva, E. V., Barris, W. C., Hinrichs, A. S., Molenaar, A. J., Pollard, K. S., Maqbool, N. J., Singh, K., Murney, R., Zdobnov, E. M., Tellam, R. L., Medrano, J. F., German, J. B. and Rijnkels, M. (2009) 'The bovine lactation genome: insights into the evolution of mammalian milk', *Genome Biology,* 10(4).

Levin, H. L. and Moran, J. V. (2011) 'Dynamic interactions between transposable elements and their hosts', *Nature Reviews Genetics,* 12(9), 615-627.

Levine, R. (2011) 'i5k: The 5,000 Insect Genome Project', *American Entomologist,* 57(2), 110-113.

Lewin, H. A., Larkin, D. M., Pontius, J. and O'Brien, S. J. (2009) 'Every genome sequence needs a good map', *Genome research,* 19(11), 1925-1928.

Lewis, S. E. (2005) 'Gene Ontology: looking backwards and forwards', *Genome Biol,* 6(1), 103.

Li, L., Stoeckert, C. J. and Roos, D. S. (2003) 'OrthoMCL: identification of ortholog groups for eukaryotic genomes', *Genome research,* 13(9), 2178-2189.

Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008) 'SOAP: short oligonucleotide alignment program', *Bioinformatics,* 24(5), 713-714.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G. and Kristiansen, K. (2010) 'De novo assembly of human genomes with massively parallel short read sequencing', *Genome Research,* 20(2), 265-272.

Lim, J. K. and Simmons, M. J. (1994) 'Gross chromosome rearrangements mediated by transposable elements in Drosophila melanogaster', *Bioessays,* 16(4), 269-275.

Lin, J., Qi, R., Aston, C., Jing, J., Anantharaman, T. S., Mishra, B., White, O., Daly, M. J., Minton, K. W. and Venter, J. C. (1999) 'Whole-genome shotgun optical mapping of Deinococcus radiodurans', *Science,* 285(5433), 1558-1562.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, C. P., Galibert, F., Smith, D. R., DeJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.-W., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.-P., Parker, H. G., Pollinger, J. P., Searle, S. M. J., Sutter, N. B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-Zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn, B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill, P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., et al. (2005) 'Genome sequence, comparative analysis and haplotype structure of the domestic dog', *Nature,* 438(7069), 803-19.

Liti, G., Peruffo, A., James, S. A., Roberts, I. N. and Louis, E. J. (2005) 'Inferences of evolutionary relationships from a population survey of LTR-retrotransposons and telomeric-associated sequences in the Saccharomyces sensu stricto complex', *Yeast,* 22(3), 177-192.

Llorente, B., Malpertuy, A., Neuvéglise, C., de Montigny, J., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E. and Brottier, P. (2000) 'Genomic Exploration of the Hemiascomycetous Yeasts: 18. Comparative analysis of chromosome maps and synteny with< i> Saccharomyces cerevisiae</i>', *FEBS letters,* 487(1), 101-112.

Lodish, H. (2008) *Molecular cell biology,* Macmillan.

Lomvardas, S., Barnea, G., Pisapia, D. J., Mendelsohn, M., Kirkland, J. and Axel, R. (2006) 'Interchromosomal interactions and olfactory receptor choice', *Cell,* 126(2), 403-413.

Longo, M. S. and Carone, D. M. (2009) 'Distinct retroelement classes define evolutionary breakpoints demarcating sites of evolutionary novelty', *BMC Genomics,* 10(1), 334.

Longo, M. S., Carone, D. M., Green, E. D., O'Neill, M. J. and O'Neill, R. J. (2009) 'Distinct retroelement classes define evolutionary breakpoints demarcating sites of evolutionary novelty', *BMC Genomics,* 10(1), 334.

Lönnig, W.-E. and Saedler, H. (2002) 'Chromosome rearrangements and transposable elements', *Annual review of genetics,* 36(1), 389-410.

Lowe, C. B. and Haussler, D. (2012) '29 Mammalian Genomes Reveal Novel Exaptations of Mobile Elements for Likely Regulatory Functions in the Human Genome', *PLoS One,* 7(8), e43128.

Lu, J., Li, W.-H. and Wu, C.-I. (2003) 'Comment on" Chromosomal speciation and molecular divergence-Accelerated evolution in rearranged chromosomes"', *Science,* 302(5647), 988-988.

Lu, J., Li, W. H. and Wu, C. I. (2003) 'Comment on" Chromosomal Speciation and Molecular Divergence-Accelerated Evolution in Rearranged Chromosomes"', *Science,* 302(5647), 988-988.

Lucas, J. M., Muffato, M. and Crollius, H. R. (2014) 'PhylDiag: identifying complex synteny blocks that include tandem duplications using phylogenetic gene trees', *BMC bioinformatics,* 15(1), 268.

Luger, K. and Hansen, J. C. (2005) 'Nucleosome and chromatin fiber dynamics', *Current opinion in structural biology,* 15(2), 188-196.

Lunney, J. K. (2007) 'Advances in swine biomedical model genomics', *International journal of biological sciences,* 3(3), 179.

Lupski, J. R. and Weinstock, G. M. (1992) 'Short, interspersed repetitive DNA sequences in prokaryotic genomes', *Journal of bacteriology,* 174(14), 4525.

Lutz, B., Kuratani, S., Cooney, A. J., Wawersik, S., Tsai, S. Y., Eichele, G. and Tsai, M. J. (1994) 'Developmental regulation of the orphan receptor COUP-TF II gene in spinal motor neurons', *Development,* 120(1), 25-36.

Lysák, M. A. and Schubert, I. (2013) *Mechanisms of chromosome rearrangements,* Springer.

Ma, J., Zhang, L., Suh, B. B., Raney, B. J., Burhans, R. C., Kent, W. J., Blanchette, M., Haussler, D. and Miller, W. (2006) 'Reconstructing contiguous regions of an ancestral genome', *Genome Research,* 16(12), 1557-1565.

Ma, W., Yan, R.-T., Li, X. and Wang, S.-Z. (2009) 'Reprogramming Retinal Pigment Epithelium to Differentiate Toward Retinal Neurons with Sox2', *STEM CELLS,* 27(6), 1376-1387.

Maere, S., Heymans, K. and Kuiper, M. (2005) 'BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks', *Bioinformatics,* 21(16), 3448-3449.

Malik, H. S. and Eickbush, T. H. (1998) 'The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs', *Molecular biology and evolution,* 15(9), 1123-1134.

Malin, J., Aniba, M. R. and Hannenhalli, S. (2013) 'Enhancer networks revealed by correlated DNAse hypersensitivity states of enhancers', *Nucleic Acids Research*, gkt374.

Mardis, E. R. (2008) 'Next-generation DNA sequencing methods', *Annu Rev Genomics Hum Genet,* 9, 387-402.

Mardis, E. R. (2011) 'A decade/'s perspective on DNA sequencing technology', *Nature,* 470(7333), 198-203.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J. and Chen, Z. (2005) 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature,* 437(7057), 376-380.

Marques-Bonet, T., Caceres, M., Bertranpetit, J., Preuss, T. M., Thomas, J. W. and Navarro, A. (2004) 'Chromosomal rearrangements and the genomic distribution of gene-expression divergence in humans and chimpanzees', *Trends in genetics : TIG,* 20(11), 524-529.

Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A. and Kaessmann, H. (2005) 'Emergence of young human genes after a burst of retroposition in primates', *PLoS Biol,* 3(11), e357.

Martin, S. L. and Bushman, F. D. (2001) 'Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon', *Molecular and cellular biology,* 21(2), 467-475.

Mathias, S. L., Scott, A. F., Kazazian, H. H., Boeke, J. D. and Gabriel, A. (1991) 'Reverse transcriptase encoded by a human transposable element', *Science,* 254(5039), 1808-1810.

Matoušková, M., Veselý, P., Daniel, P., Mattiuzzo, G., Hector, R. D., Scobie, L., Takeuchi, Y. and Hejnar, J. (2013) 'Role of DNA Methylation in Expression and Transmission of Porcine Endogenous Retroviruses', *Journal of virology,* 87(22), 12110-12120.

Maxam, A. M. and Gilbert, W. (1977) 'A new method for sequencing DNA', *Proceedings of the National Academy of Sciences,* 74(2), 560.

Mayr, G. and De Pietri, V. L. (2014) 'Earliest and first Northern Hemispheric hoatzin fossils substantiate Old World origin of a "Neotropic endemic"', *Naturwissenschaften.*

McClintock, B. (1947) 'Cytogenetic studies of maize and Neurospora', *Carnegie Inst. Washington Year Book,* 46, 146-152.

McClintock, B. (1950) 'The origin and behavior of mutable loci in maize', *Proceedings of the National Academy of Sciences,* 36(6), 344-355.

McClintock, B. (1965) 'Components of action of the regulators Spm and Ac', *Carnegie Inst Wash Year Book,* 64, 527-536.

McPherson, J. D., Marra, M., Hillier, L. D., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R. and Wilson, R. K. (2001) 'A physical map of the human genome', *Nature,* 409(6822), 934-941.

Medstrand, P., Van de Lagemaat, L., Dunn, C. A., Landry, J.-R., Svenback, D. and Mager, D. L. (2005) 'Impact of transposable elements on the evolution of mammalian gene regulation', *Cytogenet Genome Res,* 110(1-4), 342-352.

Mendel, G. (1865) *Experiments in plant hybridization (1865),* translated by.

Metzker, M. L. (2009) 'Sequencing technologies—the next generation', *Nature Reviews Genetics,* 11(1), 31-46.

Meyer, A. and Schartl, M. (1999) 'Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions', *Current opinion in cell biology,* 11(6), 699-704.

Meyers, S. N., Rogatcheva, M. B., Larkin, D. M., Yerle, M., Milan, D., Hawken, R. J., Schook, L. B. and Beever, J. E. (2005) 'Piggy-BACing the human genome: II. A high-resolution, physically anchored, comparative map of the porcine autosomes', *Genomics,* 86(6), 739-752.

Michler, R. E. (1996) 'Xenotransplantation: risks, clinical potential, and future prospects', *Emerging infectious diseases,* 2(1), 64.

Miller, J. R., Koren, S. and Sutton, G. (2010) 'Assembly algorithms for next-generation sequencing data', *Genomics,* 95(6), 315-327.

Mills, R. E., Bennett, E. A., Iskow, R. C. and Devine, S. E. (2007) 'Which transposable elements are active in the human genome?', *Trends in Genetics,* 23(4), 183-191.

Minkin, I., Patel, A., Kolmogorov, M., Vyahhi, N. and Pham, S. (2013) 'Sibelia: A scalable and comprehensive synteny block generation tool for closely related microbial genomes' in *Algorithms in Bioinformatics*, Springer, 215-229.

Miyata, T., Miyazawa, S. and Yasunaga, T. (1979) 'Two types of amino acid substitutions in protein evolution', *Journal of molecular evolution,* 12(3), 219-236.

Morgan, T. H. (1910) 'Sex limited inheritance in Drosophila', *Science,* 32(812), 120-122.

Mullikin, J. C. and Ning, Z. (2003) 'The phusion assembler', *Genome Research,* 13(1), 81-90.

Mullis, K. B. (1994) 'The polymerase chain reaction', *NOBEL LECTURES IN CHEMISTRY 1991-1995*, 103.

Mullis, K. B., Ferre, F., Gibbs, R. and Morley, B. J. (1995) 'PCR-The polymerase chain reaction', *Trends in Genetics,* 11(6), 249-249.

Munoz-Lopez, M. and García-Pérez, J. L. (2010) 'DNA transposons: nature and applications in genomics', *Current genomics,* 11(2), 115.

Murphy, W. J., Larkin, D. M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F. and Gatzke, L. (2005) *Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps*, unpublished thesis

Murphy, W. J., Larkin, D. M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F. and Gatzke, L. (2005) 'Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps', *Science,* 309(5734), 613-617.

Murphy, W. J., Larkin, D. M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F., Gatzke, L., Hitte, C., Meyers, S. N., Milan, D., Ostrander, E. A., Pape, G., Parker, H. G., Raudsepp, T., Rogatcheva, M. B., Schook, L. B., Skow, L. C., Welge, M., Womack, J. E., O'Brien, S. J., Pevzner, P. A. and Lewin, H. A. (2005) 'Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps', *Science (New York, N.Y.),* 309(5734), 613-617.

Murphy, W. J., Sun, S., Chen, Z., Yuhki, N., Hirschmann, D., Menotti-Raymond, M. and O'Brien, S. J. (2000) 'A radiation hybrid map of the cat genome: implications for comparative mapping', *Genome Research,* 10(5), 691-702.

Murrell, A., Heeson, S. and Reik, W. (2004) 'Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops', *Nature genetics,* 36(8), 889-893.

Myers, E. W. (1995) 'Toward simplifying and accurately formulating fragment assembly', *Journal of Computational Biology,* 2(2), 275-290.

Myers, E. W. (2005) 'The fragment assembly string graph', *Bioinformatics,* 21(suppl 2), ii79-ii85.

Nadeau, J. H. and Taylor, B. A. (1984) 'Lengths of chromosomal segments conserved since divergence of man and mouse', *Proceedings of the National Academy of Sciences,* 81(3), 814.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A. and Fraser, P. (2013) 'Single-cell Hi-C reveals cell-to-cell variability in chromosome structure', *Nature,* 502(7469), 59-64.

Navarro, A. and Barton, N. H. (2003) 'Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes', *Science,* 300(5617), 321-324.

Nellåker, C., Keane, T. M., Yalcin, B., Wong, K., Agam, A., Belgard, T. G., Flint, J., Adams, D. J., Frankel, W. N. and Ponting, C. P. (2012) 'The genomic landscape shaped by selection on transposable elements across 18 mouse strains', *Genome Biol,* 13(6), R45.

Nelson, S. L. and Sanregret, J. D. (1997) 'Response of pigs to bitter-tasting compounds', *Chemical senses,* 22(2), 129-132.

Nevers, P. and Saedler, H. (1977) 'Transposable genetic elements as agents of gene instability and chromosomal rearrangements', *Nature,* 268(5616), 109-115.

Ng, P. C. and Henikoff, S. (2006) 'Predicting the effects of amino acid substitutions on protein function', *Annu. Rev. Genomics Hum. Genet.,* 7, 61-80.

Noor, M. A. F., Grams, K. L., Bertucci, L. A., Almendarez, Y., Reiland, J. and Smith, K. R. (2001) 'The genetics of reproductive isolation and the potential for gene exchange between Drosophila pseudoobscura and D. persimilis via backcross hybrid males', *Evolution,* 55(3), 512-521.

Nowacki, M., Higgins, B. P., Maquilan, G. M., Swart, E. C., Doak, T. G. and Landweber, L. F. (2009) 'A functional role for transposases in a large eukaryotic genome', *Science,* 324(5929), 935-938.

O'Brien, S. J. and Nash, W. G. (1982) 'Genetic mapping in mammals: chromosome map of domestic cat', *Science,* 216(4543), 257.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) 'KEGG: Kyoto encyclopedia of genes and genomes', *Nucleic Acids Research,* 27(1), 29-34.

Ohlsson, R. and Göndör, A. (2007) 'The 4C technique: the 'Rosetta stone'for genome biology in 3D?', *Current opinion in cell biology,* 19(3), 321-325.

Ohno, S. (1970) *Evolution by gene duplication,* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.

Ohno, S. (1973) 'Ancient linkage groups and frozen accidents', *Nature,* 244, 259-262.

Oliver, S. G., Van der Aart, Q., Agostoni-Carbone, M., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J. and Benit, P. (1992) 'The complete DNA sequence of yeast chromosome III'.

Onagbesan, O. M., Bruggeman, V., Van As, P., Tona, K., Williams, J. and Decuypere, E. (2003) 'BMPs and BMPRs in chicken ovary and effects of BMP-4 and -7 on granulosa cell proliferation and progesterone production in vitro', *American Journal of Physiology - Endocrinology and Metabolism,* 285(5), E973-E983.

Ophir, R. and Graur, D. (1997) 'Patterns and rates of indel evolution in processed pseudogenes from humans and murids', *Gene,* 205(1), 191-202.

Oppenheimer, S. and Richards, M. (2001) 'Fast trains, slow boats, and the ancestry of the Polynesian islanders', *Science Progress,* 84(3), 157-181.

Padian, K. and Chiappe, L. M. (1998) 'The origin of birds and their flight', *Scientific American,* 278(2), 28-37.

Pan, X., Stein, L. and Brendel, V. (2005) 'SynBrowse: a synteny browser for comparative sequence analysis', *Bioinformatics,* 21(17), 3461-3468.

Pederson, T. (2004) 'The spatial organization of the genome in mammalian cells', *Current opinion in genetics & development,* 14(2), 203-209.

Penny, D. (2012) 'Evolution: A View from the 21st Century', *Systematic Biology,* 61(4), 709-710.

Pérez-Ortín, J. E., Querol, A., Puig, S. and Barrio, E. (2002) 'Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains', *Genome Research,* 12(10), 1533-1539.

Pevzner, P. and Tesler, G. (2003a) 'Genome rearrangements in mammalian evolution: lessons from human and mouse genomes', *Genome Research,* 13(1), 37-45.

Pevzner, P. and Tesler, G. (2003b) 'Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution', *Proceedings of the National Academy of Sciences of the United States of America,* 100(13), 7672-7677.

Pevzner, P. A., Tang, H. and Waterman, M. S. (2001) 'An Eulerian path approach to DNA fragment assembly', *Proceedings of the National Academy of Sciences,* 98(17), 9748.

Pilling, D. and Rischkowsky, B. (2007) *The state of the world's animal genetic resources for food and agriculture,* Food & Agriculture Org.

Pinton, A., Ducos, A. and Yerle, M. (2003) 'Chromosomal rearrangements in cattle and pig revealed by chromosome microdissection and chromosome painting', *Genetics Selection Evolution,* 35(6), 685-696.

Pontius, J. U., Mullikin, J. C., Smith, D. R., Lindblad-Toh, K., Gnerre, S., Clamp, M., Chang, J., Stephens, R., Neelam, B., Volfovsky, N., Schäffer, A. A., Agarwala, R., Narfström, K., Murphy, W. J., Giger, U., Roca, A. L., Antunes, A., Menotti-Raymond, M., Yuhki, N., Pecon-Slattery, J., Johnson, W. E., Bourque, G., Tesler, G. and O'Brien, S. J. (2007) 'Initial sequence and comparative analysis of the cat genome', *Genome Research,* 17(11), 1675-1689.

Pop, M. and Salzberg, S. L. (2008) 'Bioinformatics challenges of new sequencing technology', *Trends in Genetics,* 24(3), 142-149.

Prather, R. S. (2013) 'Pig genomics for biomedicine', *Nature biotechnology,* 31(2), 122-124.

Pray, L. (2008) 'Eukaryotic genome complexity', *Nature Education,* 1(1).

Price, S. A., Bininda-Emonds, O. R. and Gittleman, J. L. (2005) 'A complete phylogeny of the whales, dolphins and even-toed hoofed mammals (Cetartiodactyla)', *Biological reviews,* 80(3), 445-473.

Pryszcz, L. P., Huerta-Cepas, J. and Gabaldón, T. (2010) 'MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score', *Nucleic Acids Research*, gkq953.

Purves, W. K., Orians, G. H., Sadava, D. and Heller, H. C. (2003) *Life: The Science of Biology: Volume III: Plants and Animals,* Macmillan.

Ramos, A., Helm, J., Sherwood, J., Rocha, D. and Rothschild, M. (2006) 'Mapping of 21 genetic markers to a QTL region for meat quality on pig chromosome 17', *Animal genetics,* 37(3), 296-297.

Ranz, J. M., Maurin, D., Chan, Y. S., Von Grotthuss, M., Hillier, L. W., Roote, J., Ashburner, M. and Bergman, C. M. (2007) 'Principles of genome evolution in the Drosophila melanogaster species group', *PLoS Biol,* 5(6), e152.

Rattink, A. P., De Koning, D. J., Faivre, M., Harlizius, B., van Arendonk, J. A. and Groenen, M. A. (2000) 'Fine mapping and imprinting analysis for fatness trait QTLs in pigs', *Mammalian Genome,* 11(8), 656-661.

Rattink, A. P., Faivre, M., Jungerius, B. J., Groenen, M. A. and Harlizius, B. (2001) 'A high-resolution comparative RH map of porcine chromosome (SSC) 2', *Mammalian Genome,* 12(5), 366-370.

Rebollo, R., Zhang, Y. and Mager, D. L. (2012) 'Transposable elements: not as quiet as a mouse', *Genome Biology,* 13(6), 159.

Reilly, M. T., Faulkner, G. J., Dubnau, J., Ponomarev, I. and Gage, F. H. (2013) 'The Role of Transposable Elements in Health and Diseases of the Central Nervous System', *The Journal of Neuroscience,* 33(45), 17577-17586.

Renwick, J. (1971) 'The mapping of human chromosomes', *Annual review of genetics,* 5(1), 81-120.

Revanna, K. V., Chiu, C.-C., Bierschank, E. and Dong, Q. (2011) 'GSV: a web-based genome synteny viewer for customized data', *BMC bioinformatics,* 12(1), 316.

Rieseberg, L. H. (2001) 'Chromosomal rearrangements and speciation', *Trends Ecol Evol,* 16(7), 351-358.

Rieseberg, L. H. (2001) 'Chromosomal rearrangements and speciation', *Trends Ecol Evol,* 16(7), 351-358.

Rieseberg, L. H., Linder, C. R. and Seiler, G. J. (1995) 'Chromosomal and genic barriers to introgression in Helianthus', *Genetics,* 141(3), 1163.

Rink, A., Santschi, E. M., Eyer, K. M., Roelofs, B., Hess, M., Godfrey, M., Karajusuf, E. K., Yerle, M., Milan, D. and Beattie, C. W. (2002) 'A first-generation EST RH comparative map of the porcine and human genome', *Mammalian Genome,* 13(10), 578-587.

Robertson, W. and Rees, B. (1916) 'Chromosome studies. I. Taxonomic relationships shown in the chromosomes of tettigidae and acrididae: V-shaped chromosomes and their significance in acrididae, locustidae, and gryllidae: Chromosomes and variation', *Journal of morphology,* 27(2), 179-331.

Roeder, G. S. and Fink, G. R. (1980) 'DNA rearrangements associated with a transposable element in yeast', *Cell,* 21(1), 239-249.

Romanov, M. N., Farre, M., Lightgow, P. E., Fowler, K. E., Skinner, B. M., O'Connor, R., Vignal, A., Faraut, T., Backström, N., Jarvis, E. D., Matsuda, Y., Nishida, C., Zhang, G., Houde, P., Ellegren, H., Burt, D. W., Larkin, D. M. and Griffin, D. K. (2014) 'Reconstruction of the avian genome organization and evolution from a chromosomal perspective suggests that chicken (Gallus gallus) most closely resembles the dinosaur avian ancestor', *Genome Biology,* submitted.

Ross, S. M. (1996) *Simulation,* 2nd Edition ed., San Diego: Academic Press.

Roth, C., Rastogi, S., Arvestad, L., Dittmar, K., Light, S., Ekman, D. and Liberles, D. A. (2007) 'Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms', *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution,* 308(1), 58-73.

Rothschild, M. F. and Ruvinsky, A. (2011) *The genetics of the pig,* CABI.

Rouzankina, I., Abate-Shen, C. and Niswander, L. (2004) 'Dlx genes integrate positive and negative signals during feather bud development', *Dev Biol,* 265(1), 219-33.

Ruiz-Herrera, A., Castresana, J. and Robinson, T. J. (2006) 'Is mammalian chromosomal evolution driven by regions of genome fragility?', *Genome Biology,* 7(12), R115.

Ruiz-Herrera, A., Farre, M. and Robinson, T. J. (2012) 'Molecular cytogenetic and genomic insights into chromosomal evolution', *Heredity,* 108(1), 28-36.

Ruiz-Herrera, A., Garcia, F., Giulotto, E., Attolini, C., Egozcue, J., Ponsa, M. and Garcia, M. (2004) 'Evolutionary breakpoints are co-localized with fragile sites and intrachromosomal telomeric sequences in primates', *Cytogenet Genome Res,* 108(1-3), 234-247.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences,* 74(12), 5463.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences,* 74(12), 5463-5467.

Sankoff, D. and Trinh, P. (2004) *Chromosomal breakpoint re-use in the inference of genome sequence rearrangement,* translated by ACM, 30-35.

Sankoff, D. and Trinh, P. (2005) 'Chromosomal breakpoint reuse in genome sequence rearrangement', *Journal of Computational Biology,* 12(6), 812-821.

SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P. S., Edwards, K. J., Lee, M. and Avramova, Z. (1996) 'Nested retrotransposons in the intergenic regions of the maize genome', *Science,* 274(5288), 765-768.

Schadt, E. E., Turner, S. and Kasarskis, A. (2010) 'A window into third-generation sequencing', *Human molecular genetics,* 19(R2), R227-R240.

Schibler, L., Roig, A., Mahe, M. F., Laurent, P., Hayes, H., Rodolphe, F. and Cribiu, E. P. (2006) 'High-resolution comparative mapping among man, cattle and mouse suggests a role for repeat sequences in mammalian genome evolution', *BMC Genomics,* 7.

Schmidt, B., Sinha, R., Beresford-Smith, B. and Puglisi, S. J. (2009) 'A fast hybrid short read fragment assembly algorithm', *Bioinformatics,* 25(17), 2279-2280.

Schmitz, J. M., Graham, R. M., Sagalowsky, A. and Pettinger, W. A. (1981) 'Renal alpha-1 and alpha-2 adrenergic receptors: biochemical and pharmacological correlations', *Journal of Pharmacology and Experimental Therapeutics,* 219(2), 400-406.

Schmoeckel, M., Bhatti, F. N. K., Zaidi, A., Cozzi, E., Waterworth, P. D., Tolan, M. J., Goddard, M., Warner, R. G., Langford, G. A. and Dunning, J. J. (1998) 'Orthotopic Heart Transplantation in A Transgenic Pig-To-Primate Model1', *Transplantation,* 65(12), 1570.

Schook, L. B., Beever, J. E., Rogers, J., Humphray, S., Archibald, A., Chardon, P., Milan, D., Rohrer, G. and Eversole, K. (2005) 'Swine Genome Sequencing Consortium (SGSC): a strategic roadmap for sequencing the pig genome', *Comparative and functional genomics,* 6(4), 251-255.

Schreiber, F., Pick, K., Erpenbeck, D., Wörheide, G. and Morgenstern, B. (2009) 'OrthoSelect: a protocol for selecting orthologous groups in phylogenomics', *BMC bioinformatics,* 10(1), 219.

Schwab, M. and Amler, L. C. (1990) 'Amplification of cellular oncogenes: a predictor of clinical outcome in human cancer', *Genes, Chromosomes and Cancer,* 1(3), 181-193.

Schwartz, D. C., Li, X., Hernandez, L. I., Ramnarain, S. P., Huff, E. J. and Wang, Y. K. (1993) 'Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping', *Science,* 262(5130), 110.

Schwartz, R. M. and Dayhoff, M. O. (1978) 'Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts', *Science,* 199(4327), 395-403.

Schwarzacher, T. and Heslop-Harrison, P. (2000) *Practical in situ hybridization,* BIOS Scientific Publishers Ltd.

Scordilis, G. E., Ree, H. and Lessie, T. (1987) 'Identification of transposable elements which activate gene expression in Pseudomonas cepacia', *Journal of bacteriology,* 169(1), 8-13.

Sela, N., Kim, E. and Ast, G. (2010) 'The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates', *Genome biology,* 11(6), R59.

Servedio, M. R., Doorn, G., Kopp, M., Frame, A. M. and Nosil, P. (2011) 'Magic traits in speciation:'magic'but not rare?', *Trends Ecol Evol,* 26(8), 389-397.

Shah, N. and Fedoroff, N. V. (2004) 'CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology', *Bioinformatics,* 20(7), 1196-1197.

Shedlock, A. M. (2006) 'Phylogenomic investigation of CR1 LINE diversity in reptiles', *Syst Biol,* 55(6), 902-11.

Shedlock, A. M., Botka, C. W., Zhao, S., Shetty, J., Zhang, T., Liu, J. S., Deschavanne, P. J. and Edwards, S. V. (2007) 'Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome', *Proceedings of the National Academy of Sciences,* 104(8), 2767-2772.

Shendure, J. and Ji, H. (2008) 'Next-generation DNA sequencing', *Nature biotechnology,* 26(10), 1135-1145.

Shi, G., Zhang, L. and Jiang, T. (2010) 'MSOAR 2.0: Incorporating tandem duplications into ortholog assignment based on genome rearrangement', *BMC bioinformatics,* 11(1), 10.

Shimamura, M., Abe, H., Nikaido, M., Ohshima, K. and Okada, N. (1999) 'Genealogy of families of SINEs in cetaceans and artiodactyls: the presence of a huge superfamily of tRNA (Glu)-derived families of SINEs', *Molecular biology and evolution,* 16(8), 1046-1060.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. and Birol, İ. (2009) 'ABySS: a parallel assembler for short read sequence data', *Genome Research,* 19(6), 1117-1123.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A. and Ponting, C. P. (2014) 'Sequencing depth and coverage: key considerations in genomic analyses', *Nature Reviews Genetics,* 15(2), 121-132.

Sinha, A. U. and Meller, J. (2007) 'Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms', *BMC bioinformatics,* 8(1), 82.

Sites, J. W. and Moritz, C. (1987) 'Chromosomal evolution and speciation revisited', *Systematic Biology,* 36(2), 153-174.

Siva, N. (2008) '1000 Genomes project', *Nature biotechnology,* 26(3), 256-256.

Skinner, B. M. and Griffin, D. K. (2012) 'Intrachromosomal rearrangements in avian genome evolution: evidence for regions prone to breakpoints', *Heredity,* 108(1), 37-41.

Smallwood, A. and Ren, B. (2013) 'Genome organization and long-range regulation of gene expression by enhancers', *Current opinion in cell biology,* 25(3), 387-394.

Smit, A., Hubley, R. and Green, P. (2004) 'RepeatMasker Open-3.0'.

Smit, A. F. (1999) 'Interspersed repeats and other mementos of transposable elements in mammalian genomes', *Current opinion in genetics & development,* 9(6), 657-663.

Soderlund, C., Nelson, W., Shoemaker, A. and Paterson, A. (2006) 'SyMAP: A system for discovering and viewing syntenic regions of FPC maps', *Genome research,* 16(9), 1159-1168.

Specht, D. F. (1991) 'A general regression neural network', *Neural Networks, IEEE Transactions on,* 2(6), 568-576.

Špírek, M., Yang, J., Groth, C., Petersen, R. F., Langkjær, R. B., Naumova, E. S., Sulo, P., Naumov, G. I. and Piškur, J. (2003) 'High-rate evolution of Saccharomyces sensu lato chromosomes', *FEMS yeast research,* 3(4), 363-373.

Splinter, E., Grosveld, F. and de Laat, W. (2004) '3C technology: analyzing the spatial organization of genomic loci in vivo', *Methods in enzymology,* (375), 493-507.

Stebbins, G. L. (1958) 'The inviability, weakness, and sterility of interspecific hybrids', *Advances in genetics,* 9, 147-215.

Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C. and Coghlan, A. (2003) 'The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics', *PLoS Biol,* 1(2), e45.

Strachan, T. and Read, A. P. (1999) 'Molecular pathology'.

Strimmer, K. (2008) 'fdrtool: a versatile R package for estimating local and tail area-based false discovery rates', *Bioinformatics,* 24(12), 1461-2.

Sturtevant, A. (1926) 'A crossover reducer in Drosophila melanogaster due to inversion of a section of the third chromosome', *Biologisches Zentralblatt,* 46(12), 697-702.

Sturtevant, A. H. (1913) 'The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association', *Journal of experimental zoology,* 14(1), 43-59.

Sun, H.-F., Ernst, C., Yerle, M., Pinton, P., Rothschild, M., Chardon, P., Rogel-Gaillard, C. and Tuggle, C. (1999) 'Human chromosome 3 and pig chromosome 13 show complete synteny conservation but extensive gene-order differences', *Cytogenet Genome Res,* 85(3-4), 273-278.

Sutton, W. S. (1903) 'The chromosomes in heredity', *The Biological Bulletin,* 4(5), 231-250.

Suyama, M. and Bork, P. (2001) 'Evolution of prokaryotic gene order: genome rearrangements in closely related species', *Trends in Genetics,* 17(1), 10-13.

Suzuki, T., Kobayashi, I., Kanbe, T. and Tanaka, K. (1989) 'High frequency variation of colony morphology and chromosome reorganization in the pathogenic yeast Candida albicans', *Journal of general microbiology,* 135(2), 425.

Taft, R. J. and Mattick, J. S. (2004) 'Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences', *Genome biology,* 5(1), P1-P1.

Tanabe, H., Müller, S., Neusser, M., von Hase, J., Calcagno, E., Cremer, M., Solovei, I., Cremer, C. and Cremer, T. (2002) 'Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates', *Proceedings of the National Academy of Sciences,* 99(7), 4424-4429.

Tanaka, M., Suzuki, K., Morozumi, T., Kobayashi, E., Matsumoto, T., Domukai, M., Eguchi-Ogawa, T., Shinkai, H., Awata, T. and Uenishi, H. (2006) 'Genomic structure and gene order of swine chromosome 7q1. 1→ q1. 2', *Animal genetics,* 37(1), 10-16.

Tautz, D. and Renz, M. (1984) 'Simple sequences are ubiquitous repetitive components of eukaryotic genomes', *Nucleic acids research,* 12(10), 4127-4138.

Temperley, N. D., Berlin, S., Paton, I. R., Griffin, D. K. and Burt, D. W. (2008) 'Evolution of the chicken Toll-like receptor gene family: a story of gene gain and gene loss', *BMC Genomics,* 9(1), 62.

Tesler, G. (2002) 'GRIMM: genome rearrangements web server', *Bioinformatics,* 18(3), 492-493.

Thomas Jr, C. (1971) 'The genetic organization of chromosomes', *Annual review of genetics,* 5(1), 237-256.

Tiecke, E., Bangs, F., Blaschke, R., Farrell, E. R., Rappold, G. and Tickle, C. (2006) 'Expression of the short stature homeobox gene Shox is restricted by proximal and distal signals in chick limb buds and affects the length of skeletal elements', *Dev Biol,* 298(2), 585-96.

Trachana, K., Larsson, T. A., Powell, S., Chen, W. H., Doerks, T., Muller, J. and Bork, P. (2011) 'Orthology prediction methods: a quality assessment using curated protein families', *Bioessays,* 33(10), 769-780.

Trifonov, V. A., Dementyeva, P. V., Larkin, D. M., O'Brien, P. C. M., Perelman, P. L., Yang, F., Ferguson-Smith, M. A. and Graphodatsky, A. S. (2013) 'Transcription of a protein-coding gene on B chromosomes of the Siberian roe deer (Capreolus pygargus)', *BMC Biol,* 11, 90.

Trinh, P., McLysaght, A. and Sankoff, D. (2004) 'Genomic features in the breakpoint regions between syntenic blocks', *Bioinformatics,* 20(suppl 1), i318-i325.

Tumbleson, M. E. and Schook, L. B. (1996) *Advances in swine in biomedical research,* Plenum Pub Corp.

van Belkum, A., Scherer, S., van Alphen, L. and Verbrugh, H. (1998) 'Short-sequence DNA repeats in prokaryotic genomes', *Microbiology and Molecular Biology Reviews,* 62(2), 275-293.

Van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., Dekker, J. and Lander, E. S. (2010) 'Hi-C: a method to study the three-dimensional architecture of genomes'.

Van Hoide, K., Sahasrabuddhe, C. G. and Shaw, B. R. (1974) 'A model for particulate structure in chromatin', *Nucleic Acids Research,* 1(11), 1579-1586.

Vanderpool, H. Y. (2002) 'Critical ethical issues in clinical trials with xenotransplants', *Ethical issues in biotechnology*, 351.

Venturin, M., Gervasini, C., Orzan, F., Bentivegna, A., Corrado, L., Colapietro, P., Friso, A., Tenconi, R., Upadhyaya, M. and Larizza, L. (2004) 'Evidence for non-homologous end joining and non-allelic homologous recombination in atypical NF1 microdeletions', *Human genetics,* 115(1), 69-80.

Volff, J. N. (2006) 'Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes', *Bioessays,* 28(9), 913-922.

Völker, M., Backström, N., Skinner, B. M., Langley, E. J., Bunzey, S. K., Ellegren, H. and Griffin, D. K. (2010) 'Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution', *Genome Research,* 20(4), 503-511.

Wakamatsu, Y., Maynard, T. M., Jones, S. U. and Weston, J. A. (1999) 'NUMB localizes in the basal cortex of mitotic avian neuroepithelial cells and modulates neuronal differentiation by binding to NOTCH-1', *Neuron,* 23(1), 71-81.

Walter, M. A. and Goodfellow, P. N. (1993) 'Radiation hybrids: irradiation and fusion gene transfer', *Trends in Genetics,* 9(10), 352-356.

Wang, J., Wong, G. K. S., Ni, P., Han, Y., Huang, X., Zhang, J., Ye, C., Zhang, Y., Hu, J. and Zhang, K. (2002) 'RePS: a sequence assembler that masks exact repeats identified from the shotgun data', *Genome Research,* 12(5), 824-831.

Warnefors, M., Pereira, V. and Eyre-Walker, A. (2010) 'Transposable elements: insertion pattern and impact on gene expression evolution in hominids', *Molecular biology and evolution,* 27(8), 1955-1962.

Warren, R. L., Sutton, G. G., Jones, S. J. M. and Holt, R. A. (2007) 'Assembling millions of short DNA sequences using SSAKE', *Bioinformatics,* 23(4), 500-501.

Warren, R. L., Varabei, D., Platt, D., Huang, X., Messina, D., Yang, S. P., Kronstad, J. W., Krzywinski, M., Warren, W. C. and Wallis, J. W. (2006) 'Physical map-assisted whole-genome shotgun sequence assemblies', *Genome Research,* 16(6), 768-775.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., Searle, S., White, S., Vilella, A. J., Fairley, S., Heger, A., Kong, L., Ponting, C. P., Jarvis, E. D., Mello, C. V., Minx, P., Lovell, P., Velho, T. A. F., Ferris, M., Balakrishnan, C. N., Sinha, S., Blatti, C., London, S. E., Li, Y., Lin, Y.-C., George, J., Sweedler, J., Southey, B., Gunaratne, P., Watson, M., Nam, K., Backström, N., Smeds, L., Nabholz, B., Itoh, Y., Whitney, O., Pfenning, A. R., Howard, J., Völker, M., Skinner, B. M., Griffin, D. K., Ye, L., McLaren, W. M., Flicek, P., Quesada, V., Velasco, G., Lopez-Otin, C., Puente, X. S., Olender, T., Lancet, D., Smit, A. F. A., Hubley, R., Konkel, M. K., Walker, J. A., Batzer, M. A., Gu, W., Pollock, D. D., Chen, L., Cheng, Z., Eichler, E. E., Stapley, J., Slate, J., Ekblom, R., Birkhead, T., Burke, T., Burt, D., Scharff, C., Adam, I., Richard, H., Sultan, M., Soldatov, A., Lehrach, H., Edwards, S. V., Yang, S.-P., Li, X., Graves, T., Fulton, L., Nelson, J., Chinwalla, A., Hou, S., Mardis, E. R. and Wilson, R. K. (2010) 'The genome of a songbird', *Nature,* 464(7289), 757-62.

Waterston, R. H., Lander, E. S. and Sulston, J. E. (2002) 'On the sequencing of the human genome', *Proceedings of the National Academy of Sciences,* 99(6), 3712.

Watson, J. D. and Crick, F. H. (1953) 'Molecular structure of nucleic acids', *Nature,* 171(4356), 737-738.

Welten, M. C. M., Verbeek, F. J., Meijer, A. H. and Richardson, M. K. (2005) 'Gene expression and digit homology in the chicken embryo wing', *Evol Dev,* 7(1), 18-28.

White, M. (1969) 'Chromosomal rearrangements and speciation in animals', *Annual review of genetics,* 3(1), 75-98.

Wind, A. E., Larkin, D. M., Green, C. A., Elliott, J. S., Olmstead, C. A., Chiu, R., Schein, J. E., Marra, M. A., Womack, J. E. and Lewin, H. A. (2005) 'A high-resolution whole-genome cattle–human comparative map reveals details of mammalian

chromosome evolution', *Proceedings of the National Academy of Sciences of the United States of America,* 102(51), 18526.

Wolffe, A. (1998) *Chromatin: structure and function,* Access Online via Elsevier.

Womack, J. E. (2012) *Bovine genomics,* Wiley Online Library.

Woodcock, C. L. (2005) 'A milestone in the odyssey of higher-order chromatin structure', *Nature structural & molecular biology,* 12(8), 639-640.

Yamanaka, K., Fang, L. and Inouye, M. (1998) 'The CspA family in Escherichia coli: multiple gene duplication for stress adaptation', *Molecular microbiology,* 27(2), 247-255.

Young, A. L., Abaan, H. O., Zerbino, D., Mullikin, J. C., Birney, E. and Margulies, E. H. (2010) 'A new strategy for genome assembly using short sequence reads and reduced representation libraries', *Genome Research,* 20(2), 249-256.

Zeller, R., López-Ríos, J. and Zuniga, A. (2009) 'Vertebrate limb bud development: moving towards integrative analysis of organogenesis', *Nat Rev Genet,* 10(12), 845-58.

Zeng, X., Nesbitt, M. J., Pei, J., Wang, K., Vergara, I. A. and Chen, N. (2008) *OrthoCluster: a new tool for mining synteny blocks and applications in comparative genomics,* translated by ACM, 656-667.

Zerbino, D. R. and Birney, E. (2008) 'Velvet: algorithms for de novo short read assembly using de Bruijn graphs', *Genome Research,* 18(5), 821-829.

Zerbino, D. R., McEwen, G. K., Margulies, E. H. and Birney, E. (2009) 'Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler', *PLoS ONE,* 4(12), e8407.

Zhan, X., Pan, S., Wang, J., Dixon, A., He, J., Muller, M. G., Ni, P., Hu, L., Liu, Y., Hou, H., Chen, Y., Xia, J., Luo, Q., Xu, P., Chen, Y., Liao, S., Cao, C., Gao, S., Wang, Z., Yue, Z., Li, G., Yin, Y., Fox, N. C., Wang, J. and Bruford, M. W. (2013) 'Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle', *Nat Genet,* 45(5), 563-6.

Zhan, X., Pan, S., Wang, J., Dixon, A., He, J., Muller, M. G., Ni, P., Hu, L., Liu, Y., Hou, H., Chen, Y., Xia, J., Luo, Q., Xu, P., Liao, S., Cao, C., Gao, S., Wang, Z., Yue, Z., Li, G., Yin, Y., Fox, N. C. and Bruford, M. W. (2013) 'Peregrine and

saker falcon genome sequences provide insights into evolution of a predatory lifestyle', *Nat Genet,* 45(5), 563-6.

Zhang, B., Schmoyer, D., Kirov, S. and Snoddy, J. (2004) 'GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies', *BMC bioinformatics,* 5(1), 16.

Zhang, G., Cai, L., Qiye, L., Larkin, D. M., Lee, C., Storz, J. F., Antunes, A., Meredith, R. W., Odeen, A., Cui, J., Zhou, Q., Xu, L., Pan, H., Wang, Z., Jin, L., Zhang, P., Hu, H., Yang, W. and al., e. (2014) 'Comparative genomics across modern bird species reveal insights into Pan-avian genome evolution and trait biodiversity', *Science,* submitted.

Zhang, J. (2003) 'Evolution by gene duplication: an update', *Trends Ecol Evol,* 18(6), 292-298.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J. and Shen, B. (2011) 'A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies', *PLoS ONE,* 6(3), e17915.

Zhang, Y., McCord, R. P., Ho, Y.-J., Lajoie, B. R., Hildebrand, D. G., Simon, A. C., Becker, M. S., Alt, F. W. and Dekker, J. (2012) 'Spatial organization of the mouse genome and its role in recurrent chromosomal translocations', *Cell,* 148(5), 908-921.

Zhang, Z. and Gerstein, M. (2003) 'Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes', *Nucleic Acids Research,* 31(18), 5338-5348.

Zhang, Z., Harrison, P. and Gerstein, M. (2002) 'Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome', *Genome Research,* 12(10), 1466-1482.

Zhang, Z., Harrison, P. M., Liu, Y. and Gerstein, M. (2003) 'Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome', *Genome Research,* 13(12), 2541-2558.

Zhao, H. and Bourque, G. (2009) 'Recovering genome rearrangements in the mammalian phylogeny', *Genome Research,* 19(5), 934-942.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S. and Singh, U. (2006) 'Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically

regulated intra-and interchromosomal interactions', *Nature genetics,* 38(11), 1341-1347.

Zheng, Q. and Wang, X.-J. (2008) 'GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis', *Nucleic acids research,* 36(suppl 2), W358-W363.

Zhou, X. and Su, Z. (2007) 'EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species', *BMC Genomics,* 8(1), 246.

Zhou, Y. and Landweber, L. F. (2007) 'BLASTO: a tool for searching orthologous groups', *Nucleic acids research,* 35(suppl 2), W678-W682.

Zhou, Y., Young, J. A., Santrosyan, A., Chen, K., Yan, S. F. and Winzeler, E. A. (2005) 'In silico gene function prediction using ontology-based pattern identification', *Bioinformatics,* 21(7), 1237-1245.